Kodowanie predykcyjne i reprezentacjonizm

(wersja robocza... wczesna)

Paweł Gładziejewski

Abstrakt: Zgodnie z teorią kodowania predykcyjnego (TKP), percepcja i działanie opierają się na procesie minimalizacji błędu predykcyjnego, czyli minimalizacji rozbieżności pomiędzy dochodzącymi oddolnie „ze świata" sygnałami sensorycznymi a odgórnie wygenerowanymi sygnałami „wirtualnymi". Wielu autorów sądzi, że TKP dysponuje znaczącą mocą eksplanacyjną i ma potencjał dostarczenia pierwszej w historii nauk kognitywnych niezwykle precyzyjnej, matematycznie wyrafinowanej, zunifikowanej wizji działania systemu poznawczego. Dość powszechnie przyjmuje się też, że TKP jest teorią reprezentacjonistyczną – że ma ona wbudowane założenie o tym, iż kontakt poznawczy ze światem zapośredniczony jest za pomocą wewnętrznych reprezentacji. Nikt jednak nie podjął do tej pory próby precyzyjnego wyrażenia reprezentacjonistycznych założeń TKP oraz dokonania oceny, czy zawarty w tej koncepcji sposób rozumienia reprezentacji jest prawomocny – czy spełnia on to, co William Ramsey (2007) nazywa „wymogiem opisu zadań". W artykule tym argumentuję, że reprezentacjonistyczne pretensje TKP są uzasadnione, ponieważ teoria ta postuluje byty rzeczywiście realizujące funkcję reprezentacji, mianowicie przewodzące działaniem, odłączalne, strukturalne modele pozwalające na rozpoznanie błędu reprezentacyjnego.

1. Wstęp

Historyczny związek pomiędzy rozwojem kognitywistyki a tym, jak zmieniają się stanowiska w sprawie istnienia i natury reprezentacji mentalnych, można scharakteryzować jako koewolucję. Z jednej strony, mająca filozoficzny rodowód idea, że poznanie zapośredniczone jest wewnętrznymi reprezentacjami świata, odegrała ważną rolę w tym, jak kognitywiści przez wiele lat pojmowali swoje zadanie „wypełniania" czarnej skrzynki behawioryzmu. Skrzynka ta miała zostać wypełniona właśnie (naturalistycznie pojmowanymi) reprezentacjami świata. Zarazem wydaje się, że filozoficzny antyreprezentacjonizm odegrał rolę „inspiratora" w rozwoju antyreprezentacjonistycznych stanowisk i podejść teoretycznych w naukach kognitywnych, takich, jak enaktywizm czy

niektóre formy koncepcji poznania ucieleśnionego. Z drugiej strony, rozwój teoretyczny kognitywistyki kształtował przez lata nasze pojęcie tego, czym są reprezentacje, jak również wpływał na przyjmowane stanowiska dotyczące tego, czy poznanie w ogóle reprezentacje wykorzystuje. Na przykład, przejście z kognitywistyki symbolicznej (Starej Dobrej Sztucznej Inteligencji) w kierunku kognitywistyki opartej na modelowaniu koneksyjnym oznaczało zarazem odejście od pojmowania wewnętrznych reprezentacji na wzór przekonań, pragnień i innych postaw propozycjonalnych; natomiast rozwój podejść opartych na modelowaniu dynamicznym dał nowe perspektywy dla prób sformułowania oraz obrony stanowiska antyreprezentacjonistycznego.

Związek pojęcia reprezentacji z kognitywistyką komplikuje się jednak dodatkowo ze względu na problem, któremu w ostatnich latach wyraz dała praca *Representation Reconsidered* Williama Ramseya (2007). Ramsey pokazał, że reprezentacjonistyczne pretensje różnych podejść i teorii rozwijanych w kognitywistyce są często wysoce problematyczne. Przedstawiciele kognitywistyki nierzadko posługują się terminem „reprezentacja" na tyle swobodnie, że przestaje on desygnować struktury czy mechanizmy, w których przypadku można konkluzywnie pokazać, że na takie miano zasługują. Struktury czy mechanizmy te nie czynią zadość temu, co Ramsey nazywa „wymogiem opisu obowiązków": niejednokrotnie trudno jest pokazać, w jakim sensie „reprezentacja" postulowana w ramach danej teorii kognitywistycznej rzeczywiście odgrywa w systemie poznawczym rolę funkcjonalną, którą można w zasadny sposób skategoryzować jako *reprezentowanie* czegoś. Pojęcie reprezentacji przynajmniej czasem wydaje się w naukach kognitywnych pustym ornamentem, który nie odgrywa żadnej realnej roli eksplanacyjnej.

W obecnym krajobrazie teoretycznym kognitywistyki coraz bardziej prominentną rolę odgrywa teoria kodowania predykcyjnego (dalej: TKP). Zgodnie z tą koncepcją, mózg jest maszyną predykcyjną, nieustannie zaangażowaną w generowanie odgórnych, „wirtualnych" sygnałów sensorycznych po to, by minimalizować błąd predykcyjny, rozumiany jako różnica pomiędzy sygnałem wewnętrznie wygenerowanym a sygnałem sensorycznym napływającym „ze świata" przez zmysły. Wielu autorów wiąże z TKP ogromne nadzieje (por. Clark 2013b; Friston 2010; Hohwy 2013; Huang 2008). Teoria ta ma nie tylko w matematycznie elegancki sposób wyjaśnić percepcję i działanie oraz ich wzajemne związki, ale ma też unifikować kognitywistykę, pokazując, że minimalizacja błędu predykcyjnego stanowi najważniejszą, a być może wręcz jedyną funkcję mózgu. Jeśli oczekiwania te są przynajmniej częściowo zasadne, to niewykluczone, że wchodzimy właśnie w nową, „predykcyjną" erę w historii nauk kognitywnych.

W artykule tym chcę zbadać, jak koncepcja kodowania predykcyjnego wpisuje się w nieustającą debatę na temat istnienia i natury reprezentacji mentalnych. Na ogół przyjmuje się, że TKP jest koncepcją reprezentacjonistyczną (por. Clark 2013b; Hohwy 2013). Mózg/umysł zawdzięcza w takiej perspektywie swoją zdolność do minimalizacji błędu predykcyjnego temu, że dysponuje on bogatą reprezentacją przyczynowej struktury świata. Jednakże ten reprezentacjonistyczny aspekt TKP nie został do tej pory bardziej poddany skrupulatnej i systematycznej analizie[1]. W artykule tym chcę zadać dwa pytania, które, jak sądzę, nie doczekały się jeszcze całkowicie zadowalającej odpowiedzi w literaturze. Po pierwsze, chcę zapytać o to, jak dokładnie powinniśmy rozumieć reprezentacje w świetle TKP; albo: jaka koncepcja reprezentacji jest zawarta w TKP? Po drugie, chcę przyjąć perspektywę wspomnianego Ramseya i zadać pytanie, czy reprezentacje ujmowane w świetle TKP rzeczywiście zasługują na to miano – czy rzeczywiście realizują one funkcje, które można uznać za w jakimś nietrywialnym sensie reprezentacyjne.

Artykuł będzie mieć następującą strukturę. W sekcji 2 krótko zrekapituluję zasadnicze twierdzenia koncepcji kodowania predykcyjnego. W sekcji 3 przedstawię szerzej Ramseya wymóg opisu zadań oraz omówię szereg własności funkcjonalnych charakteryzujących prototypową pozamentalną reprezentację, jaką jest mapa kartograficzna. Inaczej mówiąc, przedstawię „opis zadań" reprezentacji rozumianej jako mapa kartograficzna. W sekcji 4 pokażę, że TKP postuluje reprezentacje, których profil funkcjonalny jest nietrywialnie zbliżony do profilu funkcjonalnego map kartograficznych. Mówiąc wprost, w bronionym przeze mnie ujęciu, TKP postuluje przewodzące działaniem, odłączalne, strukturalne reprezentacje pozwalające na rozpoznanie błędu reprezentacyjnego. Broniąc tego twierdzenia nie tylko wskażę, w jakim sensie TKP postuluje wewnętrzne reprezentacje, ale pokażę zarazem, że te reprezentacjonistyczne pretensje są uzasadnione. Jeśli mam rację, to TKP jest być może teorią tak reprezentacjonistyczną, jak to tylko w kognitywistyce możliwe. Sekcja 5 zawiera krótkie podgumowanie tego artykułu.

2. Teoria kodowania predykcyjnego – zasadnicze twierdzenia

TKP wspiera się na założeniu, że aby w odpowiedni sposób kontrolować działanie, system poznawczy (mózg) musi „odgadywać" leżące po stronie świata – i *bezpośrednio* dla

[1] Ważnym wyjątkiem jest tu praca Jakoba Hohwy'ego (2013). Pewne rozwiązania zaproponowane przez Hohwy'ego wykorzystam w bronionych tu przeze mnie propozycjach.

niego niedostępne – przyczyny napływających do niego sygnałów sensorycznych (Clark 2013a, 2013b; Friston 2010; Friston, Kiebel 2009; Friston, Sephan 2007; Hohwy 2013; Huang, Rao 2011; Lee, Mumford 2003; Rao, Ballard 1999). Na przykład, aby wygenerować stosowną reakcję, system poznawczy musi oszacować, czy napływające do niego sygnały zostały przyczynowo wywołane przez tygrysa, czy może przez pluszową imitację tygrysa. Zadanie tego rodzaju jest w nieunikniony sposób obciążone niepewnością, ponieważ nie istnieje jednoznaczna przyczynowa odpowiedniość pomiędzy napływającym „ze świata" sygnałem a jego środowiskową przyczyną. Jeden i ten sam sygnał zmysłowy może mieć potencjalnie wiele różnych przyczyn, a zarazem jedna przyczyna może, w zależności od okoliczności, wywoływać wiele różnych sygnałów zmysłowych. Zgodnie z TKP, mózg radzi sobie z zadaniem wnioskowania o przyczynach sygnałów realizując czy implementując wnioskowanie Bayesowskie. To znaczy, napływające sygnały sensoryczne traktowane są jako dane, a system wybiera taką „hipotezę" na temat ukrytej przyczyny danych, która ma najwyższe prawdopodobieństwo końcowe (jest najbardziej prawdopodobna w świetle danych), oszacowywane na podstawie wiarygodności tej hipotezy (prawdopodobieństwa otrzymania danych przy założeniu, że hipoteza ta jest prawdziwa) oraz jej prawdopodobieństwa wstępnego (prawdopodobieństwa prawdziwości tej hipotezy niezależnie od posiadanych w określonych okolicznościach danych), biorąc pod uwagę prawdopodobieństwo końcowe hipotez alternatywnych.

W świetle TKP, realizowanie wnioskowania Bayesowskiego możliwe jest dzięki temu, że system poznawczy dysponuje wielopoziomowym modelem generatywnym napływającego do niego sygnału zmysłowego, który to model wykorzystywany jest w celu minimalizacji błędu predykcyjnego (Clark 2013a, 2013b; Hinton 2007; Hohwy 2013; Huang, Rao 2011; Lee, Mumford 2003; Rao, Ballard 1999). Model generatywny ma „rekapitulować" czy „odzwierciedlać" przyczynowo-probabilistyczną strukturę leżących po stronie świata przyczyn sygnałów sensorycznych. Zajmuje się on generowaniem, w odgórny sposób, wewnętrznego, „wirtualnego" sygnału sensorycznego, który przewiduje sygnał oddolnie wywołany przez przyczynę środowiskową. Proces ten ma docelowo zmierzać do minimalizacji błędu predykcyjnego, rozumianego jako rozbieżność pomiędzy obydwoma sygnałami. W domyśle, błąd predykcyjny najlepiej minimalizowany jest wtedy, gdy system przyjmie w ramach modelu poprawną hipotezę na temat przyczynowego źródła dochodzącego do niego sygnału i swoje predykcje generuje na podstawie tej hipotezy. Zakładając, że właśnie obserwujemy tygrysa, hipoteza (ukryty parametr modelu generatywnego odpowiadający tej hipotezie), że dochodzące do systemu sygnały są wywołane obecnością

tygrysa, będzie wytwarzać predykcje sensoryczne generujące mniejszy błąd predykcyjny, niż hipoteza, że przyczyną jest pluszowa zabawka lub kot domowy o specyficznym zabarwieniu sierści. Zgodnie z często przyjmowaną neuronalną interpretacją TKP, predykcje generowane na podstawie modelu przesyłane są w mózgu za pomocą zstępujących połączeń synaptycznych; połączenia wstępujące zajmują się zaś nie tyle – jak się to klasycznie przyjmuje – detekcją cech, lecz przesyłaniem sygnału informującego o rozmiarze błędu predykcyjnego, to znaczy o rozmiarze rozbieżności pomiędzy sygnałem odgórnym a sygnałem sensorycznym (Clark 2013a, 2013b; Hohwy 2013; Huang, Rao 2011; Lee, Mumford 2003; Rao, Ballard 1999). W TKP zawarta jest więc Kantowska w duchu wizja percepcji jako aktywnego („spontanicznego") procesu poznawczej interpretacji danych zmysłowych, a nie pasywnej recepcji bodźców.

Co istotne, model generatywny postulowany w ramach TKP jest wielopoziomowy, a każda jego warstwa zaangażowana jest wyłącznie w minimalizowanie błędu predykcyjnego na bezpośrednio niższym poziomie (Clark 2013b; Hinton 2007; Hohwy 2013; Lee, Mumford 2003). Każda warstwa „śledzi" też regularności przyczynowo-probabilistyczne występujące na różnych porządkach czasowych. Mówiąc obrazowo: inna, wyżej leżąca w hierarchii warstwa generuje predykcję, że po zimie nastąpi wiosna, a inna, niższa warstwa zajmuje się przewidywaniem błyskawicznie zachodzących zmian drobnych percepcyjnych detali – odcieni barw, konturów, etc. – obserwowanego obecnie obiektu. Wielopoziomowość sprawia, że proces minimalizacji błędu jest obliczeniowo wykonalny i nie wymaga bezpośredniego predykcyjnego „przechodzenia" z bardzo abstrakcyjnych hipotez na temat świata do predykcji dotyczących najdrobniejszych, szybko się zmieniających detali percepcyjnych.

Na gruncie TKP przyjmuje się, że system poznawczy ma do swojej dyspozycji dwie strategie minimalizacji błędu predykcyjnego (Clark 2013a, 2013b; Friston 2010; Friston, Kiebel 2009; Friston, Stephan 2007; Hohwy 2013). Jedną z nich jest po prostu przyjęcie hipotezy (lub rewizja hipotezy uprzednio przyjętej), która możliwie najlepiej minimalizuje błąd predykcyjny. Druga strategia polega na podjęciu działania, to znaczy takim zainterweniowaniu w świat, które skutkuje dopasowaniem sygnału dochodzącego do sygnału przewidywanego (można to na przykład zrealizować podejmując działanie urzeczywistniające przyjętą hipotezę). Pierwszą strategię nazywa się w TKP „wnioskowaniem percepcyjnym", a drugą – „wnioskowaniem aktywnym". Percepcja świata (wnioskowanie percepcyjne) oraz działanie w świecie (wnioskowanie aktywne) stanowią w takiej perspektywie dwie strony jednej monety – jednego procesu minimalizacji błędu predykcyjnego – i są opisywane za pomocą tego samego formalizmu matematycznego. Nasz kontakt poznawczy z otoczeniem

oparty jest na nieustannym, płynnym oscylowaniu pomiędzy wnioskowaniem percepcyjnym a wnioskowaniem aktywnym.

W artykule tym będę opierał się na stosunkowo mocnej wersji TKP, która sytuuje ideę kodowania predykcyjnego w szerszym kontekście biologii teoretycznej i wyraża ją za pomocą pochodzącego z fizyki statystycznej pojęcia swobodnej energii (Friston 2010, 2013; Friston, Kiebel 2009; Friston, Setphan 2007; Hohwy 2013). Z takiego punktu widzenia, cały proces minimalizacji błędu predykcyjnego jest, fundamentalnie, narzędziem samoorganizacji. Zgodnie z tą wersją TKP, minimalizacja błędu predykcyjnego – matematycznie równoważna procesowi optymizacji swobodnej energii, czyli entropii systemu – stanowi dla organizmu pośredni sposób kontroli działania w taki sposób, by unikać znajdowania się w sytuacjach, które są zaskakujące relatywnie do jego fenotypu. Są to sytuacje, które zbliżają ten system do równowagi termodynamicznej, czyli śmierci cieplnej. Taką sytuacją może być długotrwały brak pożywienia albo poważne uszkodzenie ciała własnego. Model generatywny wyposażony jest w „wiedzę" wstępną nakładającą określone granice na oczekiwany czy spodziewany sygnał sensoryczny, a percepcja i działanie stanowią dwa sposoby, w jakie system może sprawić, że sygnał będzie znajdował się w tych granicach. Błąd predykcyjny stanowi dla systemu pośrednią aproksymację tego, jak zaskakująca (relatywnie do jego fenotypu) jest sytuacja, w której ten system się znalazł – im większy błąd, tym większe zaskoczenie. Oznacza to, że predykcje generowane przez model generatywny nie są „neutralne" ze względu na (ewolucyjnie określony) „interes", jakim jest unikanie okoliczności obniżających integralność (organizację) termodynamiczną organizmu oraz poszukiwanie okoliczności, które tej integralności sprzyjają. Teza ta dotyczy także wnioskowania aktywnego i oznacza, że proces minimalizacji błędu predykcyjnego służy takiej kontroli działania, by unikać okoliczności zwiększających entropię systemu i mogących doprowadzić do jego termodynamicznej śmierci. Na przykład, biorąc pod uwagę, że bliska obecność drapieżnika takiego, jak tygrys, jest okolicznością zaskakującą relatywnie do ludzkiego fenotypu, przyjęcie (na podstawie wnioskowania percepcyjnego) hipotezy, że tygrys jest blisko nas, będzie generowało oczekiwanie, że się w takiej sytuacji *nie* znajdujemy – oczekiwanie, które może być zrealizowane za pomocą wnioskowania aktywnego, czyli działania. (Tak w TKP opisuje się sytuację, kiedy ktoś widzi w pobliżu tygrysa i zaczyna od niego uciekać). Jak okaże się w dalszej części tego artykułu, ów związek pomiędzy minimalizacją błędu predykcyjnego a adaptywnym działaniem w świecie ma istotne znaczenie, jeśli chcemy zrozumieć rolę pojęcia reprezentacji w TKP.

3. Reprezentacje, kodowanie predykcyjne i wymóg opisu zadań

3.1. Ramseya wymóg opisu zadań i strategia „porównaj-z-prototypem”

Jak już wspomniałem, zasadniczym celem tego artykułu jest określenie, jak TKP łączy się z ideą, że poznanie zapośredniczone jest wewnętrznymi reprezentacjami – i czy w TKP jest rzeczywiście miejsce dla tej idei. Podejście, jakie obiorę przy podejmowaniu tego zagadnienia, inspirowane jest przeprowadzoną przez Williama Ramseya (2007) krytyczną analizą tego, jak pojęcie reprezentacji funkcjonuje w kognitywistyce jako takiej. Ramsey wychodzi w swojej pracy od obserwacji, że kognitywiści często – zbyt często – posługują się terminem „reprezentacja” w sposób na tyle liberalny i niezobowiązujący, że przestaje ono desygnować struktury, o których możemy w zasadny, zrozumiały i eksplanacyjnie wartościowy sposób orzec, że rzeczywiście realizują one w systemie poznawczym funkcję reprezentacji. W pewnych wypadkach badacze wykazują wręcz tendencję do nazywania „reprezentacją” w zasadzie dowolnej struktury pośredniczącej przyczynowo pomiędzy środowiskiem a działaniem, przez co pojęcie reprezentacji staje się eksplanacyjnie trywialne i zbędne. Dlatego też Ramsey twierdzi, że kognitywistyczne odwołania do reprezentacji wymagają uzasadnienia i nie powinny być domyślenie traktowane jako teoretycznie „niewinne”. Reprezentacjonistyczny status dowolnej teorii kognitywistycznej nie powinien być sankcjonowany przez to, że zwolennicy tej teorii posługują się terminem „reprezentacja”, ale przez to, że teoria ta postuluje (jako eksplananda) wewnętrzne stany lub struktury, w przypadku których można pokazać, *dlaczego* zasługują one na miano reprezentacji. Ramsey proponuje, że testem wartości pojęcia reprezentacji zawartego w danej teorii (podejściu) powinno być zawsze zadanie pytania, czy pojęcie to czyni zadość czemuś, co nazywa on „wymogiem opisu zadań” [*job description challenge*]. Aby sprostać temu wymogowi, należy pokazać, w jakim sensie albo w jaki sposób struktury desygnowane przez dane pojęcie reprezentacji rzeczywiście można uznać – w naturalny, zrozumiały i zasadny sposób – za realizujące rolę funkcjonalną polegającą na *reprezentowaniu* czegoś (odgrywające „zadania” typowe dla reprezentacji).

Poddając testowi opartemu na wymogu opisu zadań różne kognitywistyczne pojęcia reprezentacji, Ramsey (2007) często wykorzystuje strategię, którą nazwę na bieżące potrzeby „strategią porównaj-z-prototypem”. Najpierw wskazuje on przykład pewnego rodzaju struktury, którą *preteoretycznie* możemy sklasyfikować jako reprezentację w sposób całkowicie niekontrowersyjny. Wskazuje on też, na czym polega funkcja tej reprezentacji: co

takiego w tym, co ta struktura „robi" dla swoich użytkowników czyni ją reprezentacją. Można powiedzieć, że struktura ta wyznacza *prototypowy* przykład reprezentacji[2]. Następnie Ramsey skupia się na określonym rodzaju „reprezentacji" postulowanym przez *kognitywistów* i stara się odpowiedzieć na pytanie, czy funkcja realizowana w systemie poznawczym przez desygnaty tego pojęcia jest jakoś nietrywialnie zbliżona czy analogiczna do tego, jak funkcjonuje opisany wcześniej preteoretyczny prototyp. Inaczej mówiąc, Ramsey pyta, czy „reprezentacje" w tym sensie realizują funkcję na tyle zbliżoną do prototypu, byśmy i je mogli poprawnie skategoryzować jako reprezentacje. Jeśli tak jest, dane pojęcie reprezentacji spełnia wymóg opisu zadań[3].

Stosowanie strategii porównaj-z-prototypem zaskakująco często prowadzi Ramseya do stawiania *negatywnych* diagnoz w sprawie statusu różnych kognitywistycznych pojęć reprezentacji. Na przykład, autor ten poświęca wiele miejsca, by wskazać ważną rozbieżność pomiędzy sposobem funkcjonowania *zewnętrznych* reprezentacji indeksowych, to jest reprezentacji opartych na współzmienności (takich, jak wskazówka kompasu) – które to pełnią tu rolę prototypu – a postulowanymi często przez kognitywistów „reprezentacjami" pojętymi jako *wewnętrzne* detektory, które to również działają w oparciu o współzmienność

[2] Warto zwrócić uwagę, że w takiej perspektywie, status reprezentacji zawdzięcza się podobieństwu do prototypu, a *nie* temu, że spełnia się jakąś *definicję* reprezentacji. Ramsey argumentuje, że dotychczasowe próby zdefiniowania reprezentacji w kontekście kognitywistyki – wskazania warunków osobno koniecznych, a razem wystarczających do bycia reprezentacją – spełzają na niczym, ponieważ wszystkie definicje otwarte są na kontrprzykłady (to jest można wskazać przykłady, które spełniają podaną definicję choć ewidentnie nie są reprezentacjami, lub przykłady, które nie spełniają definicji, lecz reprezentacjami ewidentnie są).

[3] Można oponować, że strategia taka oparta jest na przesadnym konserwatyzmie pojęciowym. Przypisuje ona bowiem zbyt znaczną rolę naszym preteoretycznym intuicjom na temat reprezentacji i nie dopuszcza możliwości, że reprezentacje w sensie istotnym dla kognitywistyki będą się radykalnie różnić od struktur, które preteoretycznie kategoryzujemy jako reprezentacje. Na uwagę tę są dwie odpowiedzi (por. także dyskusję tego problemu w: Ramsey 2007). Po pierwsze, podejście oparte na porównaniu ze stereotypem wcale nie musi być *jedyną* strategią pozwalającą na ocenę, czy dane pojęcie reprezentacji spełnia wymóg opisu zadań. Na przykład, sam Ramsey nie zawsze w swojej książce korzysta ze strategii porównaj-z-prototypem (por. na przykład jego omówienie IO-reprezentacji, czyli reprezentacji typu wejście-wyjście; Ramsey 2007). Po drugie, użycie strategii porównaj-z-prototypem *nie* wyklucza, że reprezentacje w sensie kognitywistycznym będą posiadały wiele własności, których nie posiadają prototypy – własności, których na poziomie *preteoretycznym* reprezentacjom nie przypisujemy, a być może nawet nie dysponujemy pojęciami, by je wyrazić. Ponadto reprezentacje w sensie kognitywistycznym mogą co prawda podzielać pewne własności funkcjonalne z prototypami, ale te pierwsze mogą egzemplifikować owe własności w nieco inny sposób i w nieco innym sensie, niż te ostatnie (por. sekcja 4.1).

(przykładem tu są słynne żabie detektory owadów; por. Lettvin *et al.* 1959). Ramsey przeprowadza drobiazgową analizę mającą pokazać, że pomiędzy zewnętrznymi indeksami a wewnętrznymi detektorami zachodzi ważna funkcjonalna różnica. Zwraca on uwagę, że zewnętrzne indeksy odgrywają rolę reprezentacji ponieważ są interpretowane przez swoich użytkowników i dzięki temu *informują* ich o jakichś stanach rzeczy. Tymczasem wewnętrzne detektory okazują się odgrywać rolę wewnętrznych „pośredników" przyczynowych, których funkcją polega na byciu uruchamianymi przez pewne okoliczności środowiskowe i wywoływanie wtedy pewnego efektu (na przykład, wystrzelenie żabiego języka w kierunku owada). Skoro, jak przekonuje Ramsey, taka „mediacja" przyczynowa nie może być utożsamiona z *reprezentowanie*m czegoś – w żadnym zrozumiałym i eksplanacyjnie wartościowym sensie słowa „reprezentowanie" – to wewnętrzne detektory nie spełniają wymogu opisu zadań i nie są w istocie reprezentacjami (por. zbliżoną argumentację w: Hutto, Myin 2013).

Nie chcę tu rozstrzygać, czy Ramseyowska krytyka pojęcia reprezentacji wewnętrznych jako detektorów jest konkluzywna. Przywołuję ją tu z dwóch powodów. Po pierwsze, by pokazać, że posługiwanie się strategią porównaj-z-prototypem może prowadzić do mocno rewizjonistycznych konsekwencji, to jest można za jej pomocą pokazać, że reprezentacyjny status stanów czy struktur rutynowo nazywanych w kognitywistyce „reprezentacjami" jest (co najmniej) problematyczny. Po drugie, chciałem za pomocą tego przykładu zilustrować strategię argumentacyjną, którą sam zastosuję tu wobec pojęcia reprezentacji zawartego w TKP.

3.2. Reprezentacje w TKP i mapa kartograficzna jako prototyp reprezentacji

Wróćmy teraz do kodowania predykcyjnego. Jak już wspomniałem na początku tego artykułu, w literaturze niemal rutynowo przyjmuje się, że wizja umysłu czy poznania zawarta w TKP jest reprezentacjonistyczna. Autorzy opisujący tę koncepcję często i chętnie posługują się terminem „reprezentacja", stwierdzając chociażby, że zgodnie z TKP mózg używa „reprezentacji probabilistycznych" świata; że neurony generujące odgórnie sygnał predykcyjny są „jednostkami reprezentującymi" przyczynę napływającego oddolnie sygnału sensorycznego; albo że proces minimalizacji błędu predykcyjnego jest zarazem procesem minimalizacji rozbieżności pomiędzy tym, jak się rzeczy mają w świecie, a tym, jak je *reprezentujemy* (por. Clark 2013b; Friston, Stephan 2007; Hinton 2007). Niestety, z jednym ważnym wyjątkiem (Hohwy 2013), filozofowie oraz kognitywiści piszący o TKP do tej pory

nie podejmowali prób doprecyzowania, co dokładnie mają na myśli mówiąc o reprezentacjach – czy też w jakim dokładnie sensie ich zdaniem TKP postuluje istnienie wewnętrznych reprezentacji – oraz dlaczego sądzą, że mówienie o reprezentacjach jest tu zasadne. Być może dlatego słowo „reprezentacja" wydaje się przynajmniej czasem odgrywać w dyskusji nad TKP rolę heurystycznego narzędzia ułatwiającego wyjaśnienie tej koncepcji, a nie rolę terminu, za pomocą którego oznaczana jest jakaś ważna, zawarta w tej teorii teza o relacji pomiędzy umysłem a światem.

Chcę tutaj zastosować Ramseyowską metodologiczną „podejrzliwość" wobec powoływania się przez kognitywistów na reprezentacje właśnie w kontekście TKP. Mówiąc precyzyjniej, chcę argumentować za dwiema tezami. Po pierwsze, chcę zaproponować, że w TKP zawarta jest wizja reprezentacji mentalnych jako określonego rodzaju wewnętrznych reprezentacji *strukturalnych*, to jest reprezentacji działających na podstawie podobieństwa strukturalnego zachodzącego pomiędzy samą reprezentacją a tym, co reprezentowane. W następnej sekcji wyjaśnię, co przez to rozumiem i dlaczego sądzę, że właśnie tak powinniśmy rozumieć reprezentacje w kontekście TKP. Po drugie, chcę pokazać, że reprezentacje, na jakie powołują się zwolennicy kodowania predykcyjnego, *nie* podzielają losu przypisywanego przez Ramseya wewnętrznym detektorom. W TKP zawarta jest wartościowa, spełniająca wymóg opisu zadań wizja tego, czym są reprezentacje w systemie poznawczym. Jest tak dlatego, że reprezentacje postulowane przez TKP w nietrywialnym zakresie przypominają pod względem swoich ról funkcjonalnych to, jak funkcjonują *mapy kartograficzne*.

Moja argumentacja na rzecz tezy, że reprezentacje postulowane w ramach TKP spełniają wymóg opisu zadań, wykorzystuje strategię porównaj-z-prototypem. Rolę reprezentacyjnego prototypowego „probierzu" reprezentacyjności odgrywać będzie tu reprezentacja zewnętrzna i artefaktualna, jaką jest wspomniana mapa kartograficzna. Skupię się teraz właśnie na dostarczeniu „opisu zadań" charakteryzujących taką mapę. Jak sądzę, mapy kartograficzne posiadają cztery własności funkcjonalne, które są relewantne dla ich statusu jako reprezentacji[4]. Są one (1) reprezentacjami strukturalnymi, które (2) przewodzą

[4] Co oczywiste, mapy mogą posiadać własności funkcjonalne, które *nie* są relewantne dla nich *jako reprezentacji*. Za pomocą niektórych map można odganiać muchy, zasłaniać się przed światłem słonecznym lub rozpalać ogniska – jednak żadna z tych własności funkcjonalnych nie przyczynia się do tego, że dana mapa jest *reprezentacją*. Moja argumentacja w tym artykule opiera się wskazaniu, że reprezentacje postulowane w ramach TKP dzielą z mapami te własności, które „czynią" te ostatnie reprezentacjami. Reprezentacje postulowane przez TKP mogą nie przypominać – i prawdopodobnie nie będą przypominać – map kartograficznych pod żadnym innym względem.

działaniami swoich użytkowników, są (3) odłączalne oraz które (4) pozwalają swoim użytkownikom na rozpoznanie błędu reprezentacyjnego. Omówię teraz pokrótce każdą z tych czterech własności.

Mapa stanowi jest formą reprezentacji opartej na podobieństwie – konkretnie, na podobieństwie *strukturalnym*. Jest ona zatem reprezentacją strukturalną, czy też tak zwaną S-reprezentacją (por. Cummins 1989; O'Brien, Opie 2004; Ramsey 2007; Shea 2014; Swoyer 1991). Oznacza to, że relacja pomiędzy samą mapą a terenem przez nią reprezentowanym to relacja podobieństwa strukturalnego. Gerard O'Brien i Brian Opie (2004, s. 11) w następujący sposób definiują podobieństwo strukturalne:

> 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 [*structural* – PG] 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.

W przypadku mapy kartograficznej, układ (przynajmniej niektórych) relacji przestrzennych (metrycznych czy topograficznych) zachodzących pomiędzy (przynajmniej niektórymi) elementami samej mapy odzwierciedla zatem układ odpowiadających relacji przestrzennych zachodzących pomiędzy odpowiadającymi elementami reprezentowanego terenu. Załóżmy na przykład, że pewnym budynkom A, B i C odpowiadają na mapie, kolejno, punkty A', B' i C'. Zakładając, że mapa adekwatnie odzwierciedla teren, możemy na przykład orzec, że jeśli punkt A' znajduje się *bliżej* punktu B' niż punktu C', to budynek A znajduje się bliżej budynku B niż C; jeśli A' znajduje się *pomiędzy* B' a C', to budynek A znajduje się pomiędzy budynkami B a C; i tak dalej (por. O'Brien, Opie 2004).

Ktoś mógłby mi zarzucić, że zamierzałem tu scharakteryzować mapę czysto funkcjonalnie, tymczasem podobieństwo strukturalne, choć jest własnością relacyjną mapy, to nie jest jej własnością *funkcjonalną*. To napięcie jest jednak pozorne, ponieważ mówiąc o podobieństwie strukturalnym jako relacji łączącej mapę z terenem, nie mam na myśli po prostu relacji, która *zachodzi* (lub nie) pomiędzy nimi, ale relację, która jest dodatkowo

relewantna dla tego, czy mapa poprawnie spełnia swoją funkcję. Podobieństwo strukturalne jest tu relacją „wykorzystywalną" [*exploitable relation*] dla użytkownika mapy (por. Shea 2007, 2013, 2014). Od tego, czy ono zachodzi (albo od stopnia, w jakim ono zachodzi) zależy to, czy mapa poprawnie realizuje swoją funkcję reprezentacji (albo jak dobrze ją realizuje).

Aby zrozumieć, co mam na myśli przypisując taką rolę podobieństwu strukturalnemu, musimy najpierw zwrócić uwagę na drugą wymienioną własność funkcjonalną mapy *qua* reprezentacji, czyli na fakt, że służy ona temu, by *przewodzić działaniami* jej użytkownika (por. Anderson, Rosenberg 2008; Bickhard 1999, 2004). Mapa stanowi dla swojego użytkownika poznawczy „surogat" reprezentowanego przez nią terenu, pozwalając użytkownikowi odpowiednio „nawigować" działaniami podejmowanymi względem czy w ramach tego terenu. Kiedy wykorzystujemy mapę kartograficzną wędrując po nieznanym mieście, mapa ta pozwala nam kierować naszymi bezpośrednimi interakcjami z tym miastem, to jest podejmować decyzje dotyczące tego, *jak działać*: jaką ścieżkę obrać, w którą stronę i kiedy skręcić, których miejsc i jak unikać, i tak dalej. Mapa może też kierować naszymi czysto *poznawczymi* działaniami, pozwalając nam na przykład formować sądy o odległościach pomiędzy elementami terenu albo planować różne możliwe trasy w sposób czysto kontrfaktyczny (podróżować „palcem po mapie").

Kiedy piszę o podobieństwie strukturalnym jako relacji relewantnej dla funkcjonowania mapy, mam na myśli to, że podobieństwo to jest *wykorzystywane* w celu przewodzenia działaniem. Nawigując, podejmując decyzje czy formując sądy na podstawie mapy, wykorzystujemy fakt, że pomiędzy samą mapą a reprezentowanym przez nią terenem zachodzi podobieństwo strukturalne. *Sukces* naszych działań w zupełnie nieakcydentalny sposób *zależy* od tego, czy ta relacja zachodzi – albo od stopnia, w jakim ona zachodzi[5]. Mapy wiernie odzwierciedlające strukturalnie dany teren pozwalają nam docierać do celu, unikać

[5] Nie znaczy to, że dobra mapa to zawsze (a nawet zazwyczaj) mapa idealnie odzwierciedlająca dany teren w całej jego strukturalnej złożoności (por. Borges 1998). Czasem dobre mapy pełne są uproszczeń, przeinaczeń i idealizacji. To w jakich aspektach i w jakim zakresie mapa powinna przypominać reprezentowany teren, zależy od jej zastosowania; to znaczy od tego, *jakiego rodzaju działaniem ma ona przewodzić*. Inżynier planujący budowę osiedla będzie potrzebował mapy miasta dużo dokładniejszej niż ta, która potrzeba jest turyście z jego praktycznymi interesami. Mapa przeznaczona dla przewodzenia działaniami poznawczymi inżyniera będzie pokazywać szczegóły strukturalne, które są nieobecne na mapie turystycznej (pozwalając na rozróżnienie relacji metrycznych i topograficznych, których nie da się odczytać z tej drugiej mapy). Taka mapa będzie jednak „lepsza" jako reprezentacja tylko relatywnie do praktycznych interesów inżyniera, ponieważ dla turysty jej szczegóły stanowić mogą utrudniający użytkowanie szum.

przeszkód i formować prawdziwe sądy na temat potencjalnych ścieżek lub relacji przestrzennych pomiędzy elementami reprezentowanego terenu.

Trzecia własność funkcjonalna mapy kartograficznej polega na tym, że może ona realizować swoją rolę „przewodnika działań" nawet wtedy, gdy proces jej wykorzystania jest *częściowo* lub *całkowicie odłączony* od reprezentowanego przez nią terenu, to znaczy gdy jest ona wykorzystywana pod nieobecność tego, co reprezentowane (por. Clark, Grush 1999; Grush 1997; Haugeland 1991). Kiedy przemierzamy obce miasto wykorzystując elektroniczną mapę wyposażoną w system GPS, mamy zdolność mentalnego wybiegania „wprzód" w czasie i antycypowania zakrętów czy przeszkód, które są na razie niewidoczne w terenie, lecz widnieją na mapie i na które niedługo – na przykład, w odstępie kilku sekund – rzeczywiście natrafimy. Nasze bieżące decyzje nawigacyjne podejmowane są na podstawie mapy, a następnie wprowadzane w życie. Choć mapa w takiej sytuacji pozwala nam orientować się względem nieobecnych dla nas w danym momencie elementów terenu, to realizuje ona swoją funkcję w trakcie, gdy nawigujemy swoje działania w ramach reprezentowanego przez nią terenu. Jest ona odłączona, lecz jest to odłączenie *częściowe*, ponieważ proces wykorzystania mapy dokonuje się w trakcie bezpośrednich interakcji z przemierzanym terenem. Mapy kartograficzne mogą jednak funkcjonować także w sposób *całkowicie* odłączony, to znaczy zupełnie poza kontekstem bezpośrednich praktycznych interakcji z reprezentowanymi przez nie terenami. Siedząc w fotelu znajdującym się w warszawskim budynku, możemy sięgnąć po mapę kartograficzną Tokio. Oczywiście w takiej sytuacji mapa nie może przewodzić naszymi podróżami po Tokio. Może ona jednak przewodzić naszymi działaniami *poznawczymi*. Za jej pomocą możemy dowiedzieć się, jaka jest najkrótsza możliwa droga z jednego punktu do drugiego, jakie są odległości pomiędzy określonymi dzielnicami, jak zmienia się ukształtowanie terenu z południa na północ miasta, i tak dalej. Dzięki temu mapa może też pozwolić nam na planowanie *przyszłych* interakcji (przyszłych tras), które będziemy mogli przemierzyć kiedy (jeśli) pewnego dnia znajdziemy się w Tokio.

Ostatnią wreszcie własnością mapy jako reprezentacji jest fakt, że pozwala nam ona na *rozpoznanie błędu reprezentacyjnego*. Mapy, co oczywiste, bywają błędne. Są one błędne wtedy, gdy na tyle odbiegają one swoją organizacją przestrzenną od organizacji przestrzennej reprezentowanego terenu, że nie pozwalają one na sprawne kierowanie działaniem w tym terenie. Tu chodzi mi jednak o to, że mapy jako reprezentacje pozwalają nam, ich użytkownikom, na *rozpoznanie* błędu reprezentacyjnego. Zauważmy, że w zdecydowanej większości sytuacji nie dysponujemy możliwością *bezpośredniej* weryfikacji, czy albo na ile

poprawna jest mapa, którą się posługujemy. Zazwyczaj nie mamy do dyspozycji helikoptera, który pozwoliłby nam wzbić się w powietrze i z takiej perspektywy porównać strukturę mapy ze strukturą terenu. Nie znaczy to jednak, że nie możemy rozpoznawać błędu reprezentacyjnego. Możliwość ta jest nam dostępna *pośrednio*, dzięki temu, jak nasza *praktyczna* relacja ze światem zależna jest od wierności mapy względem terenu, którego jest to mapa (por. Anderson, Rosenberg 2008; Bickhard 1999, 2004). Mówiąc luźno, poprawność mapy możemy pośrednio oszacować na podstawie liczby guzów, jakie sobie nabijamy poruszając się po świecie za pomocą tej mapy. Błędną mapę poznajemy po tym, jak zawodzą nasze bezpośrednie, praktyczne interakcje z teren reprezentowanym przez tę mapę. Mapy błędne utrudniają czy uniemożliwiają nam dotarcie do celu, nadmiernie wydłużają nasze podróże, czy, mówiąc ogólnie, gubią nas i sprowadzają na manowce. Choć fakty dotyczące strukturalnego podobieństwa pomiędzy terenem a mapą są nam na ogół poznawczo niedostępne, to fakty dotyczące niepowodzeń naszych działań już są. Błąd reprezentacyjny poznajemy nie poprzez porównanie mapy z terenem, lecz na podstawie niepowodzenia działań, którymi ta reprezentacja przewodzi.

Zauważmy, że w przypadku map możliwe są *dwa* różne rodzaje błędu reprezentacyjnego. Pierwszy rodzaj błędu wynika z posługiwania się błędną mapą, która na tyle odbiega swoją organizacją strukturalną od reprezentowanego terenu, że uniemożliwia nam ona sprawną nawigację po tym terenie. Drugi rodzaj błędu wynika z błędnej *aplikacji* mapy. Chodzi tu o sytuację, kiedy co prawda dysponujemy mapę wiernie odzwierciedlającą teren, lecz niepoprawnie sytuujemy *nasze własne położenie* w tym terenie, co uniemożliwia nam sprawne działanie i realizowanie naszych praktycznych celów. Z taką sytuacją mielibyśmy do czynienia, gdybyśmy swoje działania opierali na fragmencie (skądinąd poprawnej) mapy Warszawy reprezentującym Mokotów, nie zdając sobie sprawy, że tak naprawdę chodzimy po Ochocie. Oba rodzaje błędu rozpoznajemy po niepowodzeniu naszych działań. Sam fakt, że mapa zawodzi nas jako przewodnik działania, nie pozwala jednoznacznie rozstrzygnąć, który z tych błędów popełniliśmy – jest to problem interpretacyjny, z którym musimy się zmagać jako użytkownicy map[6].

[6] Należy tu dodać dwa komentarze. Po pierwsze, wydaje się, że współczesna kartografia jest dobrze rozwiniętą dziedziną, zatem większość map, jakimi się posługujemy, to mapy wiernie odzwierciedlające reprezentowany teren. (Wyjątkiem są tutaj mapy GPS, które czasem nie są uaktualniane na tyle często, by „nadążyć" za zmianami organizacji ruchu drogowego). Większość błędów reprezentacyjnych w przypadku map to zatem błędy wynikające z aplikacji. Po drugie, wydaje się, że oba wymienione rodzaje błędu można rozróżniać za pomocą prostej heurystyki. Po pierwszych niepowodzeniach naszych działań zaczynamy od założenia, że nasz

4. Reprezentacje w TKP jako przewodzące działaniem, odłączalne modele strukturalne pozwalające na rozpoznanie błędu reprezentacyjnego

4.1. Uwagi wstępne

Chcę teraz rozwinąć i obronić tezę, że TKP postuluje struktury wewnętrzne, które w zupełnie nietrywialnym stopniu przypominają funkcjonalnie mapy kartograficzne. Można powiedzieć, że jeśli moja interpretacja reprezentacjonistycznych założeń jest TKP poprawna, to z koncepcji kodowania predykcyjnego wynika, iż systemy poznawcze nawigują swoimi działaniami wykorzystując wewnętrzne przyczynowo-probabilistyczne „mapy" świata. Sądzę, że to właśnie owe wewnętrzne mapy powinniśmy utożsamić z „reprezentacjami" postulowanymi w TKP. Rolę takich wewnętrznych map odgrywają zaś w TKP *modele generatywne*. To właśnie modele generatywne są, podobnie jak mapy kartograficzne, *przewodzącymi działaniem, odłączalnymi reprezentacjami strukturalnymi, pozwalającymi na rozpoznanie błędu reprezentacyjnego*.

Jak sądzę, takie odczytanie natury „reprezentacji" postulowanych w TKP jednocześnie rozstrzyga także kwestię tego, czy zawarte w tej teorii odwołania do reprezentacji są prawomocne biorąc pod uwagę Ramseyowski wymóg opisu zadań. Zgodnie z moją propozycją, modele generatywne posiadają cztery wymienione i opisane wcześniej własności funkcjonalne charakteryzujące mapy kartograficzne *qua* reprezentacje. Oznacza to, że sposób funkcjonowania modeli generatywnych bardzo przypomina sposób funkcjonowania map kartograficznych jako prototypowych reprezentacji. Owa bliskość do prototypu sprawia, że modele generatywne spełniają wymóg opisu zadań i zasługują na miano reprezentacji. TKP posługuje się więc „pełnokrwistym", eksplanacyjnie wartościowym pojęciem reprezentacji mentalnych (por. Gładziejewski, w druku-b).

---

błąd wynika z błędnej aplikacji mapy. Zmieniamy wtedy hipotezę dotyczącą naszego realnego położenia w ramach przemierzanego terenu i zaczynamy działać na podstawie tej hipotezy. Po kilku niepowodzeniach tej strategii dochodzimy do wniosku, że nasze błędy wynikają z błędnej *mapy*, a nie z błędu *aplikacji* mapy do terenu. Trzeba więc zmienić *mapę*, a nie *sposób użycia* mapy, którą już dysponujemy. Skuteczność takiej heurystyki wynika z dość oczywistego faktu. Kiedy posiadamy niepoprawną mapę, nasze działania będą zawodzić *systematycznie*; kiedy nasze niepowodzenia wynikają z błędu aplikacji, z czasem powinniśmy „odnaleźć się" i zacząć odnosić sukcesy. (Dziękuję tu Przemkowi Nowakowskiemu ze cenne uwagi dotyczące rozpoznawania źródeł błędu reprezentacyjnego przy używaniu map).

Chcę teraz kolejno skupić się na każdej z wymienionych wcześniej własności funkcjonalnych map kartograficznych *qua* reprezentacji, wyjaśniając, dlaczego i w jakim sensie modele generatywne również egzemplifikują tę własność. Zanim do tego przejdę, warto poczynić trzy wstępne uwagi, które, mam nadzieję, rozjaśnią moje stanowisko.

Po pierwsze, choć sądzę, że modele generatywne egzemplifikuję cztery własności funkcjonalne map kartograficznych, nie mam przez to na myśli, że te drugie funkcjonują *dokładnie* tak samo, jak te pierwsze. Nie chcę oczywiście argumentować, że modele generatywne są *dosłownie* mapami. Oprócz ogólnych podobieństw, występują tu też rozbieżności dotyczące detali. Na przykład, mapy to artefakty kulturowe, które odgrywają funkcję przewodników działań dzięki temu, że są interpretowane i wykorzystywane w takim celu przez ludzi, czyli pełnoprawne podmioty intencjonalne. Pod groźbą popełnienia błędu homunkularnego, nie możemy oczywiście stwierdzić, że modele generatywne swoją funkcję przewodzenia działaniem również zawdzięczają temu, że są przez „kogoś" interpretowane. Ich funkcja musi być determinowana w inny sposób (por. przyp. 9). By wziąć inny przykład, choć z proponowanej tu perspektywy zarówno o mapach kartograficznych, jak i modelach generatywnych można powiedzieć, że reprezentują na podstawie relacji podobieństwa strukturalnego, to zachodzą tu oczywiste różnice dotyczące tego, jak w obu przypadkach rozumiane są argumenty tej relacji. W przeciwieństwie do map, modele generatywne z pewnością *nie* odzwierciedlają swoją strukturą *przestrzenną* struktury *przestrzennej* jakiegoś terenu.

Po drugie, trzeba pamiętać, że sama TKP na obecnym etapie nie jest teorią kompletną. Na przykład kwestia tego, jak modele generatywne są kodowane w ośrodkowym układzie nerwowym pozostaje obecnie w znacznym stopniu w obszarze spekulacji (Pouget, Beck, Ji Ma, Latham 2013). Ponadto nie dysponujemy obecnie z pewnością danymi, które by pozwalały przesądzać o prawdziwości TKP (Clark 2013b). Niekompletność naszej wiedzy dotyczącej kodowania predykcyjnego oznacza zarazem, że nasza wiedza o tym, jak działają zgodnie z TKP *reprezentacje* jest również niekompletna. Bronione tu reprezentacjonistyczne odczytanie koncepcji kodowania predykcyjnego siłą rzeczy „odziedziczy" braki tej koncepcji. Moim celem nie jest jednak uzupełnienie tej koncepcji – to jest dodanie do teorii nowych empirycznie weryfikowalnych twierdzeń – ani podanie nowych argumentów na rzecz tej teorii. Chodzi mi raczej o *rozjaśnienie* niektórych filozoficznych założeń i konsekwencji TKP. Chcę tu nie tyle rozwinąć samą koncepcję kodowania predykcyjnego, co raczej pomóc w lepszym *zrozumieniu* tego, co ma ona do powiedzenia na temat relacji umysłu do świata.

Po trzecie, chcę podkreślić, że dla mojej argumentacji zasadnicze znaczenie ma fakt, iż modele generatywne postulowane przez TKP posiadają *wszystkie cztery* wymienione wcześniej własności. Gdyby argumentacja za reprezentacyjnym statusem TKP odwoływała się do jednej z tych własności (lub nawet do kombinacji *niektórych* z nich), bronione tu stanowisko byłoby potencjalnie otwarte na kontrprzykłady. Jak przekonują niektórzy autorzy, do bycia reprezentacją z pewnością nie wystarcza fakt, że poprawne funkcjonowanie czegoś zależy od relacji podobieństwa strukturalnego z czymś innym; nie uznalibyśmy, że klucze są reprezentacjami, ponieważ sukces w posługiwaniu się nimi do otwarcia drzwi zależy od tego, czy zachodzi izomorfizm pomiędzy kształtem klucza i zamka (Tonneau 2012). Analogicznie, jak pokazują niektórzy zwolennicy podejścia dynamicznego w kognitywistyce, sam fakt, że jakaś wewnętrzna, poznawcza struktura może funkcjonować w oderwaniu od określonych czynników środowiskowych, nie oznacza jeszcze, że można w eksplanacyjnie wartościowy sposób przypisać jej funkcję polegającą na reprezentowaniu czegoś (Chemero 2009). Zgodnie z moim stanowiskiem, o reprezentacyjnym statusie modeli generatywnych decyduje jednak *nie* fakt, że posiadają one *jedną* czy *niektóre* własności funkcjonalne map kartograficznych *qua* reprezentacji, ale fakt, że egzemplifikują one *wszystkie cztery* własności. Sądzę, że w TKP zawarte jest pojęcie reprezentacji wystarczająco mocne, by nie podlegało ono różnym trywializującym argumentom czy kontrprzykładom.

Przejdę teraz do wyjaśnienia, w jakim sensie modele generatywne postulowane w TKP egzemplifikuję cztery omówione wcześniej „prototypowo" reprezentacyjne własności funkcjonalne.

4.2. Podobieństwo strukturalne

Aby pokazać, w jakim sensie modele generatywne są reprezentacjami strukturalnymi (S-reprezentacjami), posłużę się tu uproszczonym przykładem systemu poznawczego, który wyposażony jest w model dwupoziomowy[7]. System ten posiada określone sensorium i

[7] To znaczy, model ten będzie wyposażony tylko w dwie warstwy: warstwę ukrytych parametrów i warstwę sensoryczną. Ograniczenie się do dwóch poziomów to zabieg, który znacząco upraszcza mój wywód. Musimy jednak pamiętać, że w realnych systemach poznawczych modele generatywne posiadają więcej niż dwie warstwy. Każda warstwa odzwierciedla zależności przczynowo-probabilistyczne zachodzące na różnych porządkach czasowych. Z kolei rozróżnienie „warstwa ukryta/warstwa sensoryczna" jest zrelatywizowane do każdej pary warstw znajdujących się bezpośrednio nad/pod sobą (warstwa sensoryczna relatywnie do warstwy bezpośrednio wyższej funkcjonuje zarazem jako warstwa ukrytych parametrów relatywnie do warstwy

zamieszkuje świat wypełniony przedmiotami fizycznymi średniej wielkości, które wchodzą ze sobą w określone interakcje przyczynowe. Wzorce interakcji pomiędzy przedmiotami w świecie mają charakter probabilistyczny: to, czy X wpłynie przyczynowo na Y ma zawsze określone prawdopodobieństwo, zależne od kontekstu (nawet tygrysy pożerają ludzi tylko z pewnym prawdopodobieństwem). Te przyczynowe i probabilistyczne wzorce w świecie determinują wzorce statystyczne występujące w pobudzeniu sensorium systemu. Statystyka sygnału zmysłowego docierającego do systemu jest jednym bezpośrednio dostępnym temu systemowi „śladem" tego, co dzieje się w świecie. Przyjmujemy w duchu TKP, że zadaniem naszego systemu jest wykorzystanie wzorców występujących w sygnale zmysłowym aby wytworzyć na tej podstawie wewnętrzny model świata będącego przyczyną tego sygnału. To znaczy, jeśli zakładamy, że istnieje pewna funkcja z możliwych stanów świata do możliwych stanów sensorium systemu, to zgodnie z TKP zadaniem systemu poznawczego jest dokonanie wewnętrznej *inwersji* tej funkcji i stworzenie w ten sposób wewnętrznego „surogatu" zewnętrznego świata. Ów model-surogat jest generatywny w tym sensie, że nie jest po prostu statycznym „obrazem" relacji świat-zmysły, ale rodzajem *symulatora doświadczenia*, czy, jak to poetycko ujmuje Andy Clark (2013a, s. 476), „generatorem wirtualnej rzeczywistości". Model ten ustawicznie aktywuje sensorium w taki sposób, by przebieg takiej „wirtualnej" aktywności zmysłowej dynamicznie symulował przebieg tej aktywności kiedy jest ona wywoływana przez świat. Im lepsza symulacja, tym mniejszy błąd predykcyjny.

Postulowany w TKP model generatywny przypomina zaimplementowaną w mózgu hierarchiczną sieć Bayesowską, której struktura przynajmniej w jakimś zakresie *odzwierciedla* strukturę świata zewnętrznego i, zarazem, strukturę zależności świat-sensorium (por. Pearl 2000). Wyższa warstwa modelu zawiera ukryte parametry, które odpowiadają ukrytym „za zasłoną zmysłów" przedmiotom fizycznym. Dolna warstwa sieci odpowiada sensorium systemu. Sieć taka odzwierciedla świat w trzech aspektach. Po pierwsze, dynamiczne wzorce interakcji pomiędzy ukrytymi parametrami wyższej warstwy mają odzwierciedlać wzorce zależności przyczynowo-probabilistycznych pomiędzy odpowiadającymi tym parametrom przedmiotami fizycznymi w świecie (między innymi wzorce zależności pomiędzy tygrysami i człekokształtnymi ssakami). Na poziomie ukrytych parametrów modelowana jest zatem przyczynowa struktura samego świata zewnętrznego (por. Friston, Kiebel 2009; Hohwy 2013). Po drugie, każdemu ukrytemu parametrowi

bezpośrednio niższej). Analogicznie, podobieństwo strukturalne pomiędzy modelem a światem nie będzie dwu- ale wielopoziomowe, to znaczy będzie obejmowało całą hierarchię zagnieżdżonych w sobie przyczynowo-probabilistycznych zależności zachodzących na różnych porządkach czasowych.

odpowiada dystrybucja prawdopodobieństw wywołania na niższym poziomie określonych obserwacji, to jest potencjalnych wzorców aktywności sensorium (por. Kemp, Tenenbaum 2008; Tenenbaum *et al.* 2011). Te dystrybucje prawdopodobieństw mają odpowiadać dystrybucjom prawdopodobieństw otrzymania określonych obserwacji w wyniku interakcji z określonymi, odpowiadającymi poszczególnym parametrom modelu przyczynami środowiskowymi (między innymi prawdopodobieństwom otrzymania określonych sygnałów sensorycznych w wyniku interakcji z tygrysami). Po trzecie wreszcie, skoro system ma implementować wnioskowanie Bayesowskie, ukryte parametry modelu muszą także odzwierciedlać prawdopodobieństwa wstępne charakteryzujące przyczyny środowiskowe (por. Clark 2013b; Friston, Kiebel 2009; Hohwy 2013), to znaczy prawdopodobieństwa wystąpienia tych przyczyn niezależnie od napływającego aktualnie sygnału sensorycznego (na przykład, prawdopodobieństwo wstępne natknięcia się na tygrysa).

Ujmując powyższe intuicyjne uwagi w nieco bardziej precyzyjny sposób, możemy powiedzieć, że pomiędzy wewnętrznymi modelami generatywnymi a przyczynowo-probabilistyczną strukturą świata zachodzi podobieństwo strukturalne[8] w trzech aspektach czy też na trzech poziomach:

a) Układ związków czy zależności dynamicznych pomiędzy ukrytymi parametrami modelu odzwierciedla układ zależności przyczynowo-probabilistycznych pomiędzy ukrytymi elementami środowiska. To znaczy, ukryte „po stronie świata" przyczyny sygnału są w modelu charakteryzowane poprzez ich wzajemne zależności przyczynowo-probabilistyczne. Bardziej technicznie:

   Zachodzi relacja podobieństwa strukturalnego pomiędzy (1) strukturą zależności dynamicznych (Z) pomiędzy ukrytymi parametrami modelu (Prm) a (2) strukturą powiązań przycznowo-probabilistycznych (Pp) pomiędzy ukrytymi przyczynami środowiskowymi (Prz). Inaczej mówiąc zachodzi podobieństwo strukturalne pomiędzy M

[8] Podobieństwo strukturalne rozumiem tu zgodnie z przytoczoną wcześniej definicją Opie i O'Briena (2004). Odwołuję się do tej relacji, ponieważ jest ona słabsza niż izomorfizm czy homorfizm. Podejrzewam, że charakteryzowanie związku pomiędzy modelem generatywnym a światem za pomocą tych ostatnich relacji byłoby zbyt restryktywne (podobnie zresztą jak charakteryzowanie w ten sposób relacji pomiędzy mapami i reprezentowanymi przez nie terenami, por. O'Brien, Opie 2004). Wydaje się, że nawet modele generatywne wyposażone w pewne uproszczenia i przeinaczenia – modele, które nie są *lustrami* świata i nie odzwierciedlają go w całej jego złożoności – mogą skutecznie minimalizować błąd predykcyjny, podobnie jak uproszczone i przeinaczone mapy mogą mimo to skutecznie przewodzić naszymi działaniami.

= (Prm, Z) a Św = (Prz, Pp). To znaczy, dla przynajmniej niektórych elementów Prm i przynajmniej niektórych elementów Z, istnieje funkcja wzajemnie jednoznaczna przypisująca elementy Prm elementom Prz i istnieje funkcja wzajemnie jednoznaczna przypisująca elementy Z elementom Pp, tak, że jeśli określona relacja z Z zachodzi pomiędzy określonymi elementami Prm, to odpowiadająca relacja z Pp zachodzi pomiędzy odpowiadającymi elementami Prz. Na przykład, jeśli X i Y są elementami Prz i P(X|Y) = 0.6, to pomiędzy odpowiadającymi X i Y elementami (parametrami) X' i Y' w Prm zachodzi określona, odpowiadająca prawdopodobieństwu warunkowemu 0.6 relacja Z_1 w Z.

b) Parametry modelu generatywnego są powiązane z potencjalnymi obserwacjami (wzorcami pobudzenia sensorium systemu, czyli sygnałami zmysłowymi), a układ tych związków odzwierciedla relacje przyczynowo-probabilistyczne pomiędzy przyczynami środowiskowymi a potencjalnymi obserwacjami. To znaczy, przyczyny środowiskowe są w modelu charakteryzowane poprzez dystrybucje prawdopodobieństw wywołania określonych obserwacji w systemie. Bardziej technicznie:
Zachodzi relacja podobieństwa strukturalnego pomiędzy (1) strukturą zależności (Z) pomiędzy parametrami modelu (Prm) a możliwymi obserwacjami (O) a (2) strukturą zależności przyczynowo-probabilistycznych (Pp) pomiędzy przyczynami środowiskowymi (Prz) a możliwymi obserwacjami (O). Inaczej mówiąc zachodzi podobieństwo strukturalne pomiędzy M = (Prm, O, Z) a Św = (Prz, O, Pp). To znaczy, dla przynajmniej niektórych elementów Prm, i przynajmniej niektórych elementów Z, istnieje funkcja wzajemnie jednoznaczna przypisująca elementy Prm elementom Prz, i istnieje funkcja wzajemnie jednoznaczna przypisująca elementy Z elementom Pp, tak, że jeśli określona relacja z Z zachodzi pomiędzy określonymi elementami Prm a określonymi elementami O, to odpowiadająca relacja z Pp zachodzi pomiędzy odpowiadającymi elementami Prz a tymi samymi elementami O. Na przykład, jeśli X jest elementem Prz, O_1 jest elementem O i $P(O_1|X) = 0.9$, to pomiędzy odpowiadającym X elementem X' w Prm a O_1 zachodzi należąca do Z i odpowiadająca prawdopodobieństwu warunkowemu 0.9 określona relacja Z_1.

c) Ukryte parametry modelu posiadają własności, które odzwierciedlają układ prawdopodobieństw wstępnych wystąpienia określonych przyczyn środowiskowych

(prawdopodobieństw wystąpienia tych przyczyn niezależnie od napływającego ze świata sygnału zmysłowego). Bardziej technicznie:

Zachodzi relacja podobieństwa strukturalnego pomiędzy (1) strukturą parametrów modelu (Prm) i własności (W) a (2) strukturą przyczyn środowiskowych (Prz) i ich prawdopodobieństw wstępnych (Pw). Inaczej mówiąc zachodzi podobieństwo strukturalne pomiędzy M = (Prm, W) a Św = (Prz, Pw). To znaczy, dla przynajmniej niektórych elementów Prm i przynajmniej niektórych elementów W, istnieje funkcja wzajemnie jednoznaczna przypisująca elementy Prm elementom Prz i istnieje funkcja wzajemnie jednoznaczna przypisująca elementy W elementom Pw, tak, że jeśli określony parametr z Prm posiada określoną własność z W, to odpowiadająca własność Pw zachodzi posiadana jest przez odpowiadający element Prz. Na przykład, jeśli X jest elementem Prz i P(X)=0.3 (to jest, prawdopodobieństwo wstępne X wynosi 0.3), to odpowiadający X element X' z Prm posiada odpowiadającą prawdopodobieństwu wstępnemu 0.3 własność W_1 z W.

Powyższej charakterystyce rzecz jasna daleko do pełnej precyzji i kompletności. Jaka *dokładnie* jest relewantna struktura reprezentowanego *świata* biorąc pod uwagę *rzeczywiste* systemy poznawcze? Które dokładnie elementy przyczynowej struktury świata mają swoje odpowiedniki w modelu? Jaka jest *dokładnie* relewantna struktura samej *reprezentacji*? Jak postulowane w TKP sieci probabilistyczne – to jest parametry modelu generatywnego oraz przypisane im dystrybucje prawdopodobieństw obserwacji i prawdopodobieństwa wstępne – są kodowane w tkance nerwowej? Te pytania dotyczące szczegółów domagają się rzecz jasna szczegółowych odpowiedzi i pozostaje mieć nadzieję, że odpowiedzi takie zostaną z czasem udzielone (w kwestii kodowania reprezentacji probabilistycznych w mózgu, por. Pouget *et al*. 2013). Tutaj jednak chciałem nie tyle rozszerzyć i uzupełnić TKP, co raczej wskazać i wyrazić określone zawarte w tej koncepcji *zobowiązanie* teoretyczne. Jak się wydaje, postulując modele generatywne, zwolennicy TKP postulują zarazem reprezentacje, które opierają się na wieloaspektowym odzwierciedlaniu struktury tego, co podlega reprezentowaniu. Jeśli TKP jest prawdziwa, to struktura zależności przyczynowo-probabilistycznych w świecie jest przynajmniej częściowo odzwierciedlana w strukturze i dynamice mózgu.

Zanim przejdę dalej, chcę uporać się z następującym problemem, na który potencjalnie natrafia odczytanie reprezentacji postulowanych w TKP jako S-reprezentacji. Sformułowania TKP odwołujące się do teorii informacji podkreślają znaczenie, jakie dla procesu

minimalizacji błędu predykcyjnego ma *informacja wzajemna* (Hohwy 2013). Oznacza to, że system tym lepiej minimalizuje błąd predykcyjny, im lepsza jest korelacja czy współzmienność pomiędzy parametrami modelu generującymi symulowane doświadczenie a środowiskowymi przyczynami napływającego sygnału zmysłowego. Na przykład, system powinien działać tak, by aktywować hipotezę (ukryty parametr modelu jej odpowiadający) o tygrysie jako przyczynie sygnału wtedy i tylko wtedy, gdy napływający do niego sygnał został wywołany przez tygrysa. Skoro tak, to czy nie powinniśmy uznać, że TKP postuluje reprezentacje detektorowe, to jest takie, w których relacja pomiędzy reprezentacją a tym, co reprezentowane, to relacja współzmienności, a nie relacja podobieństwa? Dlaczego powinniśmy preferować interpretację TKP odwołującą się do S-reprezentacji? To napięcie interpretacyjne zyskuje dodatkową wagę jeśli zwrócimy uwagę, że wspomniany Ramsey (2007) przytacza argumenty na rzecz tezy, iż reprezentacje rozumiane jako detektory nie spełniają wymogu opisu zadań. Jeśli okaże się, że TKP tak naprawdę postuluje nie wewnętrzne mapy, a wewnętrzne detektory, to jest jej reprezentacjonistyczny status może okazać się zagrożony.

Oczywiście można powiedzieć po prostu, że TKP postuluje zarówno reprezentacje receptorowe, jak i S-reprezentacje (por. Hohwy 2013), albo wręcz że te dwie opcje teoretyczne są w istocie nieodróżnialne (por. Morgan 2014). Sądzę jednak, że interpretacja TKP odwołująca się do S-reprezentacji ma określoną przewagę *eksplanacyjną* nad interpretacją odwołującą się do detektorów. Proces doboru hipotez/parametrów i generowania na tej podstawie „wirtualnego doświadczenia" kontrolowany jest wewnętrznie, stanowiąc efekt *endogenicznej* dynamiki systemu wykorzystującego modelu generatywny środowiska. Proces predykcji jest wynikiem wewnętrznej aktywności, a świat jedynie modyfikuje czy kalibruje ten wewnętrzny, odgórny proces, „motywując" system do rewizji hipotez jeśli błąd predykcyjny jest odpowiednio wysoki. Zauważmy, że model generatywny tym lepiej symuluje i przewiduje doświadczenie, a zatem tym lepiej minimalizuje błąd predykcyjny, im lepiej odzwierciedla on, na trzech wymienionych wcześniej poziomach, strukturę probabilistyczno-przyczynową świata. Zauważmy wreszcie – i to jest centralna obserwacja – że zachodzi *asymetria eksplanacyjna* pomiędzy detektorowymi a S-reprezentacyjnymi wyjaśnieniami tego, jak systemy poznawcze minimalizują błąd predykcyjny. Im lepiej model odzwierciedla środowisko, tym mniejszy błąd predykcyjny, czyli tym większa jest informacja wzajemna. Zatem to, czy zachodzi informacja wzajemna pomiędzy stanami systemu a stanami świata jest wyjaśniane przez to, jak dalece model odzwierciedla strukturalnie świat. Jednak wyjaśnianie nie przebiega tu w drugą stronę: współzmienność (informacja wzajemna)

pomiędzy parametrami modelu a stanami świata nie wyjaśnia, dlaczego zachodzi tu także podobieństwo strukturalne. Choć więc pomiędzy stanami systemu a stanami świata zachodzi zarówno informacja wzajemna, jak również podobieństwo strukturalne, to jednak ta ostatnia relacja jest tu eksplanacyjnie pierwotna względem tej pierwszej. Ostatecznie to podobieństwu przysługuje pierwszeństwo, jeśli chcemy wiedzieć, *jak* systemom poznawczym udaje się minimalizować błąd predykcyjny. Sądzę, że jest to wystarczający powód by stwierdzić, iż reprezentacje postulowane w ramach TKP są *pierwotnie* lub *przede wszystkim* S-reprezentacjami.

4.3. Przewodzenie działaniem

Podobnie jak to miało miejsce w mapach kartograficznych, w przypadku modeli generatywnych relacja podobieństwa strukturalnego jest relacją *wykorzystywalną* (Shea 2007, 2013, 2014). Znaczy to, że *poprawne funkcjonowanie* modelu generatywnego jako reprezentacji jest nieakcydentalnie zależne od tego, czy oraz w jakim stopniu odzwierciedla on przyczynowo-probabilistyczną strukturę świata. Podobnie, jak to było w przypadku map kartograficznych, funkcję realizowaną w ten sposób przez modele generatywne określić można jako *przewodzenie działaniem*[9]. Wielu autorów piszących o kodowaniu predykcyjnym

[9] Gdy charakteryzuje się modele generatywne funkcjonalnie, a zatem teleologicznie, naturalnie powstaje pytanie: skąd bierze się ta teleologia? Funkcja map kartograficznych jako przewodników naszych działań jest określana przez intencje twórców i transmitowana kulturowo. W przypadku modeli generatywnych potrzebujemy jakiejś *stricte* naturalistycznej koncepcji funkcji, która nie presuponuje kategorii intencjonalnych. Choć nie mam tu miejsca by bronić tej tezy, wydaje się, iż przypisanie modelom generatywnym funkcji przewodzenia działaniem jest całkowicie spójne z przynajmniej niektórymi współczesnymi naturalistycznymi teoriami funkcji. Po pierwsze, jest ono spójne z teoriami etiologicznymi odwołującymi się do historii danej cechy czy struktury (por. Millikan 1984; Wright 1973), ponieważ w TKP przyjmuje się, że modele generatywne oraz zawarte w nich wysokopoziomowe założenia uprzednie zostały wyselekcjonowane na drodze doboru naturalnego; to znaczy, modele te *istnieją dlatego*, że kiedyś pozwalały organizmom minimalizować błąd predykcyjny, a zatem unikać sytuacji zwiększających ich entropię termodynamiczną. Po drugie, broniona tu koncepcja jest spójna z teorią mówiącą, że funkcja danej struktury określana jest jej relewantnością dla pewnej innej zdolności lub tym, że struktura ta wchodzi w skład (poprawnego) wyjaśnienia tego, jak dany system realizuje określoną zdolność (por. Craver 2001; Cummins 1975). Można bowiem wskazać, że aktywność modelu generatywnego jest relewantna dla zdolności do minimalizacji błędu predykcyjnego; albo że (poprawne) wyjaśnienie tego, jak organizmy minimalizują błąd predykcyjny, odwołuje się do roli modeli generatywnych. Po trzecie, skoro TKP zakłada, że modele generatywne są narzędziem samoorganizacji – pozwalają one organizmom podtrzymywać się w istnieniu dzięki unikaniu równowagi termodynamicznej – to przypisanie im roli przewodników działań jest

przypisuje co prawda modelom funkcję minimalizacji błędu predykcyjnego. Pamiętajmy jednak, że zgodnie z przyjętą tu przeze mnie wersją TKP, minimalizacja błędu nie jest „autoteliczna". Systemy poznawcze nie minimalizują rozbieżności pomiędzy przewidywanymi a rzeczywistymi sygnałami sensorycznymi dla samej minimalizacji. Minimalizacja błędu predykcyjnego jest środkiem do realizacji innego celu, jakim jest takie sterowanie własną aktywnością w świecie, by unikać sytuacji cechujących się dużym zaskoczeniem relatywnie do danego fenotypu, to znaczy zbliżających dany system do równowagi termodynamicznej (Friston 2010, 2013; Friston, Kiebel 2009; Friston, Setphan 2007; Hohwy 2013).

Modele generatywne przewodzą zatem działaniem w następującym sensie. „Działanie" w relewantnym tu znaczeniu to inaczej wnioskowanie aktywne: takie praktyczne interweniowanie w świat, by uzgodnić napływające oddolnie sygnały sensoryczne z sygnałami przewidzianymi, a tym samym pozostawać w mało zaskakujących sytuacjach[10]. Jak pamiętamy, zgodnie z TKP wnioskowanie aktywne jest ściśle powiązane z wnioskowaniem percepcyjnym. System poznawczy ustawicznie przewiduje, jak będzie zmieniał się sygnał sensoryczny pod wpływem jego działań, a działając ustawicznie weryfikuje te predykcje; predykcje te zaś formowane są w oparciu o hipotezę przyjętą na podstawie wnioskowania percepcyjnego. Nie mamy tu jednak do czynienia tylko z zależnością przyczynową, ale też zależnością *funkcjonalną*. *Powodzenie* wnioskowania aktywnego – mierzone błędem predykcyjnym powstającym na skutek tego wnioskowania – jest zależne od tego, jak wiernie model generatywny, jakim posługuje się system generując hipotezę, odzwierciedla przyczynową strukturę świata. Jeśli system przyjął na podstawie modelu hipotezę, że właśnie wchodzi w interakcje z pluszową imitacją tygrysa, gdy w istocie stoi przed nim tygrys z krwi i kości – lub gdy zgodnie z jego modelem świata tygrysy nie uśmiercają znajdujących się przed nimi ludzi – to rośnie prawdopodobieństwo, że zaangażuje się on w działania drastycznie mało efektywnie zmniejszające błąd predykcyjny. Jak to

spójne z teoriami sytuującymi funkcje w tym, czy dane cechy lub struktury przyczyniają się do autonomii systemów biologicznych (por. Bickhard 2004; Christensen, Bickhard 2002).

[10] Jak zatem widać, nie chodzi tu o „działanie" w tym samym sensie, do którego odwołujemy się przypisując rolę przewodników „działań" mapom kartograficznym. W przypadku tych ostatnich posługiwaliśmy się bardziej „ludowym" pojęciem działania jako, mówiąc z grubsza, aktywności podejmowanej ze względu na *pragnienia* lub *intencje* podmiotu. W przeciwieństwie do takiego rozumienia działania, pojęcie działania jako wnioskowania aktywnego (endogenicznie kontrolowanej aktywności zmierzającej do minimalizacji błędu predykcyjnego) *nie* presuponuje kategorii intencjonalnych, takich, jak intencje czy pragnienia. To dobra wiadomość biorąc pod uwagę, że rozjaśniając reprezentacjonistyczne założenia TKP musimy unikać błędu homunkularnego.

ujmuje Hohwy (2013, s. 91, tłum. PG), jeśli system „angażuje się we wnioskowanie aktywne na podstawie błędnej koncepcji tego, jaki jest świat, w dłuższych odstępach czasu sprawia to, że maleje prawdopodobieństwo, iż organizm ten będzie znajdował się w mało zaskakujących sytuacjach". Im wierniej parametry modelu odzwierciedlają „sensoryczno-generatywną" strukturę odpowiadających im po stronie świata przyczyn napływającego sygnału, a także im wierniej odzwierciedlają one strukturę zależności zachodzących pomiędzy tymi przyczynami oraz ich prawdopodobieństwa wstępne, tym mniejszy błąd predykcyjny generować będą działania (wnioskowania aktywne) na tym modelu oparte; a co za tym idzie – tym bardziej te działania przyczyniać się będą do utrzymania systemu w stanie dalekim od równowagi termodynamicznej. Jest to analogiczne do tego, jak szanse powodzenia naszych działań systematycznie zależą od tego, jak wiernie mapa, na podstawie której nawigujemy swoimi działaniami, odzwierciedla strukturalnie dany teren[11].

4.4. Odłączalność

Podobnie jak mapy kartograficzne, modele generatywne mogą realizować swoją reprezentacyjną funkcję w sposób, który jest odłączony od tego, co przez te modele reprezentowane. To znaczy mogą one przewodzić działaniami nawet wtedy, gdy to, co przez nie reprezentowane, jest aktualnie nieobecne. Fakt ten wpisany jest w samo sedno TKP, która przecież zasadniczą funkcję mózgu sytuuje w *przewidywaniu przyszłości*. Modele generatywne ustawicznie „wybiegają w przyszłość", w tym sensie, że generowany przez nie

[11] Niektórzy współcześni autorzy przeciwstawiają sobie dwa sposoby myślenia o naturze reprezentacji wewnętrznych, mianowicie podejście odwołujące się do reprezentacji jako kodów oraz podejście odwołujące się do reprezentacji jako przewodników działań (Anderson, Rosenberg 2008; Bickhard 1999, 2004). Pierwsze podejście jest skupione na „wejściu" i wspiera się na założeniu, że reprezentacje są konstytuowane przez jakiegoś rodzaju relację zachodzącą pomiędzy samą reprezentacją a tym, co przez nią reprezentowane. Drugie podejście skupione jest na „wyjściu" i wspiera się na założeniu, że reprezentacje są konstytuowane przez ich rolę w regulowaniu działań czy interakcji ze światem. Choć nie chcę się tu wdawać w szczegóły tego złożonego problemu, wydaje się, że każde z tych podejść natrafia na znaczące problemy (por. Gładziejewski, w druku-a). Proponowane tu odczytanie natury reprezentacji w TKP wymyka się jednak tej opozycji, przynajmniej jeśli ją rozumiemy jako alternatywę rozłączną. Modele generatywne są przewodnikami działań, ale takimi, których sukces jako przewodników zależy od tego, jak wiernie odzwierciedlają one przyczynową/probabilistyczną strukturę świata. Dla ich statusu jako reprezentacji ważna jest zarówno relacja pomiędzy samym modelem a tym, co reprezentowane, jak również rola odgrywana w przewodzeniu działaniami. Analogicznie jest zresztą z mapami kartograficznymi, które przewodzą naszymi działaniami dzięki temu, że w określonym stopniu odzwierciedlają w sobie przestrzenne własności reprezentowanego terenu.

odgórnie sygnał sensoryczny *wyprzedza* sygnał zmysłowy dochodzący ze świata [potrzebny cytat]. Zgodnie z TKP, nasze działanie kontrolowane jest głównie *wewnętrznymi predykcjami*, a rola świata zewnętrznego polega na *korekcji* tego procesu wtedy, gdy okaże się, że wewnętrznie wygenerowane predykcje nie odpowiadają sygnałowi dochodzącemu ze świata. Kiedy przemierzamy określone otoczenie, wielopoziomowa strukturalna reprezentacja, jaką jest model generatywny, nieustannie wytwarza i wysyła w dół hierarchii predykcje dotyczące tego, jak będzie się zmieniał sygnał zmysłowy w zależności od podejmowanych przez nas działań. Proces ten można porównać do przemierzania miasta za pomocą elektronicznej mapy wyposażonej w GPS. Podobnie jak mapa w takiej sytuacji, model generatywny decyduje o decyzjach praktycznych i wybiega w przyszłość na relatywnie niewielkie interwały czasowe [cytat potrzebny]. Podobnie jednak jak mapa, model, choć kontroluje nasze działanie, to realizuje on tę funkcję w kontekście bieżących interakcji ze światem i podlegając ustawicznej korekcie na podstawie przebiegu tych interakcji. Można zatem powiedzieć, że model generatywny przewodzący naszym działaniem świecie jest co prawda odłączony, ale jedynie *słabo*.

Warto zapytać o to, czy taka analogia z mapami kartograficznymi idzie dalej. Czy modele generatywne mogą, podobnie jak mapy, działać w sposób *mocno* odłączony? Czy można pokazać, że reprezentacje, do których odwołuje się TKP, pozwalają systemom poznawczym „odkleić" się bieżących interakcji ze światem i reprezentować stany rzeczy ulokowane w odleglejszej przyszłości, w przeszłości lub zaledwie kontrfaktyczne? Biorąc pod uwagę obecny stan badań, odpowiadając na to pytanie nieuchronnie wchodzimy w obszar spekulacji. TKP w obecnej postaci jest koncepcją percepcji i opartej na percepcji regulacji działania; inne funkcje poznawcze, w tym funkcje realizowane całkowicie *off-line*, nie wchodzą w zakres zjawisk wyjaśnianych przez tę teorię. Sądzę jednak, że istnieją poszlaki pozwalające sądzić, że odpowiedź na postawione pytanie może okazać się twierdząca, a reprezentacje postulowane w ramach TKP mogą działać poza kontekstem bieżących interakcji ze światem.

Mam na myśli poszlaki dwojakiego rodzaju. Po pierwsze, wiemy, że pewne obszary mózgu normalnie czy pierwotnie zaangażowane w percepcję i regulację działania odgrywają również rolę w niektórych funkcjach poznawczych o charakterze *off-line*, takich, jak wyobraźnia (Moulton, Kosslyn 2009) i myślenie pojęciowe (Barsalou 2009). Po drugie, zaczynają pojawiać się propozycje teoretyczne, które każą rozumieć aktywność poznawczą *off-line* – w tym myślenie pojęciowe – jako rodzaj symulacji przeprowadzanych na wewnętrznych sieciach probabilistycznych (por. Chater, Oaksford 2013; Goodman,

Tenenbaum, Gerstenberg, w druku). Zwolennicy takich podejść starają się między innymi pokazać, jak symulacje przeprowadzane na takich sieciach mogą cechować się produktywnością, to znaczy, jak można za ich pomocą symulować nowe sytuacje (ciągi zależności przyczynowych), które nie odpowiadają żadnej przeszłej percepcji.

Możemy spekulować, że pomiędzy tymi dwoma „przesłankami" zachodzi następująca zależność. Otóż aktywność *off-line* systemów percepcyjnych i motorycznych w mózgu może stanowić rezultat symulacji czysto kontrfaktycznych zdarzeń/procesów, przeprowadzanych na wewnętrznej sieci probabilistycznej, jaką jest model generatywny. Taka symulacja może być wykonywana na wysokich poziomach modelu, jednak przynajmniej w niektórych okolicznościach wywoływać będzie ona kaskadę wirtualnych sygnałów sensorycznych spływających stosunkowo nisko w dół hierarchii (por. Clark 2013b). Jeśli taka interpretacja jest poprawna, to okazywałoby się, że systemy poznawcze wykorzystują modele generatywne poza kontekstem bieżących działań, analogicznie do tego, jak ludzie czasem wykorzystują mapy kartograficzne poza kontekstem bieżących „potrzeb" nawigacyjnych. Choć jesteśmy obecnie dalecy od potwierdzenia takiej hipotezy, to niewykluczone, że reprezentacje postulowane przez TKP mogą działać w sposób mocno odłączony, kierując – podobnie jak mapy – nie tylko działaniami praktycznymi, ale też aktywnościami *stricte* poznawczymi.

4.5. Rozpoznanie błędu reprezentacyjnego

Podobnie jak mapy kartograficzne, modele generatywne mogą błędnie reprezentować świat, pozwalając zarazem na *rozpoznanie* tego błędu. Rzecz jasna aby wyjaśnić, jak reprezentacje postulowane w TKP pozwalają na rozpoznanie błędu reprezentacyjnego, należy najpierw wyjaśnić, co ma się na myśli mówiąc w tym kontekście o błędnych reprezentacjach jako takich. Wyjaśnienie błędu reprezentacyjnego to poważne filozoficzne wyzwanie. Na obecne potrzeby wskażę szkic rozwiązania, inspirowany rozważaniami Jacoba Hohwy'ego (2013). Otóż w kontekście bronionego tu odczytania TKP, błąd reprezentacyjny polega na tym, że (1) w danej sytuacji organizm angażuje się we wnioskowanie aktywne, które średnio i w większych odstępach czasowych[12] nie minimalizuje w tego rodzaju sytuacjach błędu

[12] Dodanie tu zastrzeżenia o „średnim" błędzie predykcyjnym jest teoretycznie znaczące. Jak zwraca uwagę Hohwy (2013), bywa tak, że niepoprawna hipoteza przyczynia się do minimalizacji błędu predykcyjnego. Może się na przykład zdarzyć, że hipoteza, iż przyczyną nadchodzącego sygnału sensorycznego jest obecność pluszowej imitacji tygrysa, sprawnie minimalizuje błąd predykcyjny pomimo tego, że przyczyną taką jest *de facto* obecność tygrysa z krwi i kości (może się tak zdarzyć na przykład wtedy, gdy nadchodzący sygnał jest

predykcyjnego (lub minimalizuje go nieefektywnie), (2) wynika to z tego, że na podstawie wnioskowania percepcyjnego dobrana została hipoteza (parametr modelu), której umiejscowienie w strukturze modelu generatywnego nie odpowiada umiejscowieniu faktycznej przyczyny nadchodzącego oddolnie sygnału zmysłowego w przyczynowo-probabilistycznej strukturze świata. Przykładowo: stojąc przez tygrysem, system obiera na podstawie wnioskowania percepcyjnego hipotezę, że stoi przed pluszową imitacją tygrysa oraz angażuje się na tej podstawie w określone działania; ponieważ przyjęta hipoteza nie odpowiada miejscu przyczyny środowiskowej w strukturze świata, działania te – na przykład, próba przytulenia rzekomego pluszaka – wywołują oddolny sygnał zmysłowy, który znacząco odbiega od sygnału przewidzianego (tym samym, model zawodzi jako przewodnik działań).

Zauważmy, że czynnik wymieniony wyżej jako (2), to jest niepoprawność wybranej hipotezy, może mieć dwa źródła. Z jednej strony, może ona wynikać z tego, że model generatywny, jakim dysponuje system poznawczy, odbiega swoją organizacją strukturalną od przyczynowej organizacji świata (odbiega na tyle, że w danych okolicznościach nie pozwala na skuteczną minimalizację błędu predykcyjnego). To znaczy, niepowodzenie minimalizacji błędu może wynikać z tego, że (a) dynamiczne zależności pomiędzy parametrami nie odzwierciedlają poprawnie przyczynowych zależności pomiędzy czynnikami środowiskowymi odpowiadającymi tym parametrom, lub (b) model przypisuje poszczególnym parametrom niepoprawną strukturę generatywną (dystrybucje prawdopodobieństw obserwacji), lub (c) model przypisuje poszczególnym parametrom niepoprawne prawdopodobieństwa wstępne. Jest to analogiczne do sytuacji, w której ktoś porusza się po danym terenie opierając się na niepoprawnej mapie. Z drugiej strony, system może dysponować co prawda poprawnym modelem, a jednak – na przykład, ze względu na fakt, że dochodzący do niego sygnał jest pełen szumów – aplikować w danych okolicznościach hipotezę, która nie odpowiada leżącej po stronie świata przyczynie sygnału sensorycznego. W takiej sytuacji, system dysponuje poprawną reprezentacją przyczynowo-probabilistycznej struktury świata, lecz z tego czy innego powodu, w danych okolicznościach

---

mocno zniekształcony albo gdy tygrys, w którego pobliżu się znaleźliśmy, ma nadwagę i akurat śpi). Właśnie dlatego Hohwy proponuje, że powinniśmy raczej rozważać *średni* błąd predykcyjny generowany przez niepoprawną hipotezę na przestrzeni większych odstępów czasowych (większej liczby prób). Choć może się jednorazowo zdarzyć, że kategoryzacja tygrysa jako pluszowej zabawki minimalizuje błąd predykcyjny bardzo sprawnie, to średni błąd predykcyjny wynikający z takiej błędnej kategoryzacji na przestrzeni większej liczby interakcji z tygrysami będzie skutkował znaczącym błędem predykcyjnym – z pewnością większym, niż wtedy, gdy tygrys jest każdorazowo kategoryzowany jako tygrys.

niepoprawnie sytuuje on *swoje własne położenie* w tej strukturze (tak właśnie jest wtedy, gdy ktoś skategoryzuje tygrysa jako tygrysa-pluszaka). Jest to analogiczne do sytuacji, w której ktoś dysponuje co prawda poprawną mapą, ale niepoprawnie sytuuje siebie w ramach danego terenu, przez co popełnia błąd wynikający z niepoprawnej aplikacji.

Mózgi czy systemy poznawcze rzecz jasna nie dysponują zdolnością do spojrzenia „z zewnątrz" celem weryfikacji, czy zachodzi relacja podobieństwa strukturalnego pomiędzy światem a modelami generatywnymi, za których pomocą systemy te się po świecie poruszają (por. Bickhard 1999, 2004). Mimo to sądzę, że reprezentacje postulowane przez TKP pozwalają na rozpoznanie błędu reprezentacyjnego. Podobnie jak to jest w przypadku map kartograficznych, w świetle TKP system poznawczy rozpoznaje błąd reprezentacyjny dzięki temu, że może wypróbowywać swoje reprezentacje *w działaniu*. System ustawicznie generuje hipotezy, które są aktywnie testowane w oparciu o wnioskowanie aktywne. Błąd reprezentacyjny rozpoznawany jest nie przez bezpośrednie porównanie modelu ze światem, ale pośrednio, na podstawie rozmiaru *błędu predykcyjnego* powstającego w wyniku *praktycznych interakcji* ze światem. Rozmiar błędu pośrednio świadczy o tym, jak bardzo aktywność systemu oddala się od realizacji zadania polegającego na unikaniu zaskakujących sytuacji; to znaczy, świadczy on o powodzeniu lub braku powodzenia wykonywanych działań. Biorąc pod uwagę, że działanie przewodzone jest modelem (czyli reprezentacją), jego niepowodzenie może zostać wykorzystane jako świadczące o tym, że model (czy przyjęta hipoteza) nie jest poprawny[13]. Jest to analogiczne do mierzenia poprawności mapy kartograficznej liczbą guzów, jakie sobie nabijamy wykorzystując tę mapę jako „przewodnika" naszych działań. Można więc powiedzieć, że podobnie jak to jest w przypadku map kartograficznych, na gruncie TKP błąd *reprezentacji* rozpoznaje się w oparciu o to, jak zawodzi *praktyczna* relacja ze światem.

Warto też zauważyć, że, podobnie jak to było w przypadku map kartograficznych, sam fakt, iż model zawodzi w swoim zadaniu przewodzenia działaniem (minimalizowania błędu predykcyjnego), nie pozwala sam w sobie definitywnie rozstrzygnąć pomiędzy dwoma

[13] Trzeba tu dodać, że zgodnie z TKP proces modyfikacji hipotez (rewizji parametrów generujących predykcje sensoryczne) na podstawie błędu predykcyjnego zależny jest od precyzji przypisywanej temu błędowi (por. Clark 2013b; Hohwy 2013). Mówiąc ogólnie, im mniejszą precyzję system przypisuje błędowi predykcyjnemu, tym większy musi być ten błąd, jeśli hipoteza ma zostać zmodyfikowana. Intuicyjnie mówiąc, szacowanie precyzji zależy od tego, jak „wiarygodny" jest dla systemu sygnał sensoryczny wywoływany przez środowisko. Im mniej wiarygodny ten sygnał, tym bardziej proces przyjmowania hipotez będzie zależał od predykcji modelu generatywnego i tym mniejszy będzie „korekcyjny" wpływ sygnału sensorycznego dochodzącego ze świata.

wymienionymi wcześniej interpretacjami. Niepowodzenie minimalizacji błędu predykcyjnego może wynikać z przyjęcia błędnej hipotezy dotyczącej przyczynowej etiologii sygnału zmysłowego (co odpowiada błędnej *aplikacji* mapy) lub z faktu, że system dysponuje modelem generatywnym, który niewystarczająco odzwierciedla strukturę przyczynowo-probabilistyczną środowiska (co odpowiada posługiwaniu się *błędną mapą*). Rozróżnienie pomiędzy tymi możliwościami jest problem interpretacyjnym, przed którym stoją organizmy poruszające się po świecie minimalizując błąd predykcyjny. Wydaje się też, że TKP posiada zasoby pojęciowe pozwalające odróżnić sytuację, w której system „przypisuje" niepowodzenie minimalizacji błędnej hipotezie od sytuacji, gdy niepowodzenie jest „interpretowane" jako wynikające z błędnego modelu. Pierwszej okoliczności odpowiada sytuacja, w której system poznawczy pod wpływem niepowodzenia działania zmienia hipotezę dotyczącą przyczyny środowiskowej napływającego sygnału. Drugiej opcji odpowiada sytuacja, w której system pod wpływem niepowodzenia działania modyfikuje sam model generatywny. Na przykład, wykorzystując metodę „empirycznego Bayesa", system może wykorzystywać błąd predykcyjny aby modyfikować parametry modelu generatywnego lub przypisywane im prawdopodobieństwa wstępne (Clark 2013b; Friston 2003).

5. Podsumowanie

Celem tego artykułu było przedstawienie propozycji dotyczącej tego, jak powinniśmy rozumieć naturę reprezentacji wewnętrznych w kontekście teorii kodowania predykcyjnego (TKP) oraz dokonanie oceny, czy reprezentacje w tym sensie spełniają Ramseyowski wymóg opisu zadań, a zatem czy na miano „reprezentacji" rzeczywiście zasługują. Argumentowałem, że reprezentacje postulowane przez TKP powinny być rozumiane jako pewna forma wewnętrznych, przewodzących działaniem, odłączalnych reprezentacji strukturalnych, które pozwalają na rozpoznanie błędu reprezentacyjnego. Profil funkcjonalny takich reprezentacji nietrywialnie przypomina profil funkcjonalny prototypowych reprezentacji, jakimi są mapy kartograficzne, dzięki czemu te pierwsze w pełni spełniają wymóg opisu zadań.

Jak wspomniałem na początku, historia kognitywistyki jest spleciona z historią reprezentacjonizmu. Wydaje się, że przyszły los filozoficznej idei, iż poznanie opiera się na wewnętrznym reprezentowaniu świata, zależy obecnie od losu pojęcia reprezentacji w staraniach kognitywistów zmierzających do dostarczenia *stricte* naturalistycznego wyjaśnienia naszych zdolności poznawczych. To znaczy, los ten zależy od tego, czy najlepsze istniejące koncepcje z zakresu kognitywistyki powołują się na struktury, które rzeczywiście

odgrywają rolę reprezentacji, a zatem od tego, czy teorie powołują się na „reprezentacje" w jakimś eksplanacyjnie wartościowym sensie.

Jeśli przeprowadzona w tym artykule analiza i ocena reprezentacjonistycznych założeń TKP jest poprawna, to ma ona konsekwencje, które powinny dodać optymizmu zwolennikom reprezentacjonizmu. Wydaje się, że jeśli mam rację, to TKP jest tak reprezentacjonistyczna, jak to tylko w kognitywistyce możliwe. Trudno mi wyobrazić sobie Ramseyowski w duchu argument na rzecz tezy, że zawarte w TKP pojęcie reprezentacji jest trywialne i że „reprezentacje" postulowane w ramach tej koncepcji nie zasługują na takie miano. Jednocześnie zwróćmy uwagę, że wielu autorów żywi nadzieję, iż TKP stanowi, o ile można tak górnolotnie powiedzieć, przyszłość kognitywistyki. Jeśli nadzieje te zostaną spełnione, teoria ta może zunifikować sporą część nauk kognitywnych i zdominować współczesną dyskusję na temat natury poznania – podobnie do tego, jak zdominował ją koneksjonizm na przełomie lat 80. i 90. poprzedniego wieku. Zakładając, że tezy wyrażone w tym artykule są poprawne, sytuacja taka wyraźnie przechyla szalę sporu reprezentacjonizm-antyreprezentacjonizm na rzecz tego pierwszego stanowiska, jednocześnie podważając stawianą czasem diagnozę mówiącą, że kognitywistyka współczesna zmierza w kierunku antyreprezentacjonizmu (por. Ramsey 2007). Ewentualna unifikacja kognitywistyki „pod banderą" kodowania predykcyjnego będzie zarazem unifikacją „pod banderą" reprezentacjonizmu.

Literatura

Anderson, M. L., & Rosenberg, G. (2008). Content and action: The guidance theory of representation. *Journal of Mind and Behavior*, 29: 55–86.

Barsalou, L. (2009). Simulation, situated conceptualization and prediction. *Philosophical Transactions of Royal Society B*, 364: 1281–1289.

Bickhard, M. H. (1999). Interaction and representation. *Theory and Psychology*, 9: 435–458.

Bickhard, M. H. (2004). The dynamic emergence of representation. In: H. Clapin, P. Staines, & P. Slezak. (Eds.). *Representation in Mind: New Approaches to Mental Representation* (pp. 71–90). Oxford: Elsevier Science.

Borges, J. L. (1998). O ścisłości w nauce. W: Tegoż. *Twórca*. Przeł. Z. Chądzyńska, K. Rodowska. Warszawa: Prószyński i S-ka.

Chater, N., & Oaksford, M. (2013). Programs as causal models: Speculations on mental programs and mental representation. *Cognitive Science*, 37: 1171–1191.

Chemero, A. (2009). *Radical Embodied Cognitive Science*. Cambridge (MA): The MIT Press.

Christensen, W. D., & Bickhard, M. H. (2002). The process dynamics of normative function. *The Monist*, 85: 3–28.

Clark, A. (2013a). Expecting the world: Perception, prediction and the origins of human knowledge. *The Journal of Philosophy*, : 469–496.

Clark, A. (2013b). Whatever next? Predictive brains, situated agents and the future of cognitive science. *Behavioral and Brain Sciences*, 36: 181–204.

Clark, A., & Grush, R. (1999). Towards a cognitive robotics. *Adaptive Behavior*, 7: 5–16.

Craver, C.F. (2001). Role functions, mechanisms, and hierarchy. *Philosophy of Science*, 68: 31–55.

Cummins, R. (1975). Functional analysis. *The Journal of Philosophy*, 72: 741–765.

Cummins, R. (1989). *Meaning and Mental Representation*. Cambridge (MA): The MIT Press.

Friston, K. J. (2003). Learning and inference in the brain. *Neural Networks*, 16: 1325–1352.

Friston, K. J. (2010). The free-energy principle: a unified brain theory? *Nature Neuroscience*, 11: 127–138.

Friston, K. J. (2013). Life as we know it. *Journal of the Royal Society Interface*, 10: 20130475.

Friston, K. J., & Kiebel, S. (2009). Predictive coding under the free-energy principle. *Philosophical Transactions of Royal Society B*, 364: 1211–1221.

Friston K. J., & Stephan, K. E. (2007). Free-energy and the brain. *Synthese*, 159: 417–458.

Gładziejewski, P. (w druku-a). Action guidance is not enough, representations need correspondence too: A plea for a two-factor theory of representation. *New Ideas in Psychology*.

Gładziejewski, P. (w druku-b). Explaining mental phenomena with internal representations. A mechanistic perspective. *Studies in Logic, Grammar and Rhetoric*.

Goodman, N. D., Tenenbaum, J. D., & Gerstenberg, T. (in press). Concepts in a probabilistic language of thought. In E. Margolis, S. Laurence (Eds.). *The Conceptual Mind: New Directions in the Study of Concepts*. The MIT Press.

Grush, R. (1997). The architecture of representation. *Philosophical Psychology*, 10: 5–23.

Haugeland, J. (1991). Representational genera. In: W. Ramsey, S. Stich, and D. Rumelhart (Eds.). *Philosophy and Connectionist Theory*. Hillsdale (N.J.): Erlbaum.

Hinton, G. E. (2007). Learning multiple layers of representation. *Trends in Cognitive Sciences*, 11: 428–434.

Hohwy, J. (2013). *The Predictive Mind*. Oxford: Oxford University Press.

Huang, G. T. (2008). Is this a unified brain theory? *New Scientist*, 2658: 30–33.

Huang, Y., & Rao, R. P. N. (2011). Predictive coding. *Wiley Interdisciplinary Reviews: Cognitive Science*, 2: 580–93.

Hutto, D. D., & Myin, E. (2013). *Radicalizing Enactivism: Basic Minds without Content*. Cambridge (MA): The MIT Press.

Kemp, C., & Tenenbaum, J. B. (2008). The discovery of structural form. *Proceedings of the National Academy of Sciences*, 105: 10687–10692.

Lee, T. S., & Mumford, D. (2003). Hierarchical Bayesian inference in the visual cortex. *Journal of Optical Society of America*, 20: 1434–1448.

Lettvin, J., Maturana, H., McCulloch, W., & Pitts, W. (1959). What the frog's eye tells the frog's brain. *Proceedings of the Institute of Radio Engineers*, 47: 1940–1951.

Millikan, R. G. (1984). *Language, Thought, and Other Biological Categories*. Cambridge (MA): The MIT Press.

Morgan, A. (2014). Representations gone mental. *Synthese*, 191: 213–244.

Moulton, S. T., Kosslyn, S. M. (2009). Imagining predictions: mental imagery as mental emulation. *Philosophical Transactions of the Royal Society B*, 364: 1273–1280.

O'Brien, G., & Opie, J. (2004). Notes toward a structuralist theory of mental representation. In H. Clapin, P. Staines, & P. Slezak. (Eds.). *Representation in Mind: New Approaches to Mental Representation* (pp. 1–20). Oxford: Elsevier Science.

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge: Cambridge University Press.

Pouget, A., Beck, J. M., Ji Ma, W., Latham, P. E. (2013). Probabilistic brains: knowns and unknowns. *Nature Neuroscience*, 16: 1170–1178.

Ramsey, W. (2007). *Representation Reconsidered*. Cambridge: Cambridge University Press.

Rao, R. P. N., & Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2: 79–87.

Shea, N. (2007). Consumers need information: supplementing teleosemantics with an input condition. *Philosophy and Phenomenological Research*, 75: 404–435.

Shea, N. (2013). Millikan's isomorphism requirement. In D. Ryder, J. Kingsbury, & K. Williford (Eds.). *Millikan and Her Critics* (pp. 63–86). Oxford: Wiley-Blackwell.

Shea, N. (2014). Exploitable isomorphism and structural representation. *Proceedings of the Aristotelian Society*, 64.

Swoyer, C. (1991). Structural representation and surrogative reasoning. *Synthese*, 87: 449–508.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure and abstraction. *Science*, 331: 1279–1285.

Tonneau, F. (2012). Metaphor and truth: A review of 'Representation Reconsidered' by W. M. Ramsey. *Behavior and Philosophy*, 39/40: 331–343.

Wright, L. (1973). Functions. *The Philosophical Review*, 82: 139–168.