UNIWERSYTET JAGIELLOŃSKI
WYDZIAŁ FILOZOFICZNY
Instytut Psychologii

Borysław Paulewicz

Interakcja i adaptacja

Propozycja metateoretyczna w badaniach nad zachowaniem

Praca doktorska
przygotowana pod kierunkiem
prof. Edwarda Nęcki

Kraków, 2009

Pomimo bardzo wielu badań, psychologom nie udało się jak dotąd ustalić jednej, powszechnie akceptowanej definicji tego zjawiska.

(niepublikowana praca magisterska)

Spis treści

Wprowadzenie

Testowanie wielu hipotez dotyczących przebiegu nie dających się bezpośrednio obserwować procesów przetwarzania informacji wydaje się niemożliwe, o ile warunki w jakich przeprowadzane jest badanie nie będą odpowiednio ograniczone, zmniejszając tym samym wpływ właściwej ludziom elastyczności reagowania na sytuacje. Przeważającą większość psychologów poznawczych da się z grubsza podzielić na zwolenników dwóch różnych strategii radzenia sobie z tym problemem.

Do jednego obozu zaliczyć można badaczy zajmujących się formułowaniem i testowaniem teorii dotyczących względnie wyizolowanych funkcji, zarówno tych określanych czasem jako niskopoziomowe, na przykład percepcji, uwagi selektywnej czy pamięci, jak i tych określanych czasem jako wysokopoziomowe, takich jak rozumowanie, kategoryzacja, podejmowanie decyzji, rozwiązywanie problemów, indukcja i wielu innych. Strategia ta motywowana jest zwykle złożonością przedmiotu badań. System poznawczy jako całość składa się przypuszczalnie z dużej liczby teoretycznie wyodrębnialnych części i nawet wykonanie względnie prostego zadania wymaga współdziałania wielu komponentów, wchodzących w trudne do przewidzenia interakcje. W konsekwencji kontrolowanie zmiennych zakłócających staje się zadaniem wyjątkowo dokuczliwym. Trudno się wobec tego dziwić, że wiele eksperymentów przeprowadza się w warunkach laboratoryjnych, przy użyciu stosunkowo prostych zadań, skonstruowanych tak, aby angażowały przede wszystkim interesujące badacza procesy.

Na skutek stosowania tej strategii powstało wiele imponujących lokalnych teorii i modeli. Jednocześnie jednak nie sposób nie dostrzec, że teorie, które powstały aby zrozumieć poszczególne części czy aspektów działania umysłu, trudno połączyć w większe całości. Wystarczy zapoznać się z częścią poświęconą na podsumowanie treści w zasadzie dowolnego rozdziału, dotyczącego aktualnego stanu wiedzy na temat niemal dowolnej funkcji czy własności poznawczej, żeby przekonać się, że część ta zawiera często niewiele ponad *wyliczenie* niektórych hipotez, badań i wyników, o stosunkowo szczegółowym charakterze. Zarazem nietrywialne i ogólne wnioski teoretyczne wydają się należeć do rzadkości.

Ta strategia i rezultaty jej stosowania zostały poddane ostrej krytyce. Większość powtarzanych do dzisiaj argumentów przeciwko badaniu procesów poznawczych we względnej izolacji można odnaleźć w wystąpieniu Newella z 1973 roku. Opublikowany tekst

tego wystąpienia doczekał się niezliczonej liczby cytowań, czyniąc swojego autora wpływowym metateoretykiem kognitywistyki w ogólności, a psychologii poznawczej w szczególności. Zdaniem Newella, niezadowalająca efektywność procesu kumulacji wiedzy w psychologii poznawczej może znacząco wzrosnąć, jeżeli tylko podjęte zostaną próby stworzenia teorii zintegrowanych. Ponieważ względne zalety i wady obu strategii stanowią jeden z ważniejszych tematów tej pracy, zdecydowałem się zacytować niektóre fragmenty tego wystąpienia.

> Zamierzałem narysować na tablicy kreskę, a następnie, wybierając losowo jedną z osób, które dzisiejszego dnia prezentowały swoje wyniki, zanotować nad kreską moment, w którym osoba ta obroniła pracę doktorską i aktualną datę (w środku kariery naukowej). Następnie, biorąc całkowitą liczbę artykułów takich jak te, przedstawione na odbywającym się właśnie sympozjum, zamierzałem obliczyć wydajność produkcyjną tej jakże imponującej pracy. Przechodząc w końcu do daty zakończenia kariery mojego wybranego autora, zamierzałem obliczyć całkowity przyszły dodatek podobnych publikacji aż do (przypuszczalnego) zakończenia kariery naukowej. W tym momencie miałem zamiar, przyjmując rolę dyskutanta, zadać pytanie: Przypuśćmy, że zebraliśmy już te wszystkie dodatkowe publikacje, takie jak te zaprezentowane dzisiaj (pomijając to, że dotyczyłyby one nowych aspektów problemu), *gdzie wtedy będzie się znajdowała psychologia?* Czy osiągniemy wówczas poziom nauki o człowieku adekwatny w swej wielkości i oddający sprawiedliwość złożoności swojego przedmiotu? A jeśli tak, w jakim stopniu stanie się to za sprawą artykułów, które właśnie Wam przedstawiłem? Czy może będziemy raczej zadawać kolejny zestaw pytań w następnym odcinku czasu?
>
> (s. 283-284, Newell, 1973)

Zdaniem Newella, odpowiedź na powyższe pytania brzmi:

> Kiedy zastanawiam się nad dalszym losem naszych dychotomii, przyglądając się tym, które istnieją, jako wskazówce na temat tego, w jaki sposób kształtują kierunek, w którym zmierza nauka, odnoszę wrażenie, że jasność nigdy nie jest osiągana. W miarę upływu czasu sprawy stają się po prostu coraz bardziej zawikłane i mętne. (...) ten rodzaj struktury pojęciowej prowadzi raczej do stale rosnącej sterty zagadnień, które albo stają się w końcu nużące, albo się od nich odwracamy, ale których nigdy naprawdę nie rozstrzygamy.
>
> (s. 288-289, tamże)

Sytuacja nie jest jednak beznadziejna:

> Możliwe jest skonstruowanie szczegółowych modeli struktury kontrolnej, połączonych z równie szczegółowymi założeniami na temat elementarnych procesów i zawartości pamięci. W ramach takiego systemu pytanie, jaką metodę zastosuje osoba badana wykonując zadanie eksperymentalne może być badane w ten sam sposób, w jaki odkrywa się w danym języku programowania algorytm, który ma wykonać określone zadanie. Tak samo jak w przypadku programowania, kilka różnych organizacji może pozwalać na adekwatne wykonania zadania. Jednakże każda taka metoda prowadzi do określonych predykcji dotyczących wymagań czasowych i przestrzennych, dostarczając podstaw dla działań eksperymentalnych, mających na celu ustalenie, jaka metoda została faktycznie zastosowana. (...) Istnieje olbrzymia przestrzeń możliwych struktur kontrolnych, a każda dostarcza szkieletu, w ramach którego niemal dowolna metoda może być zaprogramowana. (...) Niemniej, każda struktura kontrolna posiada inne właściwości związane ze sposobem kodowania, czasowymi charakterystykami przetwarzania i obciążeniem pamięci. Dostarczają one [te charakterystyki] podstawy dla eksperymentalnej identyfikacji systemu, jeżeli tylko poddany zostanie analizie wystarczająco obszerny i zróżnicowany zestaw zadań.
>
> (s. 302, tamże)

Nawiązując do popularnej gry, krytykowaną przez siebie strategię Newell określił jako strategię dwudziestu pytań, uznając za jej cechę charakterystyczną formułowanie problemów badawczych w postaci ogólnych dychotomii, takich jak szeregowość-równoległość, natura-środowisko, zanik śladu-interferencja, automatyczne-kontrolowane i tym podobnych. Współcześnie badanie zintegrowanych modeli umysłu polega na stosowaniu tak zwanych architektur poznawczych, czyli względnie wyczerpujących, obliczeniowych modeli działania umysłu (Anderson i Lebiere, 1998; Anderson, Bothell, Byrne i Douglass, 2004; Newell, 1990; Gray, 2007). O ile mi wiadomo, lista argumentów mających świadczyć o unikalnych zaletach strategii integracyjnej nie zmieniła się zasadniczo od roku 1973. Argumenty te dotyczą przede wszystkim konieczności eksplicitnego uwzględnienia w modelu mechanizmu kontroli, problemu elastyczności sposobu wykonania zadań (różne strategie rozwiązywania), a także związku między testowalnością teorii a zróżnicowanym charakterem badań i poziomem integracji modelu.

Oba podejścia posiadają wiele niezaprzeczalnych zalet i rezygnacja z jednego na rzecz drugiego byłaby nierozsądna. Niemniej, obie strategie w swoich typowych przejawach posiadają też kilka poważnych wad. Te wady, jak i możliwe sposoby ich usunięcia, są jednym z głównych tematów podejmowanych w tej pracy. Najważniejsze z nich to nieskuteczność stosowanych zwykle rozwiązań problemu identyfikacji architektury, nie-

uchronna i kłopotliwa złożoność modeli zintegrowanych, wreszcie ogromna czarna dziura, której zdają się nie dostrzegać członkowie obu obozów, to jest brak ogólnej teorii środowiska i celowej z nim interakcji.

Rozwiązaniem wymienionych problemów nie jest moim zdaniem połączenie obu strategii. Próby takiego łączenia faktycznie są podejmowane i doprowadziły do uzyskania wielu wartościowych rezultatów, jednak usunięcie wymienionych wad wymaga nie tyle *łączenia* modeli czy teorii w większe całości, ile raczej *ogólniejszego* spojrzenia na psychologię poznawczą i jej przedmiot.

Praca, którą Czytelnik ma w ręku, jest jedną długą i znacznie mniej niżbym sobie tego życzył uporządkowaną, systematyczną i wyczerpującą próbą uzasadnienia pewnego stanowiska metateoretycznego, dotyczącego właściwej postaci ogólnej psychologicznej teorii zachowania, procesów i stanów poznawczych. Wydaje mi się, że tego rodzaju teoria może, a z pewnych względów powinna, być racjonalną teorią celowej interakcji abstrakcyjnie rozumianego agenta z abstrakcyjnie rozumianym środowiskiem.

Żeby wyjaśnić, co przez to rozumiem i uczynić uzasadnienie tej tezy względnie przekonującym, byłem zmuszony podjąć wiele, czasami dość szczegółowych zagadnień o charakterze metodologicznym i filozoficznym. Pytania filozoficzne mają niestety to do siebie, że każda próba udzielenia odpowiedzi rodzi natychmiast jeszcze więcej niewygodnych pytań, a liczba trudnych do uzgodnienia, możliwych rozwiązań wydaje się być zbliżona do liczby podejmujących tego rodzaju problemy autorów. Aby poradzić sobie jakoś z tym kłopotliwym bogactwem postanowiłem poddać analizie z konieczności ograniczony zbiór teorii, badań, problemów i stanowisk. Sądzę, że zbiór ten jest wystarczająco reprezentatywny, aby można było oddalić zarzut, jakoby przedmiot krytyki miał być w istocie wytworem mojej fantazji.

Pierwszy rozdział zawiera omówienie znaczenia i historii rozwoju metodologii selektywnego wpływu i bardzo wybiórczy zarys historii badań dotyczących mechanizmu przeszukiwania pamięci krótkoterminowej. Zdecydowałem się na nieco szersze, choć uproszczone, przybliżenie teorii identyfikacji architektur Townsenda, rezultaty uzyskane przez tego autora są bowiem nadal często ignorowane, nawet przez tych badaczy, którzy je w swoich pracach cytują. W rozdziale drugim starałem się w taki sposób przedstawić modele kumulacji świadectw, aby Czytelnik mógł uchwycić zarazem ich prostotę jak i głębokie teoretyczne uzasadnienie, dlatego modele te omawiam jako konieczne uogólnienia teorii detekcji sygnałów. Przedmiotem rozważań w rozdziale trzecim jest przede wszystkim kwestia ilościowej oceny hipotez, rozumianej jako ocena trafności wynikających z nich predykcji. W rozdziale tym próbuję przekonać Czytelnika o ograniczonej użyteczności dostępnych metod oceny ilościowej, kładąc przy tym szczególny nacisk na problem złożoności modeli matematycznych. Korzystając z dogodnego dla tego celu kontekstu (testowanie istotności hipotezy zerowej) umieściłem w tym rozdziale wprowadzenie do wnioskowania bayesowskiego i bayesowskiej metody selekcji modelu, dzięki czemu mogłem w dalszych częściach pracy omówić bliżej wiele zagadnień, których bez pewnej orientacji we wnioskowaniu bayesowskim zrozumieć nie sposób. W rozdzia-

le czwartym zajmuję się jakościową oceną hipotez, to jest oceną ich mocy wyjaśniającej, przyglądając się przy tym krytycznie praktyce ogólnie rozumianej oceny i rewizji hipotez w psychologii. W rozdziale piątym opisałem, należące w moim odczuciu do najważniejszych dokonań we współczesnej psychologii poznawczej, teorie generalizacji Sheparda, Griffithsa i Tenenbauma i teorię oceny kauzalnej Griffithsa i Tenenbauma. Teorie te próbowałem zaprezentować w sposób umożliwiający uchwycenie unikalnych walorów podejścia racjonalnego, pozostawiając jednocześnie nieco mniej domyślności Czytelnika, niż tego wymaga lektura publikacji wymienionych autorów. Rozdział szósty zawiera rys historyczny rozwoju programu analizy racjonalnej. Sugeruję w nim możliwość zrewidowania niektórych podstawowych założeń, przyjmowanych przez zwolenników tego programu. W kolejnym, siódmym rozdziale podążam tropem intrygującej, nawet jeśli ryzykownej analogii między rolą, jaką miał początkowo ogrywać program analizy racjonalnej, a postulatami funkcjonalistów z przełomu dziewiętnastego i dwudziestego wieku. Zwracam uwagę na filozoficzne źródła i problematyczność sposobu myślenia, który zdaje się znajdować wyraz zarówno w praktyce badawczej współczesnych psychologów poznawczych jak i w dokonaniach strukturalistów. W ostatnim rozdziale omawiam dokładniej ramę pojęciową uczenia się ze wzmocnieniem i odkrytą dzięki niej teorię warunkowania klasycznego, która należy moim zdaniem do najważniejszych dokonań we współczesnych badaniach dotyczących uczenia się. Rozdział ten zatytułowałem „Funkcjonalizm racjonalny w praktyce" ponieważ to, w jaki sposób teoria warunkowania klasycznego wynika z ramy pojęciowej uczenia się ze wzmocnieniem i to, w jaki sposób ta rama pojęciowa pozwala dostrzec perspektywy i ograniczenia związane z zaproponowaną przez Suttona i Barto teorią warunkowania i jej formalizacją jest najlepszą jaką tylko potrafię sobie wyobrazić ilustracją konsekwentnego stosowania stanowiska metateoretycznego, będącego centralną konstruktywną propozycją tej pracy. Żeby ułatwić uchwycenie związków między rozmaitymi kwestiami, sześć pierwszych rozdziałów zakończyłem krótkimi podsumowaniami, zawierającymi niektóre ważniejsze wnioski, z których korzystam w późniejszych etapach rozumowania.

Już na wstępnym etapie prac stało się jasne, że nie uda się uniknąć wykorzystania formalnego języka matematyki, do czego początkowo nie byłem niestety odpowiednio przygotowany. Ponieważ jednak nie dowodzę żadnych nowych i nietrywialnych twierdzeń, ograniczyłem się do uproszczonego sposobu prezentacji, mając nadzieję, że nie wywoła to zbyt wielu nieporozumień. W stopniu, w jakim mi się to powiodło, tok rozumowania powinien być zrozumiały dla Czytelnika posiadającego elementarną znajomość rachunku prawdopodobieństwa i analizy matematycznej. Muszę w związku z tym wyrazić głęboką wdzięczność za życzliwe wsparcie i pomoc, jakiej doświadczyłem ze strony Jamesa Townsenda, Michaela Lee, Saula Sternberga, Francisa Tuerlinckxa, Joachima Vandekerckhove, Marca Buehnera, Johna Foxa, Jima Alberta, Thomasa Griffithsa, Richarda Goldena, Scotta Browna i Uffe Kjaerulffa, którzy z godną podziwu cierpliwością udzielali mi wyczerpujących wyjaśnień dotyczących spraw, z którymi sam nie mógłbym sobie zbyt szybko, o ile w ogóle, poradzić. Jeżeli ta praca zawiera jakieś poważne błędy, to

jest tak dlatego, że nie umiałem w wystarczającym stopniu skorzystać z tej nieocenionej pomocy.

Rozdział 1

Problem identyfikacji systemu poznawczego

Projektowanie i modyfikacja zadań stanowi niemal osobny przemysł w psychologii poznawczej. Można wręcz odnieść wrażenie, że przedmiotem zainteresowania badaczy są często nie tyle elementy strukturalne i funkcjonalne umysłu, ile raczej poszczególne typy zadań, służących do ich badania. Trudno się dziwić, że wiele takich paradygmatów, jak zwykło się nazywać procedury eksperymentalne, na przykład zadanie Stroopa (Stroop, 1935), Sternberga (Sternberg, 1966) i inne, doczekało się osobnych, obszernych opracowań (MacLeod, 1991; Glass, 1984).

Rozdział ten zawiera szczegółowe omówienie logiki badania za pomocą klasycznego zadania na przeszukiwanie pamięci krótkoterminowej Sternberga (Sternberg, 1966, 1969b). Przedmiotem analizy są procedury eksperymentalne, wyniki i wyjaśnienia przedstawione przez Sternberga w artykule przeglądowym z 1969 roku, a także jeden eksperyment przeprowadzony w roku 2004 przez Townsenda i Fifica. W kontekście badania pamięci krótkoterminowej pojawia się między innymi problem rozstrzygnięcia między szeregowością i równoległością przetwarzania, wielokrotnie przytaczany przez zwolenników modeli zintegrowanych jako ważny argument na rzecz stosowania takich właśnie modeli (Newell, 1990, 1973; Anderson i in., 2004; Anderson i Lebiere, 1998; Anderson, 1988/1991c; Byrne, 2007; Gray, 2007). Posługując się tym przykładem mam nadzieję nasycić konkretną treścią rozważania o bardziej ogólnym charakterze, dotyczące problemu identyfikacji systemu poznawczego. Przykład ten wybrałem dlatego, że jest on pod wieloma interesującymi mnie względami reprezentatywny dla niebagatelnej liczby badań eksperymentalnych w psychologii poznawczej.

1.1 Sternberga zadanie na przeszukiwanie pamięci krótkoterminowej

Przebieg typowego badania za pomocą komputerowej wersji zadania Sternberga wygląda następująco: osoba badana ma wykonać serię prób, z których każda polega na udzieleniu

odpowiedzi na temat tego, czy jakiś element był, czy nie był prezentowany wcześniej w trakcie danej próby. Na początku każdej próby pojawia się punkt fiksacji, czyli symbol, na którym osoba badana ma skupić wzrok, w oczekiwaniu na mające się później pojawić bodźce. Po zniknięciu punktu fiksacji kolejno prezentowane są elementy zestawu do zapamiętania (na przykład spółgłoski, cyfry, inne symbole, obrazki lub zdjęcia). Po tym, jak zaprezentowany zostanie ostatni element, pojawia się bodziec testowy. Osoba badana ma w tym momencie za zadanie udzielić odpowiedź, czy bodziec ten znajdował się w zestawie do zapamiętania, na przykład naciskając jeden z dwóch klawiszy, oznaczających odpowiedzi „tak" lub „nie". Rejestrowane są czas (w milisekundach) i poprawność reakcji.

Nadal być może najważniejszym narzędziem służącym do testowaniu hipotez dotyczących przebiegu procesów przetwarzania informacji są przeprowadzane w warunkach laboratoryjnych eksperymenty, w których od osób badanych wymaga się wykonywania mniej lub bardziej złożonych zadań. Zdecydowanie najczęściej analizowanymi zmiennymi zależnymi w tego rodzaju eksperymentach są czas reakcji i poprawność. Zdaniem wielu badaczy czas reakcji jest względnie czułą miarą, pozwalającą wnioskować nie tylko na temat trudności wykonania całego zadania, ale także na temat charakteru i kolejności nie dających się bezpośrednio obserwować procesów poznawczych, a w pewnych warunkach również na temat czasu ich trwania (Luce, 1986; Ashby i Townsend, 1980).

Odtąd przez kompozycję systemu lub procesu będę rozumiał właśnie hipotetyczny charakter i kolejność procesów składowych i występujące między nimi związki logiczne. Przez system należy natomiast rozumieć zbiór elementów strukturalnych i funkcjonalnych, dający się potraktować jako pewna całość, posiadająca wejście i wyjście. Systemem może więc być uwaga selektywna lub pamięć robocza, można też mówić o systemach bardziej podstawowych, z których składa się system złożony, na przykład o podsystemie uwagi w pamięci roboczej.

Na ogół zadania eksperymentalne projektuje się w taki sposób, że albo liczba reakcji błędnych jest stosunkowo mała, często nie większa niż pięć procent wszystkich reakcji, albo stosunkowo duża, zwykle nie przekraczająca jednak poziomu przypadku, aby uniknąć problemów wynikających z utraty informacji na skutek osiągnięcia maksymalnej możliwej liczby błędów (efekt sufitowy lub podłogowy, zależnie od przyjętej skali). Gdy poprawność jest wysoka, zakłada się najczęściej, że sam czas reakcji można interpretować w kategoriach kompozycji procesu, a wyniki dla reakcji błędnych traktuje się jako zanieczyszczenia, usuwane przed przeprowadzeniem analizy statystycznej.

Zanim Sternberg sformułował podstawy metody czynników addytywnych (Sternberg, 1969a), bodaj jedyną metodą ogólnego zastosowania, służącą do wnioskowania na temat kompozycji procesu na podstawie danych w postaci czasów reakcji, była tak zwana metoda substraktywna Dondersa (1869/1969). Odkryta później i znacznie od tego czasu rozwinięta metodologia selektywnego wpływu odegra ważną rolę w tej pracy. Wyjaśnienie, na czym polega i jakie jest znaczenie metodologi selektywnego wpływu wymaga odwołania się do metody substraktywnej, dlatego poświęcę najpierw kilka słów na omówienie

tej ostatniej.

1.1.1 Metoda substraktywna

Metodę substraktywną stosuje się w celu oszacowania czasu trwania jednego z procesów składowych odpowiedzialnych za wykonanie zadania. Polega to na zastosowaniu dwóch zadań takich, że wykonanie jednego z nich wymaga wszystkich operacji potrzebnych do wykonania drugiego, prostszego zadania, a także pojedynczej dodatkowej operacji, która stanowi główny przedmiot zainteresowania badacza. Zakłada się, że czas wykonania zadania bardziej złożonego jest sumą czasu trwania operacji wymaganych przez zadanie prostsze i czasu trwania operacji dodatkowej:

$$RT_1 = A$$
$$RT_2 = A + B$$

gdzie RT_i to zmienna losowa reprezentująca czas wykonania zadania i, a A i B to zmienne reprezentujące czas trwania procesów składowych. Gdyby udało się znaleźć takie dwa zadania, oczekiwany czas trwania interesującej badacza operacji byłby różnicą między oczekiwanymi czasami reakcji dla obu zadań i dałoby się go szacować, odejmując od siebie średnie obserwowane czasy reakcji, stąd określenie „metoda substraktywna"[1]

Kluczowe dla zastosowania metody substraktywnej jest założenie dotyczące kompozycji zadania drugiego, określane jako założenie o czystej różnicy (ang. *pure insertion*). Oba zadania muszą się różnić jedynie obecnością dodatkowego etapu, co oznacza, że w zadaniu drugim proces A musi przebiegać dokładnie tak samo jak w zadaniu pierwszym. Nawet jeżeli zadanie drugie faktycznie wymaga tylko jednego dodatkowego procesu B, ale jednocześnie zmianie ulega czas trwania procesu A, różnica w średnich czasach reakcji dla obu zadań nie będzie oszacowaniem czasu trwania procesu B.

Jak podaje Sternberg (1969b, 2001), na początku dwudziestego wieku pod adresem metody substraktywnej sformułowano szereg uzasadnionych zastrzeżeń. Oszacowanie czasu trwania poszczególnych operacji poznawczych okazało się znacznie utrudnione, między innymi z powodu dużej zmienności wyników uzyskiwanych w ramach różnych eksperymentów z wykorzystaniem tych samych zadań, a także zmienności obserwowanej u tej samej osoby badanej, wykonującej to samo zadanie. Decydujący zarzut dotyczył jednak założenia o czystej różnicy. Świadczyły przeciwko niemu między innymi pochodzące od osób badanych dane introspekcyjne, przede wszystkim jednak poddano w wątpliwość aprioryczną wiarygodność tego założenia. Faktycznie, trudno znaleźć wie-

[1]Zarówno w tym, jak i wielu innych przypadkach omawianych w tej pracy równań, określających zależności między zmiennymi losowymi, znak równości oznacza identyczność rozkładów, a nie po prostu równość. Gdyby było inaczej, czas reakcji w zadaniu drugim byłby skorelowany z czasem reakcji w zadaniu pierwszym, dlatego że oba zależałyby od zmiennej A, podczas gdy są to czasy wykonania dwóch różnych zadań.

le przekonujących przykładów dwóch takich zadań, co do których można by zasadnie przypuszczać, że założenie o czystej różnicy jest dla nich spełnione.

Na skutek coraz częściej pojawiających się głosów krytycznych, zakwestionowana została sama idea dekompozycji systemu w oparciu o wyniki zastosowania prostych zadań z pomiarem czasu reakcji. Renesans tego rodzaju badań przyszedł dopiero na kilka lat przed wspomnianymi eksperymentami Sternberga, a kluczową rolę w tym procesie odegrała oparta na znacznie słabszych założeniach metoda czynników addytywnych (Sternberg, 1969a).

1.1.2 Metoda czynników addytywnych

Metoda czynników addytywnych pozwala zrezygnować z założenia o czystej różnicy, badacz traci więc zwykle możliwość oszacowania czasu trwania poszczególnych operacji. Jednocześnie jednak wnioski oparte na metodzie addytywnej są często bardziej przekonujące, a sama metoda znajduje znacznie szersze zastosowanie.

Podobnie jak metoda substraktywna, metoda addytywna ma służyć do testowania hipotez dotyczących kompozycji procesu. Zakłada się, że wykonanie określonego zadania wymaga na pewnym etapie dwóch, następujących jeden po drugim, choć niekoniecznie natychmiast po sobie, procesów składowych A i B. Żeby uzyskać wsparcie dla tej hipotezy, poszukuje się dwóch czynników eksperymentalnych α i β takich, że na podstawie obserwowanego wzorca czasów reakcji można wnioskować o selektywnym wpływie obu czynników na czas trwania tych procesów. Selektywność wpływu oznacza tutaj, że α wpływa na czas trwania A, ale nie na czas trwania B, natomiast β wpływa na czas trwania B, ale nie A ($A = A(\alpha), B = B(\beta)$).

Nawet jeżeli dwa hipotetyczne procesy przebiegają szeregowo, czas reakcji nie będzie najczęściej sumą czasu ich działania, ponieważ wykonanie zadania będzie prawie zawsze wymagało dodatkowych, częściowo nieznanych procesów, nazywanych zwykle resztowymi (R):

$$RT_{\alpha\beta} = A(\alpha) + B(\beta) + R$$

$RT_{\alpha\beta}$ oznacza tu czas reakcji w warunku eksperymentalnym wyznaczonym przez wartości czynników α i β. Czasy trwania procesów A i B mają być niezależne stochastycznie[2], zakłada się również, że czas trwania procesów resztowych jest niezależny zarówno od czasu trwania procesów A i B jak i wartości obu czynników. Na przykład, jeżeli zadanie polega na reagowaniu za pomocą klawiszy strzałek na prezentowane na ekranie strzałki skierowane w prawo lub w lewo, rozsądne wydaje się przypuszczenie, że jasność bodźców (α) i zgodność klucza reakcyjnego z kierunkiem strzałek (β) będą wpływały selektywnie na następujące po sobie etapy kodowania (A) i selekcji reakcji (B). Jeżeli efekty działania

[2]Dwie zmienne losowe A i B są niezależne stochastycznie wtedy i tylko wtedy, gdy $\forall_{a,b}[P(A = a, B = b) = P(A = a)P(B = b)]$

obu czynników są istotne, ale nie jest już istotny efekt interakcyjny, czyli efekt obu czynników jest addytywny, uzyskane wyniki uznaje się za wsparcie dla zakładanej hipotezy kompozycyjnej. W innym wypadku wyniki nie mówią nic o trafności tej hipotezy.

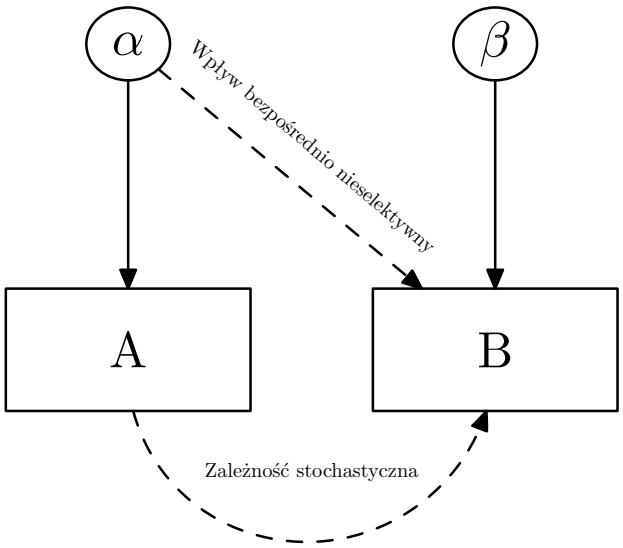
Rezultat niezgodny z założeniami metody addytywnej można wyjaśnić na wiele, wzajemnie niesprzecznych sposobów. Procesy mogą zachodzić przynajmniej częściowo równolegle, czasy ich zakończenia mogą być zależne stochastycznie, wpływ manipulacji może być nieselektywny w przyjętym znaczeniu, wreszcie jeden lub oba czynniki mogą być nieskuteczne. Gdyby jednak w rozważanym hipotetycznym zadaniu, wymagającym identyfikacji kierunku prezentowanej strzałki, udało się zaobserwować addytywny wpływ wymienionych czynników, wsparcie dla hipotezy kompozycyjnej byłoby stosunkowo silne, charakter samego zadania jak i zastosowanych czynników dostarcza bowiem wyraźnych wskazówek na temat natury procesów składowych.

1.1.3 Uogólnienia metody czynników addytywnych

Jak zauważył Townsend (1974; 1983; 1984), teoria leżąca u podstaw metody addytywnej nie pozwala udzielić odpowiedzi na wiele ważnych pytań. Interesujące badacza procesy mogą być uporządkowane inaczej niż szeregowo, na przykład równolegle albo kaskadowo (McClelland, 1979; Ashby, 1982), mogą być także zależne stochastycznie. Nie jest jasne jakiego wzorca, addytywnego czy może jakiegoś interakcyjnego, należy się w takich wypadakach spodziewać.

Nietrudno wyobrazić sobie wiarygodne modele, zarówno szeregowe jak i równoległe, w których występuje zależność stochastyczna. W systemie szeregowym dłuższy czas działania procesu poprzedzającego może prowadzić do krótszego działania procesu następującego później, dzięki temu, że informacja przekazywana do procesu późniejszego jest wtedy bardziej dopracowana. Z kolei w systemie równoległym zależność stochastyczna może wystąpić ze względu na związek logiczny między poszczególnymi operacjami lub regułę decyzyjną, w oparciu o którą wyniki działania procesów równoległych są kombinowane (J. T. Townsend i Wenger, 2004). Jeżeli na przykład równolegle działające procesy gromadzą świadectwa dla wykluczających się alternatyw, zakończenie działania jednego procesu może teoretycznie wystarczyć do podjęcia poprawnej decyzji, a więc również zakończenia działania procesu drugiego.

Dopuszczenie zależności stochastycznej między czasami działania hipotetycznych procesów składowych komplikuje zagadnienie selektywności wpływu. W najprostszym przypadku, gdy poszukiwane są dwa czynniki selektywnie oddziałujące na dwa procesy, jeden lub oba czynniki mogą wpływać na czas działania obu procesów jednocześnie tylko dlatego, że między procesami występuje zależność stochastyczna (J. T. Townsend i Ashby, 1983; J. T. Townsend, 1984; J. T. Townsend i Thomas, 1994; J. T. Townsend i Wenger, 2004). Znajdujący się poniższej diagram ilustruje bezpośrednio i pośrednio nieselektywny wpływ jednego z czynników (α) dla dwóch powiązanych logicznie (wyjście jednego jest wejściem drugiego) procesów szeregowych.



Rysunek 1.1: Pośrednio (zależność stochastyczna) i bezpośrednio nieselektywny wpływ czynnika na dwa procesy szeregowe.

Jeżeli tak jak na diagramie A i B są dwoma powiązanymi logicznie procesami szeregowymi, a f_A i f_B są rozkładami[3] czasów działania tych procesów, to łączny czas działania obu procesów będzie zmienną losową $T_{AB} = T_A + T_B$, o rozkładzie wyznaczonym przez konwolucję[4] rozkładów f_A i f_B, to jest:

$$p(T_{AB} = t) = \int_0^t f_B(t - t_A) f_A(t_A)\, dt_A$$

Uwzględniając ewentualną zależność stochastyczną i wartości czynników α i β, wpływających *bezpośrednio* tylko na odpowiednie procesy, rozkład zmiennej T_{AB} można wyrazić jako:

$$\int_0^t f_B(t - t_A|t_A, \beta) f_A(t_A|\alpha)\, dt_A$$

To jest nadal rozkład sumy dwóch zmiennych losowych T_A i T_B, reprezentujących czasy działania procesów bezpośrednio selektywnie modyfikowanych przez odpowiednie

[3]W całej pracy posługuję się terminem „rozkład prawdopodobieństwa" zarówno w odniesieniu do rozkładów zmiennych dyskretnych jak i ciągłych. W przypadku zmiennych ciągłych należałoby raczej mówić o funkcji gęstości prawdopodobieństwa, jednak to rozróżnienie nie ma znaczenia dla podejmowanych przeze mnie zagadnień, a to, czy chodzi o rozkład, czy o funkcję gęstości, zawsze jasno wynika z kontekstu.

[4]Konwolucja rozkładów dwóch zmiennych losowych jest rozkładem sumy tych zmiennych. Suma dwóch czasów może być rozłożona na nieskończenie wiele sposobów na czasy składowe, dlatego pojawia się całkowanie po zbiorze możliwych par wartości. Jednym ze sposobów ominięcia tej całki jest wyznaczenie rozkładu sumy za pomocą funkcji tworzących momenty, nie jest to jednak rozwiązanie szczególnie uniwersalne.

czynniki, teraz jednak rozkład T_B może pośrednio zależeć od wartości α, wartość ta wpływa bowiem na rozkład czasów t_A, a te z kolei mogą wpływać na rozkład t_B. Dany czynnik może oddziaływać nieselektywnie z powodu występowania zależności bezpośredniej, pośredniej, lub z obu tych powodów jednocześnie[5].

Warunek selektywności wpływu może być spełniony w przypadku dowolnego uporządkowania procesów w czasie. Nic nie stoi też na przeszkodzie, aby zastąpić czas reakcji inną zmienną zależną, na przykład poprawnością lub miarą lokalnej aktywności mózgowej (Sternberg, 2001), wreszcie selektywność można też wykazywać nie tylko na poziomie zmiennej zależnej, ale również na poziomie wolnych parametrów zakładanego modelu (Dzhafarov, 1999, 1997; Sternberg, 2001; Dzhafarov i in., 2004; Voss, Rothermund i Voss, 2004).

Na przykład, wartości parametrów standardowego modelu teorii detekcji (Green i Swets, 1966), kryterium (k) i rozróżnialność (d'), teoretycznie powinny podlegać niezależnej modyfikacji za pomocą takich manipulacji jak macierz wypłat (k) i percepcyjna rozróżnialność bodźców (d')[6]. Wnioskowanie przebiega wtedy w zasadzie bez zmian, jednak selektywność wpływu wykazuje się nie dla obserwowanych wartości zmiennych zależnych (w tym przykładzie poprawności), ale dla oszacowanych na ich podstawie wartości parametrów modelu.

Tak samo jak w przypadku metody addytywnej, wykazanie selektywnej modyfikowalności wartości parametrów stanowi stosunkowo silne wsparcie dla zakładanego modelu. Co szczególnie ważne, czasami taki rezultat dostarcza jednocześnie wsparcia dla interpretacji psychologicznej tych parametrów. Jeżeli jednak uzyskane wyniki nie są zgodne z zakładaną selektywnością oddziaływania, może to wynikać albo z nietrafności modelu, albo stąd, że nie zastosowano odpowiednich czynników, a ogólnie, odpowiedniej procedury eksperymentalnej. Z zastosowaniem metody selektywnego wpływu, zarówno w wersji addytywnej jak i uogólnionej, wiąże się więc asymetria decyzyjna, podobna do tej występującej przy wnioskowaniu na podstawie wyniku testu istotności hipotezy zerowej.

Zdaniem Sternberga (2001), selektywna modyfikowalność miałaby być decydującym kryterium modularności procesów. Mówiąc dokładniej, autor ten twierdzi, że jeżeli uznamy dwa procesy za rozróżnialne funkcjonalnie, powinny istnieć czynniki selektywnie wpływające na ich przebieg, nawet jeżeli, z powodów czysto technicznych,

[5]Pojęcie selektywnego wpływu zostało stosunkowo niedawno dodatkowo uogólnione przez Dzhafarova do wpływu warunkowo selektywnego dla procesów zależnych i niezależnych stochastycznie (Dzhafarov, 1999, 2001, 2003; Dzhafarov, Schweickert i Sung, 2004), jednak jak dotąd nie udało mi się znaleźć przykładów eksperymentalnego zastosowania tego pojęcia, co być może wynika z poziomu matematycznego zaawansowania teorii tego autora.

[6]Sternberg (2001) podkreśla, że próby odkrycia czynników wpływających selektywnie na te parametry nie były szczególnie udane, nie wziął jednak pod uwagę nieszkodliwej dla teorii, ewentualnej zależności stochastycznej między parametrami, o której będę jeszcze pisał przy okazji omawiania modeli kumulacji świadectw.

znalezienie takich czynników jest mało prawdopodobne. W przeciwnym razie hipoteza modularności rozumianej jako funkcjonalna odrębność procesów pozostaje zdaniem Sternberga spekulacją.

Podobnie jak większość innych pojęć psychologicznych, pojęcie modularności nie należy do szczególnie dobrze zdefiniowanych. Być może jedną z popularniejszych charakterystyk modularności systemu sformułował Fodor (1983). Z wymienionych przez niego ośmiu właściwości systemów modularnych, przynajmniej trzy wydają się w jakiś sposób teoretycznie związane z selektywną modyfikowalnością, to jest właściwość ograniczonej dostępności (z zewnątrz), charakterystyczna ontogeneza (regularność rozwojowa) i ustalona realizacja neuronalna. Nie są mi jednak znane żadne wyraźnie sformułowane argumenty, aby spełnienie kryterium selektywnej modyfikowalności uznać za warunek *konieczny* odrębności funkcjonalnej. Nie widzę też żadnych powodów, żeby utożsamiać odrębność funkcjonalną z modularnością. Stanowisko Sternberga wydaje się zbyt restrykcyjne i częściowo niejasne (na podstawie znanych mi publikacji tego autora trudno mi bliżej określić, czym miałaby być jego zdaniem charakterystyka funkcjonalna), niemniej prawdopodobnie nie zaproponowano dotychczas żadnego innego empirycznego kryterium odrębności funkcjonalnej lub modularności o zbliżonej ogólności zastosowań, a jego spełnienie wydaje się często dostarczać silnego wsparcia dla określonej hipotezy kompozycyjnej, z pewnością znacznie silniejszego, niż często stosowana metoda porównywania wyników między zadaniami.

Ta ostatnia, powierzchownie podobna metoda, polega na poszukiwaniu czynników wpływających na miary wykonania w *różnych zadaniach*, z których każde wymaga działania innych, złożonych procesów. Najwięcej informacji dostarcza wtedy wykrycie tak zwanej podwójnej dysocjacji, czyli selektywnego wpływu czynników na *różne zmienne zależne*, związane z odrębnymi zadaniami. Rezultat tego rodzaju nie dostarcza żadnych informacji na temat selektywnej modyfikowalności procesów składowych i wymaga zwykle przyjęcia silnych założeń dotyczących kompozycji procesu odpowiedzialnego za wykonanie poszczególnych zadań[7]. Założenia te nie są testowane, tylko przyjmowane dla potrzeb wnioskowania na temat odrębności wybranych procesów czy systemów. Metoda selektywnego wpływu dopuszcza możliwość zastosowania więcej niż jednej zmiennej zależnej, na przykład różnych miar aktywności mózgu, albo zarówno czasów reakcji jak i poprawności, jednak tylko w ramach pojedynczego zadania (Sternberg, 2001).

Zgoda ze stanowiskiem Sternberga co do znaczenia tego empirycznego kryterium odrębności funkcjonalnej lub modularności oznacza przypisanie pojęciu selektywnej modyfikowalności doniosłej roli w ramach ogólnie rozumianej metodologii badań w psychologii poznawczej. Tym samym wypada uznać, że wszelkie próby rozwinięcia samego pojęcia jak i opartych na nim metod będą miały znaczenie podstawowe dla psycholo-

[7]Publikacja Richardsona-Klavenha i Bjork (1988) zawiera szczegółowe omówienie niektórych problemów związanych z interpretacją efektu podwójnej dysocjacji i słabszych efektów uzyskiwanych w ramach porównywania zadań w kontekście badań dotyczących odrębności systemów pamięciowych.

gii poznawczej jako takiej. Dobrym przykładem zastosowania zarówno metody selektywnego wpływu jak i metody substraktywnej są eksperymenty przeprowadzone przez Sternberga w latach 60-tych.

1.2 Sternberga eksperymenty dotyczące przeszukiwania pamięci krótkoterminowej

W omawianym w tej pracy artykule przeglądowym z 1969 roku Sternberg opisał szereg eksperymentów, dotyczących przeszukiwania pamięci krótkoterminowej, w tym również najczęściej cytowany eksperyment z roku 1966. Wszystkie eksperymenty przeprowadzono według podobnego schematu. Najpierw prezentowane były, pochodzące z wylosowanego wcześniej zestawu, bodźce do zapamiętania, a następnie pojawiał się tak zwany bodziec testowy, który albo należał do prezentowanego wcześniej zestawu (bodziec pozytywny), albo nie (bodziec negatywny). Każdorazowo zadaniem osoby badanej było udzielenie odpowiedzi, możliwie szybko i poprawnie, czy bodziec testowy znajdował się czy nie w prezentowanym wcześniej zestawie. Ten sam zestaw do zapamiętania mógł być prezentowany wielokrotnie przed pojawieniem się bodźca testowego, aby zagwarantować wysoką poprawność wykonania. W takiej sytuacji, w trakcie kolejnych prób, następujących po etapie wstępnego wyuczenia zestawu, prezentowane były jedynie bodźce testowe, a procedura określana była jako zestaw ustalony (ang. *fixed-set*). Jeżeli w trakcie każdej próby prezentowane były bodźce z nowego, każdorazowo losowanego zestawu, procedura określana była jako zestaw zmienny (ang. *varied-set*).

W eksperymencie z 1966 roku zastosowano procedurę z zestawem zmiennym. Bodźcami do zapamiętania były cyfry, prezentowane kolejno wizualnie przez około 1,2 sekundy. Bodziec testowy pojawiał się po około dwóch sekundach od zniknięcia ostatniej cyfry. Za każdym razem po udzieleniu odpowiedzi osoby badane miały odtworzyć wszystkie elementy zestawu, w kolejności ich prezentacji. Zestawy do zapamiętania zawierały od 1 do 6 cyfr, a liczba prób z bodźcami pozytywnymi była taka sama jak liczba prób z bodźcami negatywnymi. W warunku pozytywnym zestaw do zapamiętania zawierał zawsze dokładnie jeden bodziec identyczny z bodźcem testowym (bodziec docelowy). Proporcja odpowiedzi błędnych u żadnej osoby nie przekroczyła dwóch procent wszystkich reakcji.

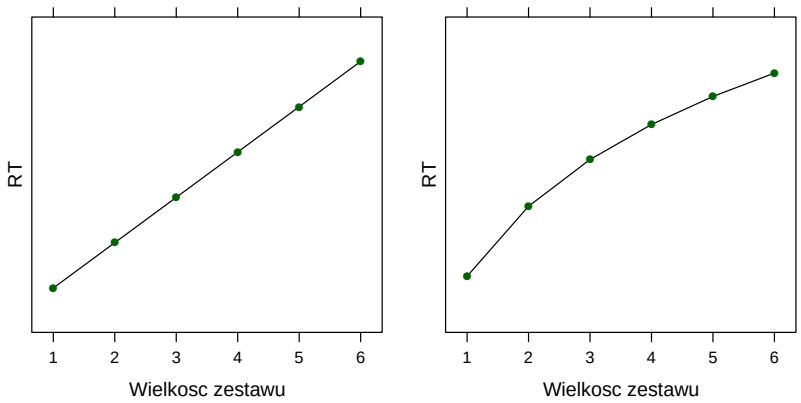
Eksperyment został przeprowadzony w celu empirycznej oceny hipotezy o szeregowym przeszukiwaniu pamięci krótkoterminowej. Zgodnie z tą hipotezą, w pewnym momencie po prezentacji bodźca (po zakończeniu etapu kodowania) rozpoczyna się proces przeszukiwania, polegający na szeregowym i samowygaszającym, czyli kończącym działanie w momencie odnalezienia elementu pasującego, porównywaniu zapisanych w pamięci elementów z bodźcem testowym. Sternberg założył, że kolejność porównywania będzie przypadkowa. Jeżeli bodziec testowy nie znajdował się w zestawie, proces przeszukiwania musiał zachodzić wyczerpująco, to znaczy sprawdzone musiały być wszystkie

elementy zapisane w pamięci. Jeżeli bodziec testowy znajdował się w zestawie, zgodnie z przyjętą hipotezą, przeszukiwanie miało zachodzić samowygaszająco, czyli aż do momentu odnalezienia elementu pasującego.

Uzasadnieniem hipotezy o samowygaszającej naturze przeszukiwania w warunku pozytywnym była domniemana efektywność takiego rozwiązania (Sternberg, 1966, 1969b). Sternberg uznał, że oczekiwany czas reakcji w systemie szeregowym powinien być liniowo zależny od wielkości zestawu, a w przypadku warunku pozytywnego nachylenie linii regresji reprezentujące efekt wielkości zestawu powinno być w przybliżeniu dwa razy mniejsze niż w warunku negatywnym. Konkretnie, chodziło mu o następujący model:

$$\begin{aligned} RT_{Pn} &= R + \sum_{n'=1}^{n} p_{n'} n' T \\ &= \frac{(1 + 2 + \cdots + n)}{n} T = \frac{n+1}{2} T \\ RT_{Nn} &= R + nT \end{aligned} \tag{1.1}$$

gdzie RT_{Pn} to czas reakcji w warunku pozytywnym, RT_{Nn} to czas reakcji w warunku negatywnym, n to wielkość zestawu, T to czas pojedynczej operacji porównywania, R to procesy resztowe, a $p_{n'}$ to prawdopodobieństwo, że znajdujący się w zestawie element docelowy będzie przetwarzany jako n'-ty. Zgodnie z założeniem o przypadkowej kolejności przetwarzania $p_{n'} = 1/n$. Na podstawie przeprowadzonych przez siebie analiz, które w tym miejscu można pominąć, Sternberg stwierdził, że gdyby przeszukiwanie miało zachodzić równolegle, zależność między wielkością zestawu a czasem reakcji powinna być monotonicznie rosnącą funkcją o przyspieszeniu ujemnym. Na poniższych wykresach przedstawiłem jakościowe wzorce zależności, które miały umożliwiać rozstrzygnięcie kwestii następstwa procesów porównywania w czasie:



Rysunek 1.2: Jakościowe wzorce efektu wielkości zestawu przewidywane przez Sternberga dla systemów szeregowych i równoległych

Jak łatwo zauważyć, przy założeniu modelu (1.1) manipulacja wielkością zestawu jest szczególnym przypadkiem zastosowania metody substraktywnej do n zadań. Zgodnie z założeniem o przeszukiwaniu samowygaszającym można też mówić o selektywności wpływu, ponieważ obecność bodźca docelowego w zestawie wpływa jedynie na łączny czas działania operacji porównywania, przebiegających od momentu odnalezienia elementu pasującego. Ten rodzaj selektywności wykracza jednak poza wąskie ramy wyznaczone przez metodę addytywną, dlatego że całkowity czas porównywania zależy wtedy zarówno od obecności bodźca docelowego w zestawie do zapamiętania jak i od wielkości tego zestawu.

Uzyskane przez Sternberga wyniki świadczyły o w przybliżeniu liniowej zależności czasu reakcji od wielkości zestawu i o braku efektu obecności elementu docelowego (nachylenia dla warunków pozytywnego i negatywnego nie różniły się istotnie). Zaobserwowane wartości nachylenia (około 400) i punktu przecięcia (około 38) odpowiadają typowym wynikom uzyskiwanym w tego rodzaju eksperymentach. Rezultaty te miały świadczyć o szeregowym i wyczerpującym przeszukiwaniu pamięci krótkoterminowej.

Zakładając model (1.1), zgodnie z logiką metody substraktywnej, wartości obu parametrów regresji pozwalają wnioskować na temat średniego czasu trwania pojedynczej operacji porównywania (38 milisekund) i łącznego czasu wymaganego do zakodowania bodźca testowego, selekcji i wygenerowania reakcji (400 milisekund). Wyniki zastosowania zestawu ustalonego były podobne.

Sternberg próbował wyjaśnić brak wyraźnych różnic w nachyleniach między warunkami pozytywnym i negatywnym spekulując, że koszty związane z zakończeniem procesu w momencie, gdy element pasujący do bodźca testowego został odnaleziony, mogą być zbyt duże. Przeszukiwanie wyczerpujące byłoby wtedy rozwiązaniem najbardziej efektywnym. Niemniej, wynik ten uznał za zaskakujący, ponieważ, jak zauważył, na jego podstawie można również przypuszczać, że proces przeszukiwania przebiega w sposób daleki od optymalności.

W kolejnym eksperymencie Sternberg manipulował nie tylko wielkością zestawu, ale również stopniem degradacji bodźca. Badanie to miało dostarczyć częściowej odpowiedzi na pytanie o formę reprezentacji bodźca testowego. Jeżeli bodziec testowy reprezentowany jest w postaci względnie nieprzetworzonej, na przykład wizualnej, degradacja powinna zwiększyć nachylenie linii regresji, ponieważ każda indywidualna operacja porównywania powinna zajmować więcej czasu, natomiast punkt przecięcia powinien przypuszczalnie pozostać bez zmian. Warto zauważyć, że w tym przypadku selektywność wpływu nie sprowadza się do addytywnego efektu czynników ze względu na zmienną zależną (czas reakcji), tylko ze względu na wolny parametr modelu regresji (nachylenie). Z drugiej strony, jeżeli bodziec testowy jest wstępnie przetwarzany do reprezentacji symbolicznej, efekt degradacji powinien ujawnić się tylko w zmianie punktu przecięcia. Zakodowanie w postaci symbolicznej powinno zajmować więcej czasu dla bodźca testowego zdegradowanego, ale czas porównywania nie powinien wtedy ulec zmianie.

Według tego autora oba możliwe wzorce zależności miały stanowić dodatkowe wsparcie dla hipotezy szeregowej. Uzyskane wyniki okazały się trudne w interpretacji. Gdy zastosowano zestaw ustalony, wpływ degradacji ujawnił się tylko w zmianie punktu przecięcia, natomiast gdy zastosowano zestaw zmienny, degradacja wywołała różnice zarówno w punkcie przecięcia jak i w nachyleniu, jednak tylko w pierwszej sesji eksperymentu. W sesji drugiej z zestawem zmiennym dało się zaobserwować wyłącznie zmianę w nachyleniu. Gdyby efekt degradacji w każdej wersji eksperymentu zaobserwowano tylko dla punktu przecięcia, albo tylko dla nachylenia, można by wnioskować, znowu zakładając model szeregowy, że oba czynniki (wielkość zestawu i degradacja) wpłynęły selektywnie na dwa odrębne etapy (przeszukiwanie i kodowanie bodźca lub tylko przeszukiwanie). Uzyskane wyniki nie były zgodne z żadną z tych alternatyw.

We wszystkich eksperymentach opisanych w artykule z 1969 roku powtarza się w przybliżeniu liniowy efekt wielkości zestawu, podobny dla warunku negatywnego i pozytywnego. Wynik ten został przez Sternberga uznany za przekonujące wsparcie dla hipotezy szeregowego i wyczerpującego przeszukiwania pamięci krótkoterminowej. Kilka lat później (J. T. Townsend, 1972, 1974), na podstawie analizy metateoretycznej, do-

tyczącej identyfikowalności systemów szeregowych i równoległych, wniosek ten został skutecznie zakwestionowany.

1.3 Warunki identyfikowalności architektury

Wyczerpująca charakterystyka mechanizmu działania dowolnego systemu przetwarzającego informacje w kategoriach procesów składowych wymaga rozstrzygnięcia kwestii następstwa tych procesów w czasie. Pytanie, czy jakieś procesy przebiegają równolegle, szeregowo, czy jeszcze inaczej, należy więc do podstawowych.

Wielokrotnie podejmowano próby rozstrzygnięcia między hipotezami szeregową i równoległą na podstawie efektów manipulacji wielkością przetwarzanego zestawu. Logika rozumowania na której oparte są najważniejsze, sformułowane w latach sześćdziesiątych przez Sternberga wnioski, choć w znacznie mniejszym stopniu same te wnioski, współcześnie nadal uznawana jest przez autorów niektórych ważnych modeli przeszukiwania wzrokowego za względnie przekonującą (Treisman i Gelade, 1980; Treisman i Gormican, 1988; Wolfe, 2007, 1994; Wolfe, Cave i Franzel, 1989). Przytoczę wkrótce argumenty świadczące o tym, że rozumowanie to jest błędne, jednak moim zamiarem jest przede wszystkim zwrócenie uwagi na ogólne warunki, jakie muszą być zwykle spełnione, aby rozumowanie *tego rodzaju* było poprawne.

Przez identyfikację systemu będę tutaj rozumiał procedurę empiryczną, której rezultaty stanowią wsparcie dla dokładnie jednej klasy modeli spośród ogółu klas rozważanych. Terminem „architektura" będę się na razie posługiwał dla określenia możliwego następstwa w czasie procesów składowych, na przykład równoległego, szeregowego lub kaskadowego. Problem identyfikowalności architektur szeregowych i równoległych, prawdopodobnie bardziej niż jakiekolwiek podobne zagadnienie w psychologii poznawczej, stanowił przedmiot powoli postępujących, formalnych analiz metateoretycznych (Logan, 2002). Autorom tych analiz w ciągu trwających niemal czterdzieści lat badań udało się ustalić wiele na temat warunków, w jakich udzielenie względnie przekonującej odpowiedzi na temat architektury systemu wydaje się być w ogóle możliwe.

Prowadzone przez Luce'a, Townsenda, Ashby'ego, Schweickerta, Dzhafarova, Logana, Van Zandt, MacLeoda, Millera, Ratcliffa i innych badania metateoretyczne doprowadziły do powstania nowych metod analizy, na przykład rozkładów czasów reakcji (Luce, 1986; Schweickert i Giorgini, 1999; J. T. Townsend, 1990b; McClelland, 1979; Ashby i Townsend, 1980; Ratcliff i Rouder, 1998; Zandt i Ratcliff, 1995; Zandt, 2002; Ratcliff, 1988b, 1988a, 1978; Logan, 2002), udoskonalenia metody substraktywnej Dondersa (Miller, Ham i Sanders, 1995; Ratcliff, 1988b; Ashby i Townsend, 1980), rozmaitych uogólnień metodologii selektywnego wpływu (Sternberg, 1969a, 2001; J. T. Townsend i Wenger, 2004; J. T. Townsend i Thomas, 1994; J. T. Townsend, 1990b; J. T. Townsend i Schweickert, 1989; J. T. Townsend, 1984; J. T. Townsend i Ashby, 1983; Ashby i Townsend, 1980; Schweickert i Townsend, 1989; Schweickert i Giorgini, 1999; Dzha-

farov i Schweickert, 1995; Dzhafarov i Cortese, 1996; Dzhafarov, 1999, 2001, 2003; Dzhafarov i in., 2004) i wielu innych, mniej lub bardziej kluczowych dla zagadnienia identyfikacji systemu propozycji. W kontekście przeszukiwania pamięci krótkoterminowej i przeszukiwania wzrokowego szczególną rolę odegrała tak zwana metoda podwójnego planu czynnikowego (J. T. Townsend i Ashby, 1983; J. T. Townsend, 1984; J. T. Townsend i Fific, 2004). Pozwala ona na jednoczesne rozstrzygnięcie kwestii architektury, reguły stopu i wydajności w ramach jednego eksperymentu, dla ogólnie zdefiniowanych klas systemów operujących na skończonej liczbie elementów, o ile tylko spełniony jest osłabiony warunek selektywności wpływu. Zastosowanie tej metody nie wymaga żadnych założeń na temat rozkładu procesów elementarnych. Wyniki zastosowanej po raz pierwszy dopiero stosunkowo niedawno do problemu przeszukiwania pamięci krótkoterminowej metody podwójnego planu czynnikowego omówię pod koniec tego rozdziału.

Wszystkie przeprowadzone przez Sternberga w latach sześćdziesiątych eksperymenty dotyczące pamięci krótkoterminowej były próbami empirycznej identyfikacji architektury systemu. Podobnie jak w przypadku każdego innego problemu badawczego, rozwiązanie problemu identyfikacji architektury wydaje się być możliwe wyłącznie w ramach określonej przestrzeni alternatywnych hipotez, a konkluzywność wyników zdaje się zależeć krytycznie od tego, czy przestrzeń ta została poprawnie sformułowana. Wyprowadzone przez Sternberga wnioski były błędne, ponieważ rozważany przez niego zbiór alternatywnych nie uwzględnia niektórych teoretycznie akceptowalnych możliwości.

Względnie przekonująca odpowiedź na pytanie o następstwo operacji porównywania w czasie wymaga wprowadzenia co najmniej kilku dodatkowych rozróżnień. Historia badań dotyczących przeszukiwania pamięci krótkoterminowej jest jednym z motywów przewodnich tego rozdziału, należy jednak pamiętać, że wymienione dalej rozróżnienia dotyczą w zasadzie dowolnych procesów przetwarzania skończonej liczby elementów, niezależnie od tego, czy będzie to przeszukiwanie pamięci krótkoterminowej, wzrokowe, czy na przykład identyfikacja wielowymiarowych bodźców. Następujące dalej omówienie oparłem przede wszystkim na publikacjach Logana (2002), Townsenda (J. T. Townsend, 1984; J. T. Townsend i Ashby, 1983; J. T. Townsend, 1974; J. T. Townsend i Wenger, 2004; J. T. Townsend i Nozawa, 1995) i McClellanda (McClelland, 1979).

1.3.1 Równoległość-szeregowość

Do tej pory dosyć swobodnie posługiwałem się pojęciami równoległości i szeregowości. Żeby uniknąć nieporozumień, wypada je teraz dokładniej zdefiniować. Podobnie jak czynią to Townsend i Logan, przez system szeregowy będę rozumiał następujące jeden po drugim procesy dyskretne. W systemie takim proces poprzedzający musi się zakończyć, zanim rozpocznie się proces następny. Przez system równoległy będę rozumiał procesy rozpoczynające się w tym samym momencie. W systemach równoległych moment za-

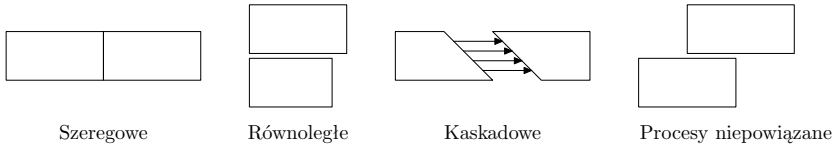
kończenia może, chociaż nie musi, być różny dla różnych procesów. O ile mi wiadomo, niewiele udało się ustalić na temat własności systemów hybrydowych, to znaczy takich, które mogą w tych samych warunkach działać czasami szeregowo, a czasem równolegle, chociaż w wielu wypadkach nie ma żadnego teoretycznie uzasadnionego powodu, aby takie systemy z góry wykluczać.

Rozważane w dalszej części pracy kryteria identyfikacji dotyczą czystych (to jest niehybrydowych) systemów równoległych i szeregowych. W praktyce wyniki wyraźnie zgodne z jedną z tych możliwości i wyraźnie niezgodne z drugą wydają się wystarczające do stwierdzenia, że badany system znajduje się w pobliżu krańca kontinuum wyznaczonego przez przypadki czyste. Ogólne systemy hybrydowe są oczywiście bardziej złożone niż ogólne czyste systemy szeregowe i równoległe, ponieważ wymagają dodatkowego parametru, określającego prawdopodobieństwo z jakim system działa w danych warunkach szeregowo, a także osobnych parametrów, wyrażających charakterystyki wydajnościowe procesów szeregowych i równoległych.

1.3.2 Dyskretność-ciągłość

Kolejne rozróżnienie dotyczy tego, czy hipotetyczne procesy przebiegają w sposób ciągły, czy dyskretny. Proces A jest dyskretny, jeżeli przekazuje informację do następującego po nim, logicznie związanego procesu B natychmiast i w całości. W przypadku procesów dyskretnych moment rozpoczęcia i zakończenia procesu poprzedzającego (A) jest jasno określony. Z kolei proces A jest ciągły, jeżeli przekazuje informację do procesu B stopniowo, w serii mniejszych, być może nawet nieskończenie małych kroków. Przy założeniu ciągłości, drobne zmiany zachodzące w procesie A mogą być natychmiast przekazywane do B, wobec czego efekty działania A pojawiają się przed tym, jak proces ten zakończy działanie. Dopuszczając dowolną wielkość porcji przekazywanej informacji („ziarnistość"; Miller, 1982, 1988), procesy dyskretne i ciągłe można potraktować jako bieguny wyznaczające kontinuum.

Gdy procesy są ciągłe, porządek czasowy i logiczny odpowiada szeregowemu następstwu etapów, natomiast jednoczesna aktywność obu procesów zgodna jest raczej z przebiegiem równoległym, wypada więc przyjąć dla nich osobną nazwę. McClelland (1979) nazywa je procesami kaskadowymi. Termin „kaskadowy" wskazuje na to, że jeden z procesów rozpoczyna się (a być może również kończy) pierwszy, jednak oba procesy mogą przebiegać przez pewien czas równolegle. Kaskadowe następstwo w czasie może dotyczyć zarówno procesów związanych jak i niezwiązanych logicznie, ciągłych lub dyskretnych. Nawet gdy istnieją wyraźne powody aby przypuszczać, że proces przebiega w sposób ciągły, dla uproszczenia najczęściej zakłada się dyskretność. Możliwe uporządkowania dwóch procesów w czasie przedstawia poniższy diagram.



Rysunek 1.3: Równoległe i szeregowe procesy dyskretne i ciągłe

1.3.3 Wydajność

Następne ważne rozróżnienie dotyczy wydajności (ang. *capacity*). Wydajność jest właściwością systemu, o którym można powiedzieć, że wykonuje pewną pracę. Zagadnienie wydajności odgrywa kluczową rolę w kontekście badania wielu funkcji poznawczych, takich jak percepcja, uwaga, kontrola poznawcza, czy pamięć. Za Townsendem i Ashbym (1983) przyjmuję tu definicję wydajności jako tempa (ang. *rate*) przetwarzania informacji, czyli stosunku ilości informacji przetworzonej do czasu przetwarzania.

Typowe miary wydajności opierają się na efektach zmian liczby elementów, które muszą być przetworzone do podjęcia poprawnej decyzji, określanych często jako efekty ładunku. Ponieważ wydajność zdefiniowana jest w kategoriach tempa przetwarzania, a nie na przykład poprawności, rozważane dalej miary wydajności będą oparte na obserwowanym czasie reakcji.

Jeżeli wraz ze wzrostem ładunku zwiększa się czas wykonania zadania, wydajność jest ograniczona. Wydajność jest nieograniczona, jeżeli zmiana ładunku nie powoduje zmian w czasach reakcji. System charakteryzuje się superwydajnością, gdy wzrost liczby elementów powoduje skrócenie czasu reakcji, wreszcie wydajność jest ustalona, gdy łączna wydajność wszystkich procesów składowych jest stała, niezależnie od liczby elementów. Wydajność jest ograniczona, nieograniczona, lub jest superwydajnością ze względu na wpływ liczby elementów na łączny czas przetwarzania, czego nie można powiedzieć o wydajności ustalonej, klasyfikacja ta jest zatem wadliwa, ale przy zachowaniu minimum ostrożności można ją z pożytkiem stosować. Należy pamiętać, że wydajność systemu jako całości jest logicznie niezależna od wydajności procesów składowych, wobec czego charakterystyka wydajnościowa systemu zawsze będzie zależała od przyjętego poziomu abstrakcji.

Przykładem systemu o ograniczonej wydajności jest system szeregowy, w którym średnie tempo przetwarzania każdego elementu jest stałe i nie ulega zmianie wraz ze wzrostem ładunku. W przypadku takiego systemu oczekiwany czas przetwarzania będzie liniowo zależny od liczby elementów niezależnie od tego, czy system ten działa wyczerpująco, czy samowygaszająco[8].

[8] O ile wykluczyć z góry takie egzotyczne scenariusze, jak możliwość występowania zawsze jednego elementu docelowego, który jest przetwarzany samowygaszająco zawsze jako pierwszy, niezależnie od liczby dystraktorów.

Do systemów o wydajności nieograniczonej należy między innymi system równoległy samowygaszający o stałych, identycznych i niezależnych tempach poszczególnych procesów. Jeżeli zadanie na to pozwala, na przykład występuje tylko jeden bodziec docelowy wśród kilku dystraktorów, w przypadku takiego systemu czas działania będzie niezależny od liczby elementów (dystraktorów), ponieważ będzie całkowicie zdeterminowany przez niezależny od liczby elementów czas działania pojedynczego procesu przetwarzającego element docelowy.

Do równoległych systemów superwydajnych należą między innymi systemy określane czasem jako koaktywacyjne (J. T. Townsend i Nozawa, 1995), to jest na przykład takie, w których rezultaty przetwarzania w ramach więcej niż jednego kanału trafiają w postaci zsumowanej („zsumowana aktywacja") do kanału wyjściowego, a także systemy, w których równolegle działające procesy są stochastycznie zależne w taki sposób, że kumulacja świadectw w ramach jednego procesu wywołuje wzrost tempa przetwarzania w innym procesie, prowadząc do systematycznego wzrostu wydajności całego systemu w czasie. Nie oznacza to jednak, że systemy koaktywacyjne będą z konieczności działały superwydajnie. Jeżeli wydajność procesów składowych będzie się wystarczająco zmniejszała wraz ze wzrostem ładunku, systemy takie będą nieodróżnialne od odpowiednich systemów o nieograniczonej, lub nawet ograniczonej wydajności.

Stosowane tutaj pojęcie wydajności należy odróżnić od pojęcia zasobów. Scharakteryzowanie procesu pod względem wydajności jest bardziej neutralne teoretycznie i nie wydaje się wymagać przyjęcia dodatkowych założeń na temat wykorzystywanej w trakcie przetwarzania „energii mentalnej". Wiele teorii zasobowych opiera się na założeniach, że pula zasobów jest stała lub jakoś inaczej ograniczona, że zasoby mogą być równolegle przydzielane do różnych zadań, a także, że poziom wykonania zadania zmienia się płynnie w miarę zmian w przydzielanej do wykonania zadania ilości zasobów (Navon, 1984). Jak trafnie zauważył Logan (2002), przyjęcie któregokolwiek z wymienionych założeń dotyczących wydajności nie implikuje żadnego z wymienionych założeń na temat zasobów. Można więc mówić o wydajności jednocześnie nie zgadzając się na istnienie czegoś takiego jak zasoby.

Na podstawie przytoczonych wcześniej przykładów można wywnioskować, że wydajność systemu zależy od wydajności procesów składowych, architektury (równoległa, szeregowa), reguły stopu i zależności między procesami, daje się więc scharakteryzować tylko przy jednoczesnym uwzględnieniu *co najmniej* wszystkich wymienionych wymiarów. Bodaj najczęstszym nieporozumieniem związanym z zagadnieniem szeregowości i równoległości jest utożsamianie systemów szeregowych z systemami o ograniczonej wydajności, a systemów równoległych z systemami o wydajności nieograniczonej (J. T. Townsend, 1990a; Logan, 2002). Oba te wymiary, to jest architektura i wydajność, są logicznie niezależne. Nawet system szeregowy, przynajmniej teoretycznie, może wykazywać wydajność nieograniczoną, jeżeli tylko tempo przetwarzania dla każdego procesu składowego będzie wystarczająco rosło wraz ze zwiększeniem się liczby elementów.

O bezzasadności utożsamienia systemów równoległych z systemami o nieograniczonej wydajności łatwo się przekonać, rozważając zadanie wymuszające wyczerpującą regułę stopu, na przykład warunek negatywny w zadaniu na przeszukiwanie pamięci. Jak wykazali J. T. Townsend i Ashby (1983), czas działania systemu równoległego o niezależnych tempach i nieograniczonej wydajności na poziomie procesów składowych w warunku przetwarzania wyczerpującego będzie monotonicznie rosnącą funkcją ładunku. Wydajność takich systemów będzie więc zawsze ograniczona na poziomie całego procesu. Brak efektu ładunku, a więc nieograniczona wydajność na poziomie całego zadania w sytuacji przetwarzania wyczerpującego, wymaga aby system był superwydajny na poziomie procesów składowych, co czasem można wykluczyć na podstawie niskiej apriorycznej wiarygodności takiej hipotezy. Wnioskowanie o szeregowości architektury wyłącznie na podstawie zaobserwowanego istotnego (niekoniecznie liniowego) efektu ładunku jest błędne. Wkrótce przedstawię argumenty świadczące o tym, że błędem jest także wnioskowanie o szeregowości na podstawie *liniowego* efektu ładunku.

1.3.4 Zależność stochastyczna

Wyczerpująca charakterystyka dowolnego systemu przetwarzającego informacje wymaga także rozstrzygnięcia kwestii zależności stochastycznej między czasami działania (lub tempami) procesów składowych. Założenie o niezależności czasów lub temp znacząco upraszcza rozumowanie, co nie znaczy oczywiście, że można je przyjmować arbitralnie.

Wydaje się, że zależność stochastyczna powinna być raczej regułą niż wyjątkiem w przypadku powiązanych logicznie procesów ciągłych. Jak już wspomniałem wcześniej, w przypadku procesów równoległych zależność może przybierać między innymi postać koaktywacji - gdy równoległe działanie dwóch lub więcej procesów powoduje wzajemne zwiększenie się temp tych procesów w czasie, albo aktywacja generowana przez dwa równoległe procesy jest sumowana w pojedynczym kanale wyjściowym aż do momentu przekroczenia progu decyzyjnego - albo inhibicji, gdy równoległe działanie dwóch lub większej liczby procesów związane jest z wzajemnym zmniejszaniem się ich temp w czasie, na przykład gdy każdy z procesów gromadzi świadectwa dla wykluczających się alternatyw (Miller, 1988; J. T. Townsend i Nozawa, 1995).

1.3.5 Reguła stopu

W pewnych warunkach możliwe jest zakończenie przetwarzania i podjęcie poprawnej decyzji zanim opracowane zostaną wszystkie dostępne elementy. To, czy działanie zostanie w takiej sytuacji faktycznie zakończone, określa reguła stopu dla danego systemu. Jeżeli w zestawie do przeszukiwania znajduje się jeden element docelowy, możliwe jest zakończenie przetwarzania i podjęcie poprawnej decyzji w momencie odnalezienia tego elementu, co odpowiada regule samowygaszającej. Jeżeli mimo to przetwarzane są do końca wszystkie elementy, obowiązuje reguła wyczerpująca. Z drugiej strony, jeżeli

poprawne wykonanie zadania wymaga pełnego opracowania wszystkich elementów, reguła stopu (wyczerpująca) jest wymuszona i wyniki eksperymentu nie mogą pozwolić na jej identyfikację. Jak się wkrótce przekonamy, obie reguły stopu mogą prowadzić do zupełnie innych konsekwencji zależnie od tego, czy system jest równoległy, czy szeregowy. Zależnie od charakteru zadania teoretycznie wiarygodne mogą być jeszcze inne, bardziej złożone reguły, takie jak reguła dysjunktywna (albo A, albo B), jednak dla potrzeb dalszych rozważań wystarczy jeżeli ograniczę się do reguł samowygaszającej i wyczerpującej.

Wymienione niezależne logicznie wymiary, to jest równoległość-szeregowość, dyskretność-ciągłość, wydajność, zależność stochastyczna i reguła stopu pozwalają rozważyć bogaty zbiór ogólnych klas modeli. Omówię teraz niektóre ważniejsze konsekwencje przyjęcia różnych założeń co do tych wymiarów na tyle szczegółowo, aby Czytelnik mógł nabrać pewnej orientacji w całości problematyki, jednak wiele wątków będę zmuszony pominąć. Wszędzie tam, gdzie wydawało mi się, że nie powinno to powodować nadmiernych nieporozumień, zdecydowałem się na nieformalny styl prezentacji.

1.4 Problem mimikry na przykładzie architektury systemu przeszukiwania pamięci krótkoterminowej

W ogólnym przypadku wyprowadzenie pozwalających na identyfikację systemu predykcji, dotyczących związku między czasem reakcji a liczbą i rodzajem elementów (element docelowy lub dystraktor) wymaga rozstrzygnięcia co najmniej wszystkich wymienionych wyżej kwestii, to jest równoległości-szeregowości, reguły stopu, wydajności i zależności stochastycznej. Tym samym znacznemu poszerzeniu ulega zbiór alternatywnych hipotez, które należy uwzględnić próbując wyjaśnić między innymi zaobserwowane przez Sternberga efekty wielkości zestawu.

Townsend (1972) wykazał, że wystarczy porzucić założenie o nieograniczonej wydajności przeszukiwania równoległego, aby liniowy efekt wielkości zestawu nie pozwalał rozróżnić między systemami szeregowym i równoległym. Nierozróżnialność empiryczna modeli opartych na wzajemnie wykluczających się założeniach ochrzczona została mianem „problemu mimikry", a model Townsenda był prawdopodobnie jednym z pierwszych w historii psychologii poznawczej przypadkiem formalnej demonstracji tego problemu (Logan, 2002). Wkrótce liczba teoretycznie wiarygodnych, stosunkowo prostych systemów, opartych na zupełnie innych założeniach, które okazały się nierozróżnialne za pomocą typowych procedur eksperymentalnych i metod analizy, zaczęła niepokojąco rosnąć[9].

[9]Omówienie przykładów mimikry wykraczające poza zagadnienie szeregowości-równoległości znajduje się między innymi u Luce'a (1986) i Logana (2002).

1.4.1 Model równoległego przeszukiwania pamięci krótkoterminowej z ograniczoną wydajnością

Jak zauważyli Atkinson, Holmgren i Juola (1969), a także J. T. Townsend (1974), Sternberg założył, że wydajność przeszukiwania równoległego jest nieograniczona na poziomie operacji składowych. Inaczej mówiąc, tempo przebiegu każdego indywidualnego procesu porównywania miało być takie samo, niezależnie od liczby jednocześnie przetwarzanych elementów. Townsendowi i Ashby'emu (1983) udało się udowodnić, że przewidywana przez Sternbegra dla procesu równoległego, monotonicznie rosnąca zależność o przyspieszeniu ujemnym, faktycznie występuje w przypadku każdego systemu równoległego niezależnego o nieograniczonej wydajności. Omówię teraz kilka podstawowych własności jednej z klas takich modeli.

Dla uproszczenia dogodnie jest przyjąć, że czas zakończenia każdego spośród N równoległych procesów ma rozkład wykładniczy, czyli prawdopodobieństwo zakończenia procesu i w momencie t wynosi $v_i e^{-v_i t}$, $v_i > 0, t \geqslant 0$, gdzie v_i to tempo procesu i, którego wartość można interpretować jako liczbę elementów przetwarzanych w jednostce czasu. Jeżeli na przykład tempo v wynosi $1/2$, to jeden element zostanie przetworzony średnio w ciągu dwóch jednostek czasu (oczekiwana wartość zmiennej o rozkładzie wykładniczym i tempie v wynosi $1/v$).

Założenie o wykładniczym rozkładzie czasów zakończenia ułatwia wiele obliczeń, istnieją też (bardzo) słabe przesłanki teoretyczne, aby tak postąpić, ponieważ rozkład wykładniczy jest rozkładem o maksymalnej entropii[10] wśród wszystkich rozkładów ciągłych o wsparciu w przedziale $[0, \infty)$ i średniej $1/v$. Należy jednak pamiętać, że założenie o wykładniczym rozkładzie procesów przyjmuję w tej pracy tylko dla uproszczenia dowodów i omawiane dalej własności systemów szeregowych i równoległych są niezależne od zakładanego rozkładu czasów zakończenia procesów składowych (J. T. Townsend i Nozawa, 1995). Dowody niektórych własności wykładniczych systemów szeregowych i równoległych niezależnych mają jedynie ułatwić Czytelnikowi śledzenie toku rozumowania.

1.4.2 Niektóre własności systemów równoległych

Dla uproszczenia rozważmy przypadek dwóch procesów równoległych A i B, o niezależnych rozkładach wykładniczych i niekoniecznie jednakowych tempach, v_A i v_B. Czas od momentu rozpoczęcia przetwarzania do momentu, gdy jeden z procesów zakończy

[10]Rozkład prawdopodobieństwa o maksymalnej entropii to rozkład, którego entropia jest co najmniej tak duża jak entropia wszystkich pozostałych członków określonej klasy rozkładów. Maksymalizacja entropii minimalizuje ilość informacji „wpisanej w rozkład", co czyni takie rozkłady domyślnym wyborem, gdy niewiele wiadomo na temat badanego procesu. Rozkład normalny jest takim rozkładem o maksymalnej entropii wśród wszystkich rozkładów o wsparciu w zbiorze liczb rzeczywistych z określoną średnią i odchyleniem standardowym.

działanie, będę określał jako etap pierwszy, natomiast czas od momentu zakończenia etapu pierwszego do zakończenia działania drugiego procesu będę określał jako etap drugi. Zakładam, że zawsze zachodzą oba etapy, przetwarzanie jest więc wyczerpujące.

Zarówno proces A jak i B może zakończyć działanie w etapie pierwszym. Ponieważ oba procesy są niezależne, prawdopodobieństwo, że etap pierwszy zakończy się w momencie t to suma prawdopodobieństw zdarzenia, że proces A zakończy działanie w momencie t jako pierwszy i zdarzenia, że proces B zakończy działanie w momencie t jako pierwszy. Prawdopodobieństwo, że B zakończy działanie w jakimś momencie późniejszym niż t wynosi $p(T_B > t) = 1 - p(T_B \leqslant t) = 1 - F(t, v_B)$, gdzie T_B to zmienna losowa reprezentująca czas zakończenia procesu B, a $F(t, v_B)$ to dystrybuanta rozkładu wykładniczego o tempie v_B, równa $e^{-v_B t}$. Oznaczając przez $f_i(t, v_i)$ prawdopodobieństwo, że proces $i = A, B$ zakończy działanie w momencie t, rozkład f_1 czasów zakończenia pierwszego etapu będzie równy:

$$\begin{aligned} f_1(t) &= f_A(t, v_A)[1 - F(t, v_B)] + f_B(t, v_B)[1 - F(t, v_A)] \\ &= v_A e^{-v_A t} e^{-v_B t} + v_B e^{-v_B t} e^{-v_A t} \\ &= v_A e^{-(v_A + v_B)t} + v_B e^{-(v_B + v_A)t} \\ &= (v_A + v_B) e^{-(v_A + v_B)t} \end{aligned} \tag{1.2}$$

Jak widać, rozkład czasów zakończenia pierwszego etapu jest też wykładniczy, z tempem równym sumie tmp procesów indywidualnych, a więc oczekiwany czas zakończenia etapu pierwszego wynosi $1/(v_A + v_B)$.

Całkując $f_A(t, v_A)[1 - F(t, v_B)]$ po t można ustalić prawdopodobieństwo, że określony proces (tutaj A) zakończy działanie jako pierwszy - wynosi ono $v_i/(v_A + v_B)$ dla procesu $i = A, B$. Posługując się tą wielkością można z kolei wyznaczyć rozkład czasów zakończenia pierwszego etapu pod warunkiem, że określony proces zakończył działanie jako pierwszy. Na przykład, dla procesu A:

$$\begin{aligned} f_{A1} &= f_A(t, v_A)[1 - F(t, v_B)] / \left(\frac{v_A}{v_A + v_B} \right) \\ &= v_A e^{-(v_A + v_B)t} / \left(\frac{v_A}{v_A + v_B} \right) \\ &= (v_A + v_B) e^{-(v_A + v_B)t} \end{aligned} \tag{1.3}$$

Wynika stąd, co być może zaskakujące, że tempo pierwszego etapu jest takie samo, niezależnie od tego który proces, szybszy czy wolniejszy, zakończył w nim działanie. Jasne jest również, że tempo każdego etapu będzie rosło wraz ze wzrostem liczby jednocześnie działających procesów, o ile tylko tempo procesów indywidualnych nie będzie wystarczająco malało. Można to wyjaśnić posługując się analogią między przetwarzaniem równoległym i wyścigiem. Wraz ze wzrostem liczby ścigających się zawodników rośnie prawdopodobieństwo, że któryś z nich wcześniej dotrze do mety.

1.4. Problem mimikry na przykładzie architektury systemu przeszukiwania pamięci krótkoterminowej

Pozostaje jeszcze ustalić, jaki jest rozkład czasów etapu drugiego pamiętając, że czas ten liczony jest od momentu zakończenia etapu pierwszego. Z pomocą przychodzi tutaj pewna przyjemna własność rozkładu wykładniczego, mianowicie:

$$p(T > t + s|T > s) = p(T > t),\ t \geqslant 0, s \geqslant 0$$

Własność ta oznacza, że jeżeli proces wykładniczy o danym tempie nie zakończył jeszcze działania, to można zacząć liczyć czas od zera i rozkład liczonego w ten sposób czasu zakończenia będzie też wykładniczy z takim samym tempem. Skoro tak, to rozkład czasu etapu drugiego będzie sumą rozkładów wykładniczych o tempach v_A i v_B, ważonych przez prawdopodobieństwa, że proces B (A) zakończył działanie jako pierwszy:

$$f_2(t) = \frac{v_B}{v_A + v_B} v_A e^{-v_A t} + \frac{v_A}{v_A + v_B} v_B e^{-v_B t}$$

To jest mieszanina rozkładów wykładniczych i można zauważyć, że rozkład czasu etapu drugiego musi być mieszaniną niezależnie od tego, czy czasy poszczególnych procesów są wykładnicze, czy nie. Mieszanina takich samych rozkładów jest zwykle rozkładem innego typu, co być może dałoby się wykorzystać w celu empirycznej identyfikacji architektury systemu. Przy założeniu, że rozkłady czasów indywidualnych procesów są identyczne, rozkłady kolejnych etapów będą takie same jak rozkłady procesów indywidualnych w modelach szeregowych, ale w ogólnym przypadku będą różne w modelach równoległych.

Omówione do tej pory własności modeli równoległych z niezależnymi czasami wykładniczymi można stosunkowo łatwo uogólnić na dowolne n. Przedstawiając wyniki analiz ograniczyłem się do przypadku zestawu dwuelementowego dlatego, że jest on najprostszy i pozwala zademonstrować wiele kluczowych zagadnień związanych z zachowaniem się architektur szeregowych i równoległych. Dla pełniejszego objaśnienia własności architektur trzeba czasem wyrazić oba rodzaje modeli w ogólniejszej postaci, uwzględniającej ewentualną zależność stochastyczną między procesami i możliwą zależność temp procesów od etapu, liczby elementów i kolejności ich przetwarzania. Ta ogólniejsza postać ma jednak zastosowanie głównie metateoretyczne. Liczba wolnych parametrów przekracza wtedy zdecydowanie liczbę niezależnych punktów danych nawet bardzo wymyślnego eksperymentu, dlatego modele takie nie są identyfikowalne empirycznie w zwykłym, statystycznym znaczeniu, to znaczy wartości wolnych parametrów tych modeli nie dają się unikalnie oszacować na podstawie danych.

Korzystając z wymienionych własności modeli równoległych można teraz wyjaśnić, w jaki sposób postulowana przez Sternberga krzywoliniowa zależność czasu reakcji od wielkości zestawu może, choć nie musi, wynikać z równoległości przetwarzania. Niech wydajność v każdego indywidualnego procesu o rozkładzie wykładniczym w modelu równoległym wyczerpującym z niezależnymi czasami będzie taka sama, niezależnie od etapu i wielkości zestawu. W ramach każdej próby indywidualne procesy można uporządkować pod względem kolejności, w jakiej zakończyły działanie. Tak jak wcześniej, interwały pomiędzy momentami zakończenia kolejnych procesów będę nazywał etapami,

przy czym etap pierwszy to po prostu czas zakończenia pierwszego procesu. Na podstawie (1.2) wiemy już, że czas każdego etapu będzie miał rozkład wykładniczy. Wydajność etapu pierwszego jest teraz równa nv, wobec czego oczekiwany czas zakończenia etapu pierwszego wynosi $1/nv$. W etapie drugim „ściga się" już tylko $n-1$ procesów, wydajność wynosi więc $(n-1)v$, i tak dalej, dla wszystkich pozostałych etapów. Oczekiwany czas zakończenia wszystkich procesów $E(T)$ można przedstawić jako sumę oczekiwanych czasów trwania kolejnych etapów:

$$\begin{aligned} E(T) &= \frac{1}{nv} + \frac{1}{(n-1)v} + \frac{1}{(n-2)v} + \cdots + \frac{1}{v} \\ &= \frac{1}{v} + \frac{1}{2v} + \cdots + \frac{1}{nv} \\ &= \frac{1}{v} \sum_{i=1}^{n} \frac{1}{i} \end{aligned} \tag{1.4}$$

Dodanie każdego kolejnego elementu powoduje coraz mniejszy przyrost oczekiwanego czasu zakończenia wszystkich równoległych porównań (gdy dodany jest m-ty element przyrost wynosi $1/vm$), dlatego, zgodnie z przewidywaniami Sternberga dla przeszukiwania równoległego, zależność średniego czasu reakcji od wielkości zestawu dla takiego systemu jest monotonicznie rosnącą funkcją o przyspieszeniu ujemnym.

Pomijając kwestię przyjętego dla uproszczenia, dokładnego rozkładu czasów zakończenia procesów, model ten jest najprostszym modelem równoległym w tym znaczeniu, że posiada tylko jeden wolny parametr v, określający tempo każdego procesu. Analogiczny, ale prowadzący do innych predykcji model szeregowy miałby takie samo tempo dla każdego, niezależnego stochastycznie procesu, dla każdej liczby elementów i kolejności przetwarzania. Ze względu na ich prostotę, modele te określa się czasem jako standardowe.

Co ciekawe, ogólny model szeregowy okazuje się bardziej elastyczny, a więc bardziej złożony w sensie statystycznym[11], niż ogólny model równoległy, ponieważ w ogólnym modelu szeregowym tempo każdego procesu może teoretycznie zależeć zarówno od etapu jak i od całej kolejności przetwarzania wszystkich elementów (J. T. Townsend i Evans, 1983; J. T. Townsend, 1984). Z kolei w ogólnym modelu równoległym tempo określonego procesu w danym etapie lub też danego etapu nie może zależeć od całej kolejności w jakiej przetworzone są w danej próbie elementy, ponieważ kolejność ta ustala się stopniowo, w miarę jak procesy indywidualne kończą swoje działanie, i nie może być, tak ja to jest przynajmniej teoretycznie możliwe w modelu szeregowym, wyznaczona z góry.

[11]Złożoność modelu w znaczeniu statystycznym będzie przedmiotem moich rozważań w rozdziale trzecim.

1.4.3 Model równoległy zgodny z liniowym efektem wielkości zestawu

Rezultaty uzyskane przez Sternberga wykluczają *standardowy* model równoległy. Nie oznacza to jednak, że jakiś inny, wiarygodny model równoległy nie mógłby zachowywać się zgodnie z uzyskanymi przez niego wynikami. Aby skonstruować model równoległy wyczerpujący, nierozróżnialny na podstawie efektów wielkości zestawu od standardowego modelu szeregowego, trzeba zagwarantować, że oczekiwany czas każdego etapu będzie taki sam. Krzywoliniowy efekt wielkości zestawu w standardowym modelu równoległym bierze się między innymi stąd, że w miarę jak kolejne procesy kończą działanie, wyścig rozgrywa się między coraz mniejszą liczbą procesów, przez co tempo każdego kolejnego etapu jest coraz mniejsze. Tempo każdego etapu jest mniejsze od tempa etapu poprzedniego o v, wobec czego wystarczy za każdym razem dodać brakujące tempo i rozkłady kolejnych etapów będą wykładnicze z jednakowym tempem, tak samo jak w standardowym modelu szeregowym.

Sama realokacja temp to jednak za mało. Niech w modelu równoległym z realokacją tempo każdego procesu w trakcie etapu pierwszego wynosi v. Dla dowolnego n, wyczerpująca reguła stopu wymaga, aby odbyło się n etapów. Gdy $n = 2$, tempo każdego etapu z realokacją wynosi $2v$, a więc oczekiwany czas zakończenia całego procesu wynosi $1/2v + 1/2v = 1/v$ i oczywiście dla dowolnego n oczekiwany czas zakończenia całego procesu jest równy $n/nv = 1/v$. W ten sposób można zagwarantować jednakowe tempo każdego kolejnego etapu, niemniej model wykazuje nieograniczoną wydajność na poziomie całego zadania, co wyraźnie odróżnia go od standardowego modelu szeregowego. Ta nieograniczona wydajność wynika z łącznego wpływu liczby elementów i realokacji. Nic nie stoi jednak na przeszkodzie, aby uzależnić tempo każdego procesu od wielkości zestawu. Przyjmując, że wynosi ono $v = u/n$ dla pewnej stałej u, a więc wydajność jest ustalona, uzyskujemy predykcje identyczne jak dla standardowego modelu szeregowego dla dowolnego n na poziomie rozkładu czasów poszczególnych etapów i rozkładu czasów zakończenia całego procesu. Realokacja gwarantuje jednakowość rozkładów każdego etapu, a ustalona wydajność daje liniowy efekt wielkości zestawu, ponieważ $n/(n(u/n)) = n/u$. Identyczność predykcji na poziomie rozkładów oznacza, że niezależnie od zastosowanej statystyki, na przykład średniej albo wariancji, oba modele są całkowicie nierozróżnialne empirycznie. Taki model równoległy jest zgodny na przykład z założeniem, że „suma zasobów" jest stała, a zwalniane po każdym etapie zasoby są natychmiast ponownie przydzielane do procesów nadal aktywnych. Wbrew temu, co sądził Sternberg, liniowy efekt wielkości zestawu nie pozwala rozstrzygnąć między dwoma, stosunkowo wiarygodnymi modelami szeregowym i równoległym.

Nieidentyfikowalność architektury na podstawie liniowego efektu wielkości zestawu można uzyskać opierając się na innych założeniach, również w przypadku, gdy obserwuje się różnice w nachyleniach między warunkiem pozytywnym i negatywnym. Chociaż różnica nachyleń w tego rodzaju eksperymentach nie pozwala ustalić architektury

systemu, czasami pozwala na wyprowadzenie mocnych wniosków na temat reguły stopu (Zandt i Townsend, 1993). Jeżeli stosunek nachyleń dla warunków pozytywnego i negatywnego jest nawet nieznacznie różny od $1:1$, konieczne jest przyjęcie bardzo egzotycznych założeń, żeby taki efekt dał się uzgodnić z regułą wyczerpującą, ale już jednakowe nachylenia można łatwo uzgodnić zarówno z regułą wyczerpującą jak i samowygaszającą. Analiza uogólnionionej przestrzeni alternatywnych modeli pozwala ustalić wiele na temat identyfikowalności systemu, jednak możliwe wyniki eksperymentu nadal związane są z asymetrią decyzyjną, a pewne kwestie (reguła stopu) okazują się łatwiejsze do empirycznego rozstrzygnięcia niż inne (równoległość-szeregowość).

Problem mimikry dla zagadnienia równoległości-szeregowości jest w istocie poważniejszy, niż do tej pory sugerowałem. Nie tylko liniowy efekt wielkości zestawu, ze współwystępującym efektem obecności bodźca docelowego lub bez takiego efektu, nie pozwala rozstrzygnąć między tymi alternatywami. Również obserwowany czasem, krzywoliniowy monotonicznie rosnący z przyspieszeniem ujemnym związek czasu reakcji z wielkością zestawu może być konsekwencją działania systemu równoległego albo szeregowego. Żeby system szeregowy dawał tego rodzaju predykcje, wystarczy aby tempa procesów indywidualnych malały odpowiednio wraz ze wzrostem liczby przetwarzanych elementów.

Bodaj jedynym, rzadko występującym w tego rodzaju eksperymentach wzorcem zależności, stanowiącym silne wsparcie dla modelu równoległego, jest brak efektu wielkości zestawu (średni czas reakcji w przybliżeniu taki sam dla każdego n). Aby uzyskać model szeregowy zgodny z brakiem efektu wielkości zestawu trzeba tak uzależnić tempa etapów od tej wielkości, że suma czasu wszystkich etapów będzie stała. Taki model jest dość osobliwy i niełatwo wyobrazić sobie warunki, w których mógłby być wiarygodny.

Jeżeli spełnione jest założenia o selektywnym wpływie na czas działania procesów indywidualnych, problem mimikry dla rozważanej tutaj ogólnej przestrzeni alternatyw okazuje się rozwiązywalny. W 1995 roku Dzhafarov i Schweickert przedstawili teorię identyfikacji pewnych prostych systemów, która została później uogólniona do współcześnie najbardziej uniwersalnej, jednak nadal, o ile mi wiadomo, nie wykorzystanej w praktyce teorii selektywnego wpływu Dzhafarova (1999, 2001).

Chociaż Townsend nie odwołuje się do teorii Dzhafarova i Schweickerta wyjaśniając sens metody podwójnego planu czynnikowego, teoria ta stanowi moim zdaniem dobry punkt wyjścia dla omówienia podstaw tej metody, a także zaproponowanych przez Townsenda i Nozawę (1995) metod analizy i projektowania badań, umożliwiających przeprowadzenie stosunkowo silnych testów empirycznych, dotyczących bogatego zbioru ogólnie zdefiniowanych klas modeli. Teoria Dzhafarova i Schweickerta jest interesującym przykładem tego, co można osiągnąć posługując się pojęciem selektywnego oddziaływania, a jej zaletą w kontekście tej pracy jest również i to, że w przeciwieństwie do nowszych propozycji Dzhafarova nie wymaga wprowadzenia bardziej zaawansowanych pojęć rachunku prawdopodobieństwa.

1.4. Problem mimikry na przykładzie architektury systemu przeszukiwania pamięci krótkoterminowej

1.5 Dzhafarova i Schweickerta teoria dekompozycji czasów reakcji

Jak wyjaśniłem już wcześniej, zastosowanie metody addytywnej polega w najprostszym przypadku na poszukiwaniu dwóch czynników, wpływających selektywnie na czas działania dwóch hipotetycznych procesów. Warto jednak rozważyć dowolne zmienne losowe, reprezentujące nie tylko czas działania procesów, ale także inne ich właściwości. Selektywny wpływ będzie wtedy dotyczył rozkładu tych zmiennych, a obserwowane efekty będą wynikiem działania hipotetycznej reguły kombinacji.

Niech $T_{\alpha\beta}$ oznacza zmienną, której rozkład zależy od wartości przyjmowanych przez czynniki eksperymentalne α i β, selektywnie wpływające na rozkłady zmiennych losowych A i B. Zastosowanie metody czynników addytywnych sprowadza się wtedy do przypadku, gdy zmienne losowe A i B są stochastycznie niezależne dla wszystkich wartości obu czynników i reprezentują czas działania odpowiednich procesów, zmienna T odpowiada natomiast, pomijając bez szkody dla rozumowania procesy resztowe, czasowi reakcji:

$$T_{\alpha\beta} = A(\alpha) + B(\beta) \tag{1.5}$$

Operację sumowania można, ale nie trzeba, interpretować w kategoriach czasów działania następujących po sobie procesów składowych. Zastępując operację sumowania przez minimum uzyskujemy system równoległy samowygaszający, a w przypadku operacji maksimum system równoległy wyczerpujący.

Dzhafarov i Schweickert rozwinęli pojęcie selektywnego wpływu wzdłuż dwóch wymiarów. Po pierwsze, rozważyli niemal dowolne operacje łączne i przemienne, które nazwali operacjami prostymi[12]. Do operacji prostych należą sumowanie, minimum i maksimum, ale też mnożenie i nieskończenie wiele innych. Po drugie, zaproponowana przez nich teoria dopuszcza zarówno możliwość niezależności jak i „doskonałej zależności stochastycznej" między zmiennymi A i B.

Najważniejszym rezultatem analiz obu autorów jest ustalenie, że dla dowolnych operacji prostych można skonstruować test, którego spełnienie stanowi warunek konieczny, aby zmienna losowa T dała się odpowiednio rozłożyć na selektywnie modyfikowalne komponenty A i B. Udało im się również stwierdzić, w jakich warunkach pomyślny rezultat takiego testu stanowi warunek wystarczający, aby zmienna T dała się odpowiednio zdekomponować, a także w jakich warunkach reguła kombinacji jest identyfikowalna unikalnie na podstawie pomyślnego wyniku testu ze względu na rozważany zbiór reguł dopuszczalnych.

[12]Dokładniej, zastosowali procedurę odkrytą przez Aczéla (1966), umożliwiającą wyprowadzenie wszystkich możliwych operacji łącznych i przemiennych z operacji dodawania, minimum i maksimum. Przyjęli też pewne dodatkowe założenia, bez których klasa tych operacji byłaby zbyt obszerna, jak na potrzeby dowodzonych twierdzeń.

Przy założeniu niezależności stochastycznej okazało się, że udany wynik testu jest w ogólnym przypadku warunkiem koniecznym, ale niewystarczającym dla istnienia odpowiedniej architektury. Inaczej mówiąc, wynik testu może być pomyślny nawet wtedy, gdy zmienna T nie daje się odpowiednio rozłożyć na zmienne A i B. Jednocześnie jednak, przyjmując dosyć łagodne założenia co do dopuszczalnych operacji, żadne dwa testy, skonstruowane dla dwóch różnych operacji prostych, nie mogą być spełnione jednocześnie, ani w sytuacji niezależności stochastycznej, ani doskonałej zależności stochastycznej. Ze względu na rozważany zbiór operacji łącznych i przemiennych reguła kombinacji jest wobec tego unikalnie wyznaczona przez wynik testu. Dodatkowo, przy założeniu doskonałej zależności stochastycznej pomyślny wynik testu jest warunkiem wystarczającym istnienia odpowiedniej „architektury". Okazuje się także, że dwa testy, zaprojektowane dla tej samej operacji ale dwóch różnych rodzajów zależności, mogą być oba spełnione pomyślnie. Należy zatem pamiętać, że zarówno związek stochastyczny między A i B, jak i selektywność wpływu są tutaj zakładane, a nie testowane. Dla pełnego planu czynnikowego z dwoma poziomami obu czynników (1.5) można przedstawić jako układ równań:

$$\begin{aligned} T_{11} &= A_1 + B_1 \\ T_{12} &= A_1 + B_2 \\ T_{21} &= A_2 + B_1 \\ T_{22} &= A_2 + B_2 \end{aligned} \tag{1.6}$$

Jeżeli teraz zmienne $A_i, B_i, i = 1, 2$ są wzajemnie stochastycznie niezależne:

$$\begin{aligned} (A_1 + B_1) + (A_2 + B_2) &= T_{11} + T_{22} \\ (A_1 + B_2) + (A_2 + B_1) &= T_{12} + T_{21} \\ (A_1 + B_1) + (A_2 + B_2) &= (A_1 + B_2) + (A_2 + B_1) \\ T_{11} + T_{22} &= T_{12} + T_{21} \end{aligned} \tag{1.7}$$

Ostatnie równanie jest warunkiem koniecznym dla zachodzenia reguły sumowania. Zależność wyrażona jest wyłącznie w kategoriach obserwowalnej zmiennej losowej T, dlatego można próbować ustalić, czy równanie to jest spełnione, przeprowadzając odpowiedni test empiryczny. Wyprowadzenie tego równania wymaga jedynie, aby operacja (tutaj akurat sumowanie) była łączna i przemienna, a więc tak samo można skonstruować test dla dowolnej innej operacji łącznej i przemiennej, na przykład minimum, maksimum, mnożenia i innych. Autorzy podają przykłady, ilustrujące w jaki sposób można nadać wielu takim regułom kombinacji sensowną interpretację w kategoriach własności nieobserwowalnych procesów przetwarzania informacji.

W praktyce zastosowanie teorii Dzhafarova i Schweickerta wiąże się z pewnymi trudnościami, wynikającymi z ograniczeń metod wnioskowania statystycznego. Opisana wyżej metoda konstrukcji testu dekompozycji prowadzi do ustalenia warunku konieczne-

go, który musi być spełniony, aby zmienna zależna dała się rozłożyć w odpowiedni sposób na selektywnie modyfikowalne komponenty. Warunek ten wyraża identyczność rozkładów obserwowalnych zmiennych losowych. Przeprowadzenie testu istotności wymaga, aby warunek identyczności rozkładów był reprezentowany przez hipotezę zerową, a nie hipotezę alternatywną, ponieważ ta pierwsza, a nie ta ostatnia wyraża brak różnic. Wsparcie dla danej reguły kombinacji będzie więc tym *mniejsze*, im wyższa będzie istotność testu.

Jeżeli tylko założenia modelu statystycznego są spełnione, wynik istotny pozwala z ustalonym prawdopodobieństwem błędu odrzucić hipotezę zerową, czyli na przykład stwierdzić, że jeden lub więcej parametr modelu jest różny od zera, albo że wartości parametrów różnią się między sobą. Niestety, sam wynik nieistotny, bez dodatkowej informacji o mocy statystycznej, nie pozwala na wyciągnięcie żadnych wniosków poza stwierdzeniem, że efektu nie udało się zaobserwować. Brak wyniku istotnego może być skutkiem tego, że efekt w populacji nie występuje, albo tego, że efekt występuje, jednak zbiór danych okazał się za mały do jego wykrycia. Badanie mające na celu identyfikację architektury ma charakter podstawowy, a nie aplikacyjny, wobec czego nawet istnienie bardzo małego efektu ma znaczenie zasadnicze. Częściowym rozwiązaniem tej trudności jest zastosowanie oszacowania przedziałowego dla wartości kontrastu skonstruowanego w oparciu o test dekompozycji. Na przykład, dla reguły sumowania wartość oczekiwana rozkładu:

$$T_{12} + T_{21} - (T_{11} + T_{22}) \tag{1.8}$$

jest równa zero. Pozostaje więc obliczyć na podstawie zebranych danych wartość odpowiedniego kontrastu. Jeżeli zero znajdzie się pomiędzy przedziałami ufności dla tego kontrastu, to im mniejsze będą te przedziały, tym silniejsze będzie wsparcie dla hipotezy o występowaniu reguły sumowania[13].

Sytuacja staje się bardziej kłopotliwa, gdy zmienna T reprezentuje czas reakcji. Zastosowanie ogólnego modelu liniowego jest wtedy problematyczne, ponieważ wiadomo, że standardowe metody wnioskowania oparte na modelu liniowym są wyjątkowo mało odporne na odstępstwa od założeń co do rozkładu (Wilcox, 1998, 2009). W przypadku czasów reakcji bardzo często obserwuje się różnice w wariancjach między warunkami, zwykle wariancja czasów reakcji rośnie wraz z trudnością zadania (Wagenmakers i Brown, 2007), zawsze pojawi się też mniejsza lub większa liczba obserwacji odstających. W konsekwencji moc statystyczna jest niska lub bardzo niska i zarówno wyniki testów istotności, jak też obliczone na podstawie danych przedziały ufności mogą być obarczone rozmaitymi artefaktami. Niestety, wiele tak zwanych metod odpornych (Huber, 2004; Wilcox, 2009), pozwalających rozwiązać częściowo problem niejednorodności wariancji, obserwacji odstających o dużym wpływie i niesymetryczności rozkładu czasów reakcji, polega na szacowaniu czegoś innego niż średnia arytmetyczna, na przykład

[13]Pojawiają się tutaj pewne trudności, wynikające z zakładanej interpretacji prawdopodobieństwa, o czym będę jeszcze pisał w rozdziale trzecim.

średniej odciętej, mediany, albo jeszcze innego estymatora, co uniemożliwia zastosowanie tych metod do rozstrzygania między regułami kombinacji zgodnie z teorią Schweickerta i Dzhafarova.

1.6 Townsenda metoda podwójnego planu czynnikowego

Metoda podwójnego planu czynnikowego Townsenda polega na zastosowaniu dwóch rodzajów manipulacji - selektywnego (z założenia) oddziaływania na rozkład czasów zakończenia hipotetycznych procesów i manipulacji liczbą i pozycją elementów docelowych. W wersji podstawowej metody zestaw składa się z dwóch elementów i stosowane są dwa czynniki α i β, o dwóch poziomach każdy, o których to czynnikach zakłada się, że oddziałują selektywnie na czas działania procesów A i B. Dopuszcza się możliwość wystąpienia elementu docelowego na każdej z dwóch pozycji, na obu pozycjach lub na żadnej, co daje razem 16 warunków. Udzielenie odpowiedzi na temat architektury systemu zależy przede wszystkim od wyników uzyskanych w tych warunkach, w których oba elementy są elementami docelowymi. Warunki te można rozpatrywać łącznie jako osobny eksperyment, ze zrozumiałych względów nazywany paradygmatem redundantnych elementów docelowych.

Zastosowanie zestawu samych elementów docelowych umożliwia rozwiązanie dwóch ważnych problemów. Po pierwsze, nie wymusza przetwarzania wyczerpującego, może więc zadziałać ewentualna samowygaszająca reguła stopu. Po drugie, stała wielkość zestawu, składającego się wyłącznie z elementów docelowych, pozwala testować hipotezy architektoniczne niezależnie od kwestii wydajności.

Paradygmat redundantnych elementów docelowych dla $n = 2$ jest szczególnym przypadkiem zastosowania teorii dekompozycji Dzhafarova i Schweickerta, co gwarantuje między innymi, że na podstawie wyników tego eksperymentu można rozróżnić architektury szeregową wyczerpującą (sumowanie) i równoległą, a w przypadku równoległej da się ustalić regułę stopu (minimum albo maksimum). Oryginalna teoria wspomnianych autorów nie uwzględnia jednak możliwości niecałkowitej zależności stochastycznej między czasami procesów składowych, nie pozwala również rozróżnić między regułami samowygaszającą i wyczerpującą w przypadku systemów szeregowych. Townsendowi i Nozawie (1995) udało się udowodnić kilka ważnych twierdzeń dotyczących systemów szeregowych, równoległych i koaktywacyjnych, przy założeniu możliwej zależności stochastycznej między procesami dla reguł stopu samowygaszającej i wyczerpującej.

Ze względów praktycznych schemat wnioskowania Townsenda opiera się na zastosowaniu kontrastu innego niż (1.8), konkretnie:

$$\bar{RT}_{22} - \bar{RT}_{21} - (\bar{RT}_{12} - \bar{RT}_{11}) \tag{1.9}$$

gdzie $\bar{RT}_{ij}$ to średni czas reakcji w danym warunku, a $i = 1, 2$ i $j = 1, 2$ to odpo-

wiednio poziomy czynników α i β. Odtąd poziom 2 będzie zawsze oznaczał sytuację, gdy odpowiedni proces przebiega *wolniej* na skutek selektywnego oddziaływania danego czynnika. Zgodnie z tą konwencją, wartość szacowana przez kontrast (1.9) określa, w jaki sposób spowolnienie wywołane czynnikiem β zależy od spowolnienia wywołanego czynnikiem α (spowolnienie rośnie, maleje, albo pozostaje takie samo). Gdy wartość (1.9) jest dodatnia, spowolnienie wywołane czynnikiem β jest większe gdy czynnik α również działa spowalniająco. Gdy wartość (1.9) jest ujemna, spowolnienie wywołane przez β jest mniejsze gdy α działa spowalniająco, wreszcie gdy wartość (1.9) jest równa zeru, spowolnienie wywołane przez β działa niezależnie od α.

Przypuśćmy, że pozostałe procesy zaangażowane w wykonanie zadania (R) zachodzą albo przed, albo po procesach A i B, a czynniki α i β nie wpływają na te procesy. Oznaczając przez T_{ij} sumę, minimum, albo maksimum zmiennych A_i i B_j, wielkość szacowaną przez kontrast (1.9) dana jest przez:

$$(E(T_{22}) + E(R)) - (E(T_{21}) + E(R)) - [(E(T_{12}) + E(R)) - (E(T_{11}) + E(R))]$$

Jak łatwo sprawdzić, $E(R)$ nie ma wpływu na wartość kontrastu, dlatego na razie nie będę więcej wspominał o procesach resztowych.

Najważniejsze twierdzenia dotyczące oczekiwanej wartości kontrastu (1.9) w systemach równoległych i szeregowych można udowodnić nie przyjmując żadnych założeń na temat rozkładów A i B. Zamiast jednak cytować, sformułowane przez innych autorów, ogólniejsze dowody, zdecydowałem się przeprowadzić analizę dla modeli z rozkładami wykładniczymi. Dzięki temu mogę skorzystać z ustalonych wcześniej własności tych modeli i w prostszy sposób wyjaśnić, dlaczego kontrast (1.9) pozwala na identyfikację architektury i reguły stopu.

W przypadku modelu wykładniczego selektywny wpływ czynników α i β sprowadza się do zmiany temp odpowiednich procesów, to jest v_A i v_B. Niech v_{Ai} i v_{Bj} będą tempami procesów A i B, gdy czynnik α przyjmuje poziom i, a czynnik β poziom j. Zgodnie z przyjętą wcześniej konwencją, poziom, dla którego proces przebiega wolniej, będę oznaczał jako poziom drugi, wobec czego $v_{A2} < v_{A1}$ i $v_{B2} < v_{B1}$, a więc też $1/v_{A2} > 1/v_{A1}$ i $1/v_{B2} > 1/v_{B1}$. Dla systemu równoległego niezależnego z regułą samowygaszającą kontrast (1.9) staje się oszacowaniem wielkości:

$$E\left(\min(A_2, B_2)\right) - E\left(\min(A_2, B_1)\right) - \left[E\left(\min(A_1, B_2)\right) - E\left(\min(A_1, B_1)\right)\right]$$

Gdy system równoległy jest samowygaszający i zestaw składa się wyłącznie z elementów docelowych, czas zakończenia przetwarzania to czas, w którym jeden z procesów zakończył działanie jako pierwszy, czyli minimum wartości dwóch zmiennych losowych. W takim razie rozkład czasów zakończenia całego procesu w każdym z czterech rozważanych warunków będzie taki, jak rozkład czasów zakończenia etapu pierwszego w systemie równoległym niezależnym wyczerpującym. Na podstawie (1.2) wiemy już, że będzie

to rozkład wykładniczy, którego tempo jest sumą temp procesów indywidualnych. Ponieważ wartość oczekiwana zmiennej wykładniczej jest odwrotnością jej tempa, wielkość szacowaną przez kontrast (1.9) można przedstawić jako:

$$\frac{1}{v_{A2}+v_{B2}}-\frac{1}{v_{A2}+v_{B1}}-\left(\frac{1}{v_{A1}+v_{B2}}-\frac{1}{v_{A1}+v_{B1}}\right)$$

$$=\frac{v_{A2}+v_{B1}-(v_{A2}+v_{B2})}{(v_{A2}+v_{B2})(v_{A2}+v_{B1})}-\frac{v_{A1}+v_{B1}-(v_{A1}+v_{B2})}{(v_{A1}+v_{B2})(v_{A1}+v_{B1})}$$

$$=\frac{v_{B1}-v_{B2}}{v_{A2}v_{A2}+v_{A2}v_{B1}+v_{A2}v_{B2}+v_{B1}v_{B2}}-\frac{v_{B1}-v_{B2}}{v_{A1}v_{A1}+v_{A1}v_{B1}+v_{A1}v_{B2}+v_{B1}v_{B2}}$$

Każdy kolejny iloczyn znajdujący się w mianowniku po lewej stronie, za wyjątkiem ostatniego, jest mniejszy, niż odpowiadający mu iloczyn po stronie prawej - ostatnie iloczyny są równe. Oczekiwana wartość kontrastu jest zatem większa od zera, czyli oba czynniki oddziaływują superaddytywnie. Przeprowadzając podobne rozumowanie dla systemu równoległego wyczerpującego z procesami niezależnymi można wykazać, że oczekiwana wartość kontrastu będzie mniejsza od zera, czynniki będą więc oddziaływały subaddytywnie.

Pozostaje jeszcze przypadek systemu szeregowego samowygaszającego. Oczekiwany czas zakończenia przetwarzania będzie wynosił $1/(pv_{Ai}+(1-p)v_{Bj}),\ i=1,2,\ j=1,2,\ 1\geqslant p\geqslant 0$, gdzie p to prawdopodobieństwo, że proces A zakończył działanie jako pierwszy. Rozważany kontrast będzie wtedy oszacowaniem wielkości:

$$\begin{aligned}&\left(p\frac{1}{v_{A2}}+(1-p)\frac{1}{v_{B2}}\right)-\left(p\frac{1}{v_{A2}}+(1-p)\frac{1}{v_{B1}}\right)\\&\quad-\left(\left(p\frac{1}{v_{A1}}+(1-p)\frac{1}{v_{B2}}\right)-\left(p\frac{1}{v_{A1}}+(1-p)\frac{1}{v_{B1}}\right)\right)\\&=(1-p)\left(\frac{1}{v_{B1}}-\frac{1}{v_{B2}}\right)-(1-p)\left(\frac{1}{v_{B1}}-\frac{1}{v_{B2}}\right)\end{aligned}$$

Dla architektur szeregowych oczekiwana wartość kontrastu wynosi więc zero niezależnie od reguły stopu. Oszacowanie kontrastu (1.9) na podstawie wyników eksperymentu z redundantnymi elementami docelowymi pozwala na empiryczne rozstrzygnięcie między architekturami szeregową i równoległą, a w przypadku równoległej umożliwia dodatkowo identyfikację reguły stopu, przy założeniu, że procesy są niezależne i spełniony jest warunek selektywnego wpływu na czas działania obu procesów. Można udowodnić (J. T. Townsend i Nozawa, 1995), że własności omawianego kontrastu są takie same, gdy procesy A i B są stochastycznie zależne.

1.6.1 Zastosowanie kontrastu interakcyjnego funkcji przeżyciowej do wyników eksperymentu z redundantnymi elementami docelowymi

W 1995 roku Townsend i Nozawa przedstawili dowody dosyć zaskakujących, również dla samych autorów (Townsend 2009, korespondencja osobista), wyników analiz, dotyczących konsekwencji założeń architektonicznych dla czasów reakcji w eksperymencie z redundantnymi elementami docelowymi. Udało im się wykazać, że systemy szeregowe, równoległe i koaktywacyjne, zależne lub niezależne, z regułą wyczerpującą i samowygaszającą, prowadzą do jakościowo różnych predykcji dla kontrastu interakcyjnego funkcji przeżyciowej, przy założeniu, że czynniki bezpośrednio selektywnie wpływające na procesy indywidualne oddziaływują na odpowiednio mocnym poziomie na rozkłady czasów zakończenia tych procesów.

Udowodnione przez Townsenda i Nozawę twierdzenia są niezależne od rozkładu procesów indywidualnych. Ponieważ twierdzenia te dotyczą obserwowalnych konsekwencji na poziomie rozkładu czasów reakcji (konkretnie, funkcji przeżyciowej), pozwalają ominąć niektóre z wymienionych wcześniej problemów, związanych z zastosowaniem standardowych metod wnioskowania statystycznego.

Dowody twierdzeń dotyczących kontrastu interakcyjnego funkcji przeżyciowej opierają się na kilku dosyć ostrożnych założeniach. Po pierwsze, rozkład czasów procesów resztowych f_R jest niezależny od poziomów czynników α i β. Po drugie, dopuszczana jest zależność rozkładu czasów zakończenia każdego z procesów od czasu zakończenia drugiego procesu i od czasu procesów resztowych, wobec czego łączny rozkład prawdopodobieństwa czasów zakończenia obu procesów dany jest przez:

$$f_{ij}(t_A, t_B|t_R) = f_{A(i)j}(t_A|t_B, t_R) f_{iB(j)}(t_B|t_R)$$

gdzie i i j oznaczają poziomy czynników α i β odpowiednio, f_{ij} to łączny rozkład prawdopodobieństwa czasu zakończenia obu procesów, $f_{A(i)j}$ to rozkład czasu zakończenia procesu A, a $f_{iB(j)}$ to rozkład czasu procesu B. Zakłada się dalej, że rozkład procesu A nie zależy bezpośrednio od poziomu czynnika β, a rozkład procesu B nie zależy bezpośrednio od poziomu czynnika α, chociaż oba mogą zależeć od czasu zakończenia drugiego procesu (dopuszcza się możliwość wpływu pośrednio nieselektywnego) i od czasu procesów resztowych:

$$f_{1B(j)}(t_B|t_R) = f_{2B(j)}(t_B|t_R)$$
$$f_{A(i)1}(t_A|t_R, t_B) = f_{A(i)2}(t_A|t_R, t_B)$$

Inaczej mówiąc, dopuszczalna jest zależność czasu zakończenia procesu A od czasu zakończenia procesu B, ale zależność ta ma być taka sama dla obu poziomów β (tak samo dla procesu B i czynnika α). Wreszcie, zakłada się też, że czynniki α i β wpływają na

rozkłady odpowiednich zmiennych w taki sposób, że:

$$\begin{aligned} S_{A1}(t_A|t_B,t_R) &< S_{A2}(t_A|t_B,t_R) \\ S_{B1}(t_B|t_A,t_R) &< S_{B2}(t_B|t_A,t_R) \end{aligned} \tag{1.10}$$

S oznacza tutaj funkcję przeżyciową, to jest $S(x) = 1 - F(x) = 1 - P(X \leqslant x) = P(X > x)$, gdzie $F(x)$ to dystrybuanta. Pamiętając o przyjętej konwencji, zgodnie z którą poziom drugi odpowiedniego czynnika to poziom, przy którym selektywnie modyfikowany proces przebiega wolniej, nierówności (1.10) oznaczają, że gdy czynnik α (lub β) przyjmuje poziom drugi (spowalniający), dla każdego momentu t_A (lub t_B) prawdopodobieństwo zakończenia procesu A (lub B) w momencie późniejszym niż t_A (lub t_B) jest większe, niż gdy czynnik przyjmuje poziom pierwszy. O takich funkcjach przeżyciowych można powiedzieć, że są uporządkowane.

Uporządkowanie na poziomie funkcji przeżyciowych (a więc również dystrybuant) jest silniejszym założeniem, niż na przykład uporządkowanie na poziomie średnich. Uporządkowanie funkcji przeżyciowych implikuje uporządkowanie średnich, ale nie odwrotnie (J. T. Townsend, 1990b)[14]. Niektóre interesujące twierdzenia udowodnione przez Townsenda i Nozawę wymagają jeszcze silniejszego założenia, to jest takiego uporządkowania, że rozkłady przecinają się dokładnie w jednym punkcie, co z kolei implikuje uporządkowanie funkcji przeżyciowych, ale nie odwrotnie.

Zamiast w kategoriach oczekiwanej wartości zmiennej T_{ij}, reprezentującej czas zakończenia całego procesu, kontrast (1.9) można wyrazić w kategoriach funkcji przeżyciowej $S_{ij}(t)$:

$$S_{22}(t) - S_{21}(t) - [S_{12}(t) - S_{11}(t)] \tag{1.11}$$

Wartość kontrastu (1.11) będzie oczywiście w ogólnym przypadku różna dla różnych wartości t. Dla lepszego zrozumienia przedstawionych dalej wyników warto przyjrzeć się bliżej, jak kontrast ten zachowuje się dla kilku prostych przypadków.

Dla systemu równoległego samowygaszającego z czasami niezależnymi wartość funkcji przeżyciowej czasu zakończenia całego procesu t to prawdopodobieństwo, że żaden z procesów nie zakończył działania do momentu t, czyli, pomijając dla uproszczenia rolę procesów resztowych, $S_{ij}(t) = P(T_{Ai} > t, T_{Bj} > t)$. Ponieważ procesy są niezależne $P(T_{Ai} > t, T_{Bj} > t) = P(T_{Ai} > t)P(T_{Bj} > t) = S_{Ai}(t)S_{Bj}(t)$, wobec czego kontrast (1.11) jest równy:

$$\begin{aligned} &S_{A2}(t)S_{B2}(t) - S_{A2}(t)S_{B1}(t) - [S_{A1}(t)S_{B2}(t) - S_{A1}(t)S_{B1}(t)] \\ &= [S_{A2}(t) - S_{A1}(t)][S_{B2}(t) - S_{B1}(t)] \end{aligned}$$

Z (1.10) wynika, że oba człony ostatniego iloczynu są większe od zera, a więc wartość kontrastu też jest większa od zera dla każdego t.

[14]Publikacja Townsenda z 1990 roku zawiera o ile mi wiadomo jak dotąd ostateczną wersję zaproponowanej przez niego teorii uporządkowania rozkładów.

Dla systemu równoległego wyczerpującego, zależnego lub niezależnego, wartość funkcji przeżyciowej dla momentu t to prawdopodobieństwo, że jeden lub drugi proces nie zakończył działania do momentu t. Zdarzenia te nie są rozłączne, wobec czego:

$$\begin{aligned} S_{ij}(t) &= P(T_{Ai} > t \text{ lub } T_{Bj} > t) \\ &= P(T_{Ai} > t) + P(T_{Bj} > t) - P(T_{Ai} > t, T_{Bj} > t) \\ &= S_{Ai}(t) + S_{Bj}(t) - P(T_{Ai} > t, T_{Bj} > t) \end{aligned}$$

Na mocy wcześniejszych ustaleń $P(T_{Ai} > t, T_{Bj} > t) = S_{MINij}(t)$, gdzie $S_{MINij}(t)$ to funkcja przeżyciowa dla systemu równoległego samowygaszającego, o której wiemy już, że przybiera wartości dodatnie dla każdego t. Wystarczy teraz zauważyć, że wyrażenie $S_{Ai}(t) + S_{Bj}(t)$ zniknie na skutek zastosowania kontrastu, tak samo jak w przypadku zastosowania kontrastu interakcyjnego średniej do procesów szeregowych. W takim razie kontrast interakcyjny funkcji przeżyciowej będzie miał dla każdego t taką samą wartość, jak w przypadku systemu równoległego samowygaszającego, ale z przeciwnym znakiem, czyli będzie zawsze mniejszy od zera. Równie łatwo można wykazać, że dla systemu szeregowego samowygaszającego wartość tego kontrastu musi być równa zero.

Przytoczone wyżej dowody przedstawiłem dla zilustrowania schematu rozumowania. Jako że dowód twierdzenia dotyczącego systemu szeregowego wyczerpującego jest bardziej skomplikowany i w moim odczuciu niezbyt intuicyjny, pozwoliłem go sobie pominąć. W tabelce poniżej znajduje się sumaryczne zestawienie najważniejszych rezultatów analiz Townsenda i Nozawy.

System	Kontrast średniej	Kontrast funkcji przeżyciowej
Równoległy		
Samowygaszający	dodatni	dodatni
Wyczerpujący	ujemny	ujemny
Szeregowy		
Samowygaszający	zero	zero
Wyczerpujący	zero	ujemny $\rightarrow$ dodatni

Tabela 1.1: Wartości kontrastu średniej i kontrastu funkcji przeżyciowej dla systemów szeregowych i równoległych, samowygaszających i wyczerpujących, zależnych lub niezależnych, przy złożeniu bezpośredniej selektywności wpływu.

Jak można wywnioskować na podstawie powyższej tabelki, kontrast interakcji funkcji przeżyciowej umożliwia rozróżnienie systemów, które zachowują się w sposób nierozróżnialny na poziomie średnich, to jest systemów szeregowych samowygaszającego i wyczerpującego. Najciekawszy wniosek dotyczy systemu szeregowego wyczerpującego. Dla pewnej wartości $t^* > 0$, kontrast interakcji funkcji przeżyciowej przyjmuje wartości ujemne, jeżeli $t < t^*$ i dodatnie, jeżeli $t > t^*$.

1.6.2 Zastosowanie paradygmatu redundantnych elementów docelowych do zadania na przeszukiwanie pamięci krótkoterminowej Sternberga

Eksperyment Townsenda i Fifica (2004) jest o ile mi wiadomo pierwszą próbą zastosowania metody podwójnego planu czynnikowego do oryginalnej procedury Sternberga z 1966 roku. Bodźcami w tym eksperymencie były trigramy postaci spółgłoska-samogłoska-spółgłoska. Wykorzystano bodźce w języku serbskim, dlatego że w języku tym każda litera odpowiada osobnemu fonemowi. Dzięki tamu, manipulując podobieństwem fonetycznym, teoretycznie możliwe było selektywne oddziaływanie na łatwość porównywania zapisanych w pamięci elementów z bodźcem docelowym. Wprowadzenie dystraktorów o zróżnicowanym podobieństwie do bodźca docelowego może powodować trudne w interpretacji efekty interakcyjne w warunku pozytywnym, dlatego analiza wyników dotyczyła przede wszystkim warunku negatywnego. Tym samym reguła stopu, wymuszona w warunku negatywnym, nie mogła być zidentyfikowana.

Pięć osób, w tym trzy kobiety, biorące udział w eksperymencie wykonało po 128 prób w każdym z 44 bloków, co pozwoliło na zebranie dostatecznej liczby obserwacji, aby dokładność oszacowania funkcji przeżyciowej była stosunkowo wysoka. Podobnie jak w oryginalnej procedurze Sternberga, bodźce do zapamiętania prezentowane były przez 1,2 sekundy, przy czym zastosowano dwie długości czasu między prezentacją zestawu do zapamiętania a prezentacją bodźca testowego (ISI), to jest 2000 milisekund (tak samo jak w procedurze Sternberga) i 700 milisekund. Użyto dwóch różnych czasów ISI, ponieważ zdaniem niektórych autorów krótsze ISI mogą umożliwiać przetwarzanie równoległe, często obserwuje się też efekty świeżości - gdy ISI jest małe, ostatni prezentowany element jest często szybciej przetwarzany.

Uzyskane przez Townsenda i Fifica oszacowania kontrastu interakcji średniej i funkcji przeżyciowej wskazują wyraźnie na to, że jedna z osób przeszukiwała równolegle w obu warunkach ISI, a pozostałe cztery osoby przeszukiwały szeregowo w warunku *krótszego* ISI i równolegle w warunku dłuższego ISI. U tych samych osób przeszukiwanie przebiegało zatem czasami szeregowo, a czasami równolegle. Wyniki łącznej analizy dla wszystkich osób były z kolei zgodne z szeregowością przeszukiwania w obu warunkach, stanowiąc tym samym doskonały przykład znaczenia analizy danych na poziomie indywidualnym. Wnioski o tyle trudno zakwestionować, że opierają się na jakościowo różnych i dokładnych predykcjach, wyprowadzonych dla ogólnie zdefiniowanych klas modeli szeregowych i równoległych, wyczerpujących przestrzeń stosunkowo prostych i wiarygodnych alternatywnych modeli dyskretnych.

Wydaje się, że ograniczając się do rozważanego zbioru alternatywnych modeli, zastosowaną przez autorów formalną procedurę wnioskowania statystycznego wypada uznać za zbędną. Jeżeli zaobserwowany wzorzec jest jakościowo zgodny z jedną z dopuszczalnych alternatywnych hipotez i niezgodny z pozostałymi, hipoteza ta będzie najlepszym spośród ogółu rozważanych wyjaśnieniem uzyskanych wyników. Badanie to ma charak-

ter podstawowy, dlatego znaczenie oszacowanej wielkości efektu sprowadza się w zasadzie do kwestii mocy statystycznej, wymaganej do zreplikowania wyniku. Co więcej, znowu ograniczając się do rozważanego zbioru alternatyw, uzyskanie wysokiej zgodności z wzajemnie wykluczającymi się i dokładnymi predykcjami, ustalonymi na podstawie przytoczonych wcześniej analiz metateoretycznych, pozwala ominąć problem złożoności modelu, o którym będę mówił w rozdziale trzecim.

1.7 Podsumowanie

Teoria Townsenda sprawia wrażenie wzorcowej realizacji strategii 20 pytań. Rezultaty zastosowania metody podwójnego planu czynnikowego do przeszukiwania pamięci krótkoterminowej zostały po raz pierwszy opublikowane w roku dwutysięcznym czwartym, czyli trzydzieści osiem lat po opublikowaniu historycznego artykułu Sternberga. W ciągu trzydziestu dwóch lat od opublikowania pierwszych formalnych demonstracji problemu mimikry dla systemów równoległych i szeregowych przeprowadzono ogromną liczbę podobnych eksperymentów, a wiele z nich miało na celu udzielenie odpowiedzi na temat architektury. W oparciu o rozmaite przesłanki teoretyczne stosowano bodźce o różnej modalności, manipulowano czasem prezentacji, czasem między pojawieniem się ostatniego bodźca a prezentacją bodźca testowego, wymagano od osób badanych wykonywania jednocześnie dodatkowych zadań i stosowano wiele innych, mniej lub bardziej pomysłowych modyfikacji oryginalnej procedury (Glass, 1984)[15].

Za wyjątkiem opisanych w tej pracy eksperymentów, wnioski z przeważającej większości tego rodzaju badań nie przyczyniły się w żaden sposób do rozstrzygnięcia kwestii równoległości-szeregowości, ani w przypadku przeszukiwania pamięci krótkoterminowej, ani w przypadku przeszukiwania wzrokowego. Albo rezultaty analiz Townsenda, Ashby'ego i Nozawy bardzo powoli się upowszechniają, albo z jakiś powodów społeczność badaczy uważa, że można je zignorować. Zdaniem Logana (2002), demonstracja problemu mimikry dla architektury pamięci krótkoterminowej mogła być głównym powodem, dla którego problem ten „przestał być modny". Powrót zainteresowania zagadnieniem szeregowości-równoległości nastąpił dopiero znacznie później, wraz z rozwojem teorii i badań dotyczących przeszukiwania wzrokowego. Autorzy dwóch, należących współcześnie do najbardziej wpływowych teorii przeszukiwania wzrokowego, to jest teorii integracji cech w obiekt (Treisman i Gelade, 1980; Treisman i Gormican, 1988) i teorii przeszukiwania ukierunkowanego (Wolfe i in., 1989; Wolfe, 1994, 2007), mimo że odnoszą się wprost do rezultatów Townsenda i Ashby'ego, wydają się nimi zanadto nie przejmować i poprzestają na stwierdzeniu, że kwestia jest „złożona", a analizy Townsenda i innych „subtelne"[16].

[15]Problem szeregowości-równoległości przeszukiwania stał się do tego stopnia „niemodny", że nie udało mi się znaleźć bardziej aktualnego przeglądu.

[16]Konkretnie, Treisman i Wolfe do dzisiaj zdają się sądzić, że liniowy efekt wielkości zestawu można

Mimo licznych prób, nie udało mi się dotąd zrozumieć, dlaczego hipotezę szeregowości (lub równoległości) należałoby uznać za prostsze wyjaśnienie. Odwoływanie się do zbieżnych świadectw (ang. *converging evidence*), gdy takie udało się uzyskać dla danej hipotezy, nie wnosi wiele. Łączne wsparcie, wynikające z dowolnego zbioru rezultatów empirycznych nie zmienia się wraz z dodaniem rezultatu, który w żaden sposób nie rozstrzyga między rozważanymi alternatywami.

Historia badań dotyczących przeszukiwania pamięci krótkoterminowej jest jaskrawym przykładem komplikacji, jakie pojawiają się w sytuacji, gdy próby testowania hipotez dotyczących nieobserwowalnego mechanizmu działania procesów poznawczych przebiegają bez udziału formalnych, metateoretycznych analiz dotyczących podstawowych, ogólnych założeń, na których oparte są te hipotezy. Jeden z morałów tej historii, prawdopodobnie niezbyt zaskakujący, jest taki, że (zawsze częściowe) rozstrzygnięcie empiryczne niektórych ogólnych kwestii teoretycznych nie jest możliwe w izolacji i wymaga uwzględniania dodatkowych wymiarów zagadnienia. Mogłoby się wydawać, że nawet jeżeli w konkretnym przypadku drobiazgowa, formalna analiza przestrzeni alternatywnych hipotez nie jest konieczna, często nie da się tego stwierdzić, dopóki nie zostaną podjęte próby jej przeprowadzenia, a więc w zasadzie zawsze trzeba ją przeprowadzać.

Autorom opisanych w tym rozdziale badań można zarzucić, że nie uwzględniają jawnie mechanizmu kontroli, bez którego wykonanie zastosowanych zadań nie jest przecież możliwe. Należy jednak zauważyć, że hipotetyczny mechanizm kontroli wraz z innymi, bliżej nieokreślonymi, ale niewątpliwie zachodzącymi w przypadku tych zadań procesami, został uwzględniony nie wprost jako proces resztowy. Dzięki temu, że założenia dotyczące procesów resztowych zostały sformułowane ostrożnie, można było dokonać empirycznej oceny stosunkowo ogólnych i prostych hipotez, bez konieczności nieskrępowanego spekulowania na temat rozmaitych delikatnych kwestii.

Gdyby identyfikacja systemu poznawczego miała polegać na empirycznym rozstrzyganiu między wszystkimi alternatywami wyznaczonymi przez zbiór wymiarów, za pomocą których można opisać możliwe architektury, tworzenie modeli zintegrowanych byłoby rozwiązaniem niedorzecznym. Liczba kombinacji rośnie gwałtownie wraz z dodaniem każdego kolejnego wymiaru (szeregowość-równoległość × zależność-niezależność × dyskretność-ciągłość × wydajnosć, i tak dalej), dlatego wyczerpująca analiza metateoretyczna przestrzeni alternatyw dla modelu zintegrowanego wydaje się po prostu niemożliwa. Najznamienitsi zwolennicy stosowania modeli zintegrowanych twierdzą jednak (np. Newell, 1973, 1990; Anderson, 1988/1991c; Anderson i Lebiere, 1998; Anderson, 1990; Byrne, 2007; Gray, 2007), powołując się przy tym nieodmiennie na problem mimikry dla systemów szeregowych i równoległych, że próby tworzenia bardziej wyczerpujących modeli mają być rozwiązaniem rozsądnym właśnie ze względu na problem identyfikacji architektury. Dzięki temu, że jest względnie wyczerpujący, model zintegro-

uznać za *częściowe* wsparcie dla hipotezy przeszukiwania szeregowego.

wany ma pozwalać na uwzględnienie bogatszego, bardziej zróżnicowanego repertuaru wyników badań, z którymi predykcje modelu muszą się zgadzać, co ma rzekomo zwiększać wrażliwość poszczególnych hipotez, na których oparty jest taki model, na rezultaty testów empirycznych. Żeby zilustrować odmienne podejście do modelowania lokalnych procesów i struktur, w następnym rozdziale omówię historię badań, dotyczących mechanizmu odpowiedzialnego za wykonywanie stosunkowo prostych zadań wyboru z dwóch alternatyw w warunkach presji czasowej.

Rozdział 2

Proces wyboru ze skończonej liczby alternatyw w warunkach presji czasowej

W poprzednim rozdziale zajmowałem się między innymi analizą eksperymentów, w których główną zmienną zależną był czas reakcji. Jak już wspomniałem, jeżeli poprawność w tego rodzaju badaniach jest stosunkowo wysoka, reakcje błędne traktuje się zwykle jako zanieczyszczenia, odrzucane przed przeprowadzeniem właściwego wnioskowania statystycznego. Ludzie nie wykonują jednak zadań bezbłędnie i często sama poprawność jest głównym przedmiotem zainteresowania badacza.

Czas reakcji i poprawność są dwiema najważniejszymi zmiennymi zależnymi w psychologii poznawczej. Do zdecydowanie najczęściej stosowanych, ogólnych modeli, określających źródła błędu w stosunkowo prostych zadaniach, należą modele oparte na teorii detekcji sygnałów i modele kumulacji świadectw. Obie te klasy modeli są niewątpliwie lokalne w znaczeniu przyjętym w tej pracy, jednak powody, dla których w pewnych warunkach wydają się w przybliżeniu prawdziwe, różnią się zasadniczo od powodów, dla których uzasadnione mogą się wydawać opisane w poprzednim rozdziale wnioski Townsenda i Fifica.

2.1 Poprawność reakcji z perspektywy torii detekcji

Teoria detekcji sygnałów została po raz pierwszy zastosowana w psychologii dla wyjaśnienia zachowania się osób badanych, wykonujących zadania polegające na odróżnieniu trudno identyfikowalnego percepcyjnie bodźca od występującego w tle szumu. W klasycznej już pracy Green i Swets (1966) założyli, że osoby badane podejmują wtedy decyzje starając się optymalizować wykonanie w obecności częściowo nieprzewidywalnej zmienności. Autorzy opisali szczegółowo procedury eksperymentalne i metody analizy, które miały pozwolić na oszacowanie niezależnego wpływu czynników decyzyjnych i percepcyjnych. Rozróżnienie na czynniki decyzyjne i percepcyjne, a właściwie pozade-

cyzyjne czynniki związane z efektywnością procesu przetwarzania informacji, jest kluczowym elementem teorii detekcji.

Obecnie teoria ta wykorzystywana jest w badaniach dotyczących nie tylko procesów percepcyjnych, ale także pamięci (rozpoznawanie i odtwarzanie), uwagi, podejmowania decyzji eksperckich i innych, względnie elementarnych lub bardziej złożonych procesów poznawczych (Swets, 1996; Macmillan i Creelman, 2002), w tym również procesów podejmowania decyzji u zwierząt (np. Nachtigall, 1986) i decyzji grupowych (np. Sorkin, Hays i West, 2001). W konsekwencji poszerzenia obszaru zastosowań stało się jasne, że teoria znajduje zastosowanie nie tylko w przypadku tych zadań, które dają się naturalnie interpretować jako wymagające wykrywania trudno identyfikowalnych percepcyjnie sygnałów. Faktycznie, postulowany abstrakcyjny model przebiegu procesu przetwarzania informacji okazał się przydatny do analizy zachowania się ludzi i zwierząt w sytuacjach, które ogólnie można określić jako zadania wyboru lub identyfikacji w warunkach presji czasowej. Ta zmiana uzasadnia przyjętą przez autorów jednego z bardziej wyczerpujących podręczników dotyczących omawianych dalej modeli konwencję, aby posługiwać się określeniem „teoria detekcji" z pominięciem terminu „sygnałów" (Macmillan i Creelman, 2002).

W typowym zadaniu, do którego teoria detekcji daje się zastosować, prezentowane są bodźce, z których każdy należy do jednej ze z góry ustalonych klas. Osoba badana ma za każdym razem podjąć decyzję, do której klasy należy prezentowany bodziec i na tej podstawie wykonać odpowiednią reakcję. Na przykład, zadanie może polegać na udzielaniu odpowiedzi, czy jakieś bodziec znajdował się czy nie w prezentowanym wcześniej zestawie do zapamiętania, albo czy jest jaśniejszy czy ciemniejszy niż jednocześnie prezentowany bodziec wzorcowy. Odpowiedniość między bodźcami a ich klasami, narzucona przez eksperymentatora, dostarcza kryterium oceny poprawności obserwowanych reakcji.

Ogólnie, w teorii detekcji zakłada się, że reakcje błędne mają swoje źródło w nieuniknionej zmienności obecnej w stymulacji bodźcowej i w procesach zachodzących w umyśle obserwatora. Teoria określa, w jaki sposób błędy określonego rodzaju, takie jak fałszywe alarmy i ominięcia w zadaniach na detekcję, zależą od pewnych własności nieobserwowalnego procesu. Własności te, reprezentowane przez wolne parametry w konkretnym modelu, można oszacować na podstawie informacji o reakcjach błędnych. W sytuacji, gdy wszystkie reakcje są poprawne, teoria nie znajduje zastosowania. Jeżeli reakcje błędne występują, ale ich liczba jest stosunkowo mała w stosunku do liczby wszystkich reakcji obserwowanych, teoria może być bezużyteczna z powodu niedostatecznej mocy statystycznej.

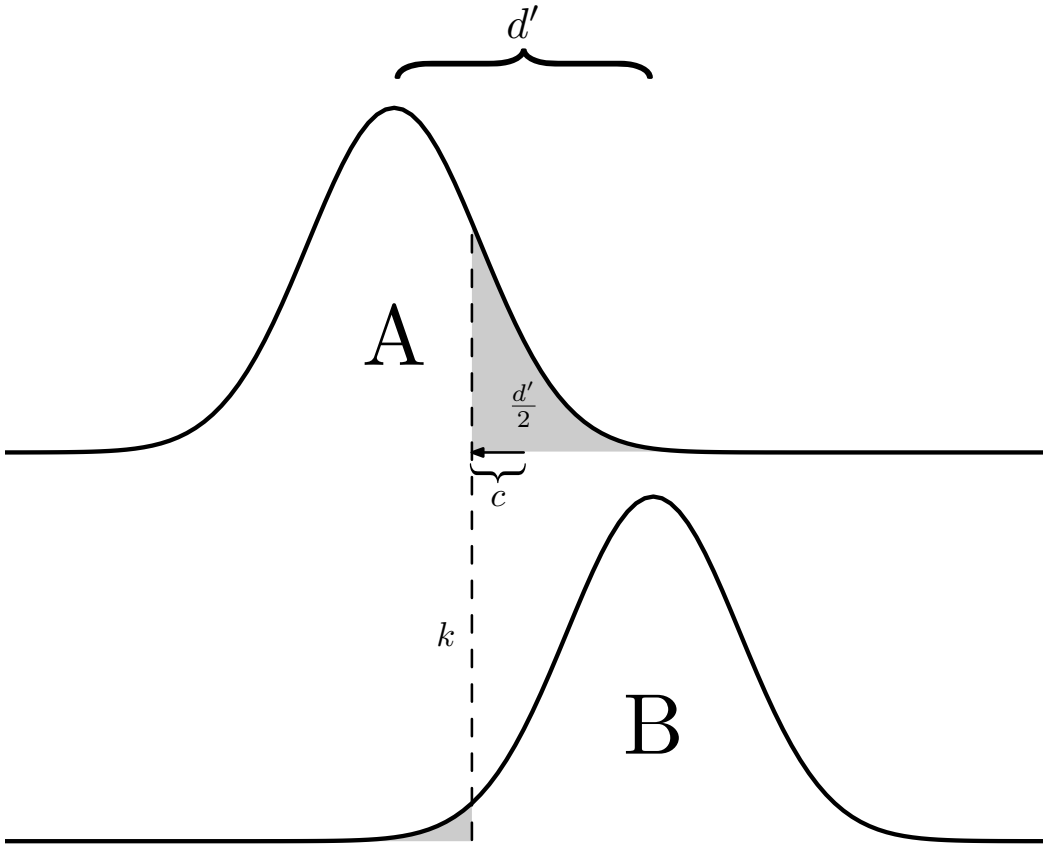
Dla przypomnienia omówię najprostszy schemat zadania eksperymentalnego, które można analizować za pomocą modelu detekcji. Składa się ono z serii prób, w trakcie których pojawiają się pojedyncze bodźce, należące do jednej z dwóch klas. Zadaniem osoby badanej jest podejmowanie w trakcie każdej próby decyzji, do której klasy należą prezentowane bodźce. W eksperymencie dotyczącym percepcji wzrokowej może to być obraz

o dużym lub małym kontraście, albo fragment twarzy kobiecej lub męskiej. Zadanie polega więc na rozróżnianiu między dwoma rodzajami bodźców. W przypadku, gdy jeden z bodźców można interpretować jako szum, powiemy, że zadanie takie dotyczy detekcji, jeżeli zaś oba bodźce można potraktować jako identyfikowalne sygnały, powiemy że osoba badana dokonuje rozpoznania.

Metody analizy wyników eksperymentu dotyczącego detekcji, rozróżniania i rozpoznawania są niezależne od modalności bodźców i charakteru ich klasyfikacji. Nie ma też znaczenia, czy według instrukcji osoba badana ma udzielać odpowiedzi typu „tak"-„nie", czy na przykład, „duży kontrast"-„mały kontrast", albo czy ma identyfikować konkretne egzemplarze. Dzięki zastosowaniu odpowiedniego modelu dane w postaci liczby błędów każdego rodzaju można interpretować w kategoriach rozróżnialności (ang. *sensitivity*) i tendencji do wybierania jednej z dwóch alternatyw (ang. *bias*).

Zakłada się, że w trakcie każdej próby, niezależnie od tego, jaki bodziec akurat się pojawił, powstaje pewien wewnętrzny sygnał, w najprostszym przypadku zdefiniowany formalnie jako jednowymiarowa ciągła zmienna losowa. O wartościach przyjmowanych przez tą zmienną ma decydować klasa bodźca. W sytuacji, gdy jeden z bodźców można interpretować jako szum, a drugi jako sygnał, zgodnie z najprostszą intuicją można założyć, że detektor będzie przyjmował zwykle mniejsze wartości w obecności szumu niż w obecności bodźca właściwego (można by go wtedy nazwać detektorem sygnału). Jeżeli zadanie polega na rozpoznawaniu, można założyć, że detektor przyjmuje zwykle wartości dodatnie w obecności bodźców z jednej kategorii i ujemne w obecności bodźców z drugiej kategorii. Teoria detekcji daje się zastosować niezależnie od założeń na temat położenia punktu zerowego skali. Liczy się tylko to, na ile odpowiednie rozkłady są rozsunięte, natomiast to, który z nich położony jest „wyżej", można rozstrzygnąć arbitralnie.

Częściowo z powodów teoretycznych, o których będzie jeszcze mowa później, a częściowo dla uproszczenia obliczeń, zwykle przyjmuje się, że w każdym z dwóch możliwych warunków wartość detektora pochodzi z jednego z dwóch różnych rozkładów normalnych. Powtórne pojawienie się bodźca należącego do tej samej klasy będzie w ogólnym przypadku prowadziło do powstania innych, choć skorelowanych wartości wewnętrznego sygnału. Abstrahuje się przy tym, od traktowanych jako komponent losowy, źródeł zmienności reprezentowanej przez oba rozkłady. Na znajdującym się poniżej diagramie przedstawiłem schematycznie główne założenia podstawowej wersji modelu:



Rysunek 2.1: Schemat podstawowego modelu detekcji dla dwóch klas bodźców

Gdyby rozkłady wartości detektora dla każdej z alternatyw w ogóle na siebie nie zachodziły, osoba badana mogłaby wykonać zadanie bezbłędnie. W typowym zadaniu, do którego teoria daje się zastosować, taki stan rzeczy jest jednak mało prawdopodobny. Zwykle, tak jak na zamieszczonym diagramie, odpowiednie rozkłady będą się pokrywały przynajmniej częściowo, osoba badana będzie więc zmuszona do przyjęcia pewnej wartości krytycznej k, która pozwoli podjąć decyzję, gdy wartość sygnału będzie się znajdowała w obszarze wspólnym dla obu rozkładów, czyli będzie niejednoznaczna.

Pomijając na razie kryterium, prawdopodobieństwo błędu zależy od stopnia, w jakim oba rozkłady się pokrywają, czego miarą jest zwykle odległość między odpowiednimi średnimi, oznaczana jako d'. Parametr d' jest miarą trudności w rozróżnianiu klas bodźca. Ponieważ skala nie ma tutaj znaczenia, dogodnie jest przyjąć, że oba rozkłady są standardowymi rozkładami normalnymi ($\sigma_A = \sigma_B = 1$). Wartość parametru d' oznacza wtedy liczbę odchyleń standardowych, o jaką oddalone są od siebie średnie obu rozkładów. Każda kombinacja wartości d' i k wyznacza jednoznacznie prawdopodobieństwa obu możliwych rodzajów błędu - obszary pod rozkładami prawdopodobieństwa odpowiadające tym błędom przedstawione są na diagramie kolorem szarym. Jeżeli przyj-

miemy, że $\mu_A = 0$, a więc $d' = \mu_B$, z elementarnego rachunku prawdopodobieństwa wynika:

$$p_{AB} = 1 - \Phi(k) = \Phi(-k)$$
$$p_{BB} = \Phi(k - d')$$

gdzie $p_{ij}, i, j = A, B$ to prawdopodobieństwo zaklasyfikowania bodźca i jako j, k to odległość wartości kryterium od średniej rozkładu A, a Φ to dystrybuanta standardowego rozkładu normalnego. Dysponując wartościami p_{AB} i p_{BB} można obliczyć $p_{BA} = 1 - p_{BB}$ i $p_{AA} = 1 - p_{AB}$. Odchylenie kryterium od punktu położonego dokładnie pomiędzy średnimi $c = k - d'/2$ można interpretować w kategoriach tendencji do przypisywania bodźcom jednej z dwóch kategorii. Jeżeli kryterium odchylone jest na lewo od tego punktu, występuje tendencja do przypisywania bodźcom kategorii B, jeżeli na prawo, kategorii A.

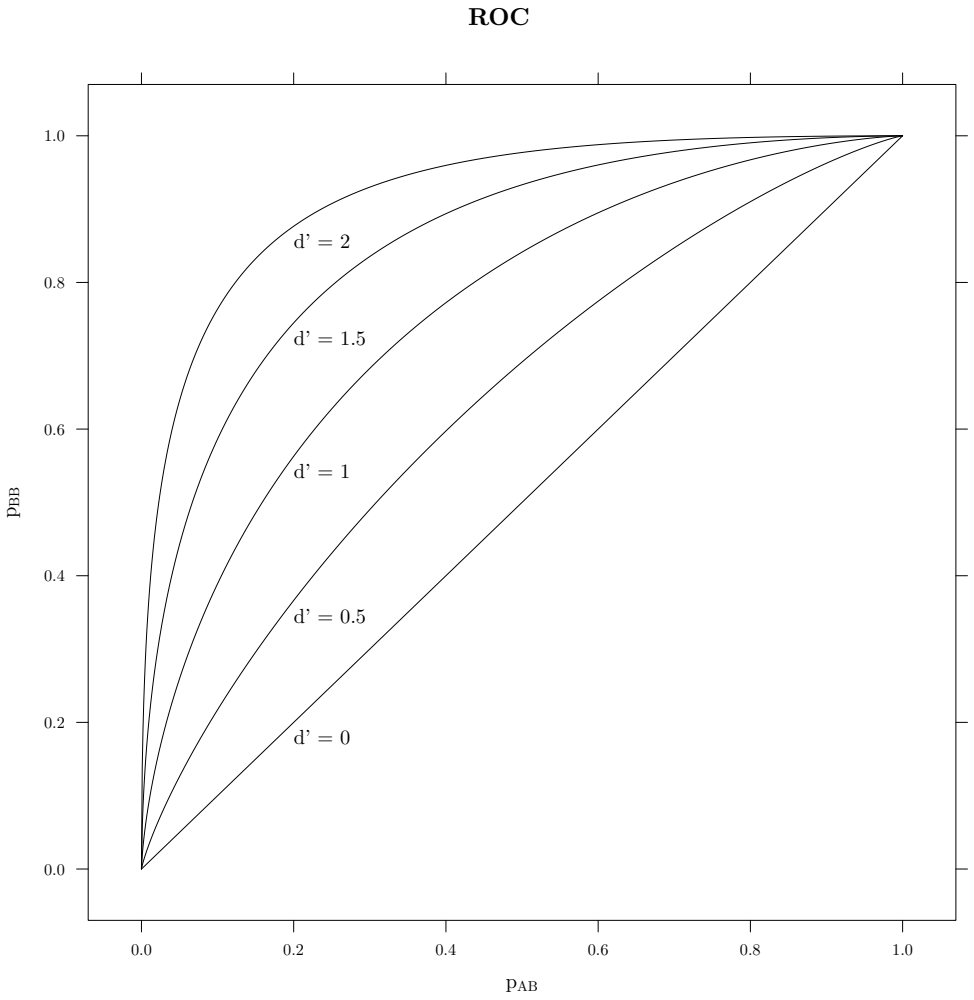
Kiedy w zadaniu obserwuje się oba rodzaje błędów, uwzględnienie tylko jednego z nich najczęściej nie jest poprawną miarą rozróżnialności. W pewnym zadaniu osoba badana może stosunkowo często błędnie klasyfikować bodźce typu A jako B. Bez dodatkowej informacji na temat tego, jak często osoba ta błędnie klasyfikuje B jako A, nie da się nic powiedzieć o trudności związanej z rozróżnianiem między klasami bodźca. Wartość p_{AB} dostarcza informacji na temat samego kryterium, a p_{BA} dostarcza informacji na temat związku między kryterium i rozróżnialnością, ale żadna z tych wielkości nie wystarcza do obliczenia d'. Jeżeli osoba badana popełnia wiele błędów typu AB, może to wynikać z niskiej rozróżnialności, z tendencji do klasyfikowania wszystkich bodźców jako B, albo z obu tych powodów.

Miary wykonania zadania oparte tylko na liczbie błędów jednego rodzaju, takie jak suma fałszywych alarmów, albo suma ominięć, są prawdopodobnie rzadziej stosowane. Mogłoby się wydawać, że zsumowanie obu rodzajów błędu rozwiązuje problem, skoro wynik sumowania zależy zarówno od przyjętego kryterium jak i od rozróżnialności. Chwila refleksji wystarczy jednak, aby stwierdzić, że miara oparta na sumie błędów nie wystarcza, ponieważ tą samą sumę można uzyskać na nieskończenie wiele sposobów, zmieniając odpowiednio wartości parametrów k i d'.

Dopiero zastosowanie teorii detekcji, o ile jej założenia są w przybliżeniu spełnione, pozwala uniknąć problemu pomylenia wpływu kryterium i rozróżnialności. Nieuchronnie pojawia się pytanie, jakie będą konsekwencje niespełnienia założeń teorii, takich jak założenie o jednakowej wariancji rozkładów, a także w jaki sposób można ustalić, na ile te założenia są spełnione. Okazuje się, że teoretycznie da się testować rozmaite hipotezy dotyczące różnic między rozkładami A i B, włączając w to założenie o jednorodności wariancji, które zresztą najczęściej spełnione nie jest (Swets, 1996; Macmillan i Creelman, 2002). Ograniczę się do omówienia zagadnienia testowania różnic w wariancjach, ponieważ testowanie hipotez dotyczących innych różnic między rozkładami, takich jak różnice w skośności, jeżeli dopuszczalne jest, żeby rozkłady nie były normalne, przebiega zwykle podobnie.

2.1.1 Testowanie założeń teorii detekcji dotyczących rozkładów na przykładzie różnic w wariancji

Aby wyjaśnić, skąd biorą się obserwowalne konsekwencje różnic w rozkładach, dobrze jest omówić dokładniej w jaki sposób, dla ustalonej wartości d', zmiana kryterium wpłynie na stosunek p_{BB}/p_{AB}, czyli stosunek poprawnego zaklasyfikowania bodźca B jako B do prawdopodobieństwa błędnego zaklasyfikowania bodźca A jako B. Ten związek przedstawia tak zwana krzywa ROC (ang. *Receiver Operating Characteristic*). Na następnym wykresie widoczne są cztery takie krzywe, z których każda odpowiada pewnej wartości d'. Interesuje nas teraz stosunek dwóch prawdopodobieństw, dlatego obszar na którym zaznaczyłem krzywe jest kwadratem jednostkowym. Każdy punkt na takiej krzywej odpowiada unikalnej wartości k. Dla $d' = 0$ bodźce są całkowicie nierozróżnialne i prawdopodobieństwa poprawnych i błędnych klasyfikacji bodźca jako A (lub B) są równe. Jak łatwo zauważyć, dla $d' > 0$ każdy wzrost prawdopodobieństwa poprawnej klasyfikacji BB wiąże się z kosztem w postaci wzrostu prawdopodobieństwa błędu AB, a koszty te rosną w miarę jak rośnie poprawność.



Rysunek 2.2: Krzywe ROC dla dwóch rozkładów normalnych o jednakowej wariancji i różnych wartości d'

Dla każdej wartości parametru d' i prawdopodobieństwa pojawienia się bodźca typu A (lub B) można wyznaczyć wartość kryterium, która będzie optymalna w tym znaczeniu, że będzie minimalizowała oczekiwaną sumę wszystkich błędów. W rozważanym tutaj przypadku, gdy prawdopodobieństwa pojawienia się bodźców należących do każdej z dwóch kategorii są jednakowe i rozkłady mają taką samą wariancję, optymalnym kryterium będzie $k = d'/2$. Wynika to wprost z kształtu rozkładu normalnego. Przesunięcie wartości k w jedną lub drugą stronę od miejsca znajdującego się dokładnie pośrodku między średnimi obu rozkładów powoduje, że zmniejsza się prawdopodobieństwo błędu jednego rodzaju, jednak szybciej rośnie prawdopodobieństwo błędu drugiego rodzaju.

Gdyby badacz miał pewność, że osoby badane będą przez cały czas wykonywania za-

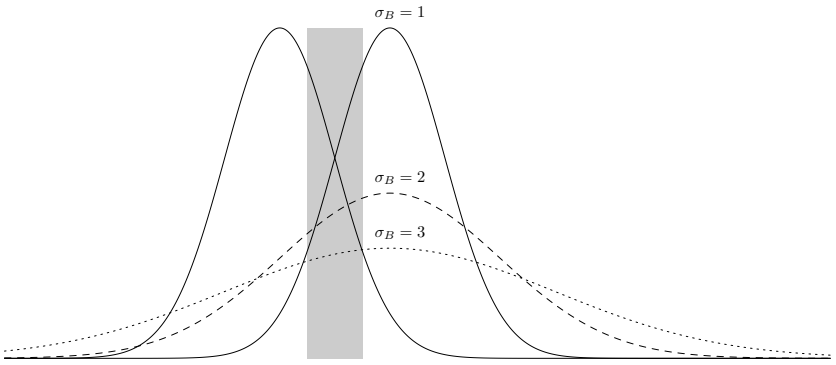
dania stosowały wartość optymalną, kryterium byłoby całkowicie zdeterminowane przez rozróżnialność bodźców i teoria detekcji byłaby znacznie mniej użyteczna. Na szczęście dla teorii, w wielu wypadkach taki scenariusz wypada uznać za niewiarygodny. Żeby od początku stosować optymalną wartość kryterium osoba badana musiałaby z góry znać i konsekwentnie uwzględnić odległość między rozkładami i prawdopodobieństwa pojawienia się bodźców każdego rodzaju, co wymagałoby od niej zdolności jasnowidzenia.

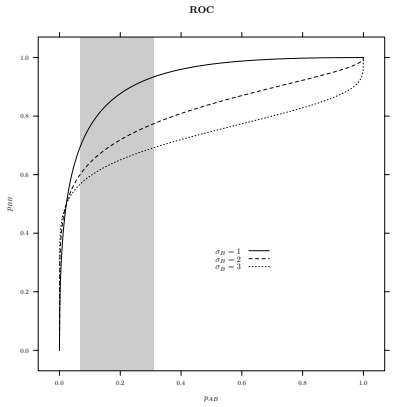
Kształt krzywych ROC jest dosyć nieoczywistą konsekwencją teorii detekcji. Dysponując odpowiednio dużym zbiorem danych można sprawdzić, na ile konsekwencje modelu wyrażone przez te krzywe zgadzają się z uzyskanymi wynikami. Krzywe zawierają informację na temat kształtu obu rozkładów, ich związek z zakładanym modelem jest więc stosunkowo silny. Inaczej mówiąc, duża zgodność obserwacji z przewidywanym kształtem krzywych stanowi silne wsparcie dla teorii detekcji, dlatego że trudno znaleźć równie proste, alternatywne wyjaśnienie dla tego rodzaju zależności.

Aby kształt krzywej ROC dał się w ogóle w przybliżeniu oszacować, badacz musi albo wpłynąć na zmianę wartości kryterium, nie zmieniając jednocześnie wartości pozostałych parametrów modelu, albo zebrać dane dla różnych wartości kryterium, zmieniającego się spontanicznie. Gdyby kryterium stosowane przez osobę badaną pozostawało niezmienne, dostępne byłoby oszacowanie tylko jednego punktu. Przy ustalonych i znanych wariancjach A i B, byłaby to informacja wystarczająca do oszacowania d', jednak teraz interesuje nas ocena trafności założeń na temat rozkładu, a do tego jeden punkt nie wystarczy.

Kształt krzywej zależy od d', k i kształtu obu rozkładów, dlatego zmiana wariancji rozkładu B wpłynie również na kształt tej krzywej. Tak jak w przypadku modelu podstawowego, skalę dla parametru d' można ustalić dowolnie, przyjmując jakiekolwiek założenie na temat wariancji rozkładu A. Przyjmowana przeze mnie tutaj, stosowana czasem konwencja $\sigma_A = 1$, ma jedynie zagwarantować porównywalność wyników pochodzących z różnych badań.

Na następnym wykresie widoczne są krzywe ROC odpowiadające różnym wartościom σ_B, dla $\sigma_A = 1$ i $d' = 2$. Na szaro zaznaczyłem wybrany podzbiór możliwych wartości kryterium ($k \in [0{,}5, 1{,}5]$). Punkty na krzywych ROC odpowiadają stosunkowi pola pod rozkładem B do pola pod rozkładem A, na prawo od punktu kryterium. Wartość p_{BB} rośnie wraz z przesuwaniem się kryterium w lewo. Im większa wariancja B, tym wolniej ta wartość będzie przyrastała, ale p_{AB} będzie się zmieniało dokładnie tak samo, niezależnie od wariancji rozkładu B. Krzywe odpowiadające różnym wariancjom rozkładu B wyraźnie różnią się kształtem.





Rysunek 2.3: Krzywe ROC dla różnych wariancji rozkładu B

Jeżeli teoretycznie wariancje obu rozkładów mogą się różnić, a trudno a priori wykluczyć taką ewentualność, oszacowania oparte na założeniu jednakowych wariancji mogą być błędne. Oszacowanie d' będzie wtedy częściowo zależało od prawdziwej rozróżnialności, a częściowo od różnic w wariancjach, które z kolei mogą być związane z interesującymi badacza zmiennymi. Wobec tego, za wyjątkiem sytuacji, w której można w sposób

uzasadniony z góry założyć równość wariancji, model uwzględniający dodatkowy wolny parametr σ_B będzie rozwiązaniem o wiele bardziej rozsądnym.

Dzięki temu, że kształt krzywych zależy od kształtu obu rozkładów, badacz może porównywać pod względem dopasowania do danych modele różniące się założeniami dotyczącymi rozkładów A i B. Charakter przyjmowanych założeń może sprawić, że ilościowa ocena alternatywnych modeli będzie wymagała stosunkowo dużej liczby obserwacji, jednak pomijając problem dostatecznej mocy statystycznej i techniczne trudności związane z uzyskaniem danych dla różnych wartości kryterium, w wielu wypadkach hipotezy te będą empirycznie testowalne.

Teoretycznie wydaje się całkiem prawdopodobne, że kryterium decyzyjne będzie związane z niektórymi wymiarami różnic indywidualnych, „semantycznymi" właściwościami bodźców, lub innymi charakterystykami zadań, warunkami w jakich przeprowadzane jest badanie, albo dowolną kombinacją wymienionych czynników. Ignorowanie problemu wpływu kryterium decyzyjnego z pewnością nie przyczyni się do zwiększenia konkluzywności wniosków. Teoria detekcji stanowi potencjalnie atrakcyjną alternatywę dla ateoretycznych, a przez to niezbyt użytecznych miar trudności wykonania zadań wyboru ze skończonej liczby alternatyw, takich jak średni czas reakcji, albo suma błędów obu rodzajów. Jej główną zaletą jest możliwość oszacowania jednocześnie rozróżnialności i stosowanego kryterium, o ile oczywiście założenia modelu są w przybliżeniu spełnione.

Często przedmiotem zainteresowania badacza jest trudność związana przede wszystkim z ogólnie rozumianym procesem rozróżniania. Na przykład, badacz zainteresowany zależnością między pojemnością pamięci roboczej a inteligencją może, podobnie jak faktycznie zrobiło to wielu jego poprzedników, podjąć próbę empirycznego ustalenia związku między wynikami uzyskanymi w wybranym teście inteligencji a poziomem wykonania jednego lub większej liczby zadań, wymagających wykorzystania „zasobów" tej pamięci. Osoby badane mogą się różnić poziomem wykonania zastosowanych zadań z powodów istotnych lub przygodnych ze względu na podjęty problem badawczy. Badacz życzyłby sobie w tym wypadku, aby różnice wynikały w możliwie największym stopniu z różnic w pojemności pamięci roboczej, a nie na przykład z różnic w subiektywnym prawdopodobieństwie wystąpienia bodźców. Gdyby jednak, podobnie jak wielu jego poprzedników, badacz poprzestał na odkryciu wysokiej korelacji między poprawnością wykonania (albo średnim czasem reakcji) niektórych z zastosowanych zadań a wynikami w teście inteligencji i na tej podstawie wnioskował o istnieniu godnej uwagi zależności między pojemnością pamięci roboczej a inteligencją, musiałby się liczyć z ryzykiem pomylenia własnych życzeń z rzeczywistością. Wykonanie każdego takiego zadania będzie *zawsze* zależało od czynników decyzyjnych, których kryterium teoriodetekcyjne jest tylko pojedynczym, stosunkowo niewinnym przykładem, i czynników przypuszczalnie pozadecyzyjnych, takich jak pojemność pamięci roboczej, percepcyjna rozróżnialność bodźców, i tym podobne.

Niestety, teoria detekcji, także w wersjach, których tu nie omówiłem[1], przy całej swojej elastyczności i ogólności warunków dopuszczających jej zastosowanie ma dosyć poważną wadę. Nie pozwala kontrolować przetargu między czasem reakcji i poprawnością, dlatego że w ogóle nie uwzględnia czasu przetwarzania.

2.2 Dynamiczna teoria detekcji

Od pewnego czasu daje się zaobserwować wzrost popularności teorii podejmowania decyzji opartych na modelach sekwencyjnego próbkowania i kumulacji świadectw, takich jak dyskretne modele błądzenia przypadkowego i ciągłe modele dyfuzyjne (Ratcliff i Rouder, 1998; Balakrishnan, MacDonald, Busemeyer i Lin, 2002; Bogacz, Brown, Moehlis, Holmes i Cohen, 2006; Wagenmakers, 2009). Jak trafnie zauważyli Balakrishnan i in. (2002), stanowią one naturalne rozwinięcie teorii detekcji do wersji uwzględniającej dynamiczną naturę procesu podejmowania decyzji. Modele kumulacji świadectw różnią się od modeli opartych na klasycznej teorii detekcji przede wszystkim tym, że pozwalają uwzględnić efekt przetargu między czasem i poprawnością wykonania. Wzorując się na publikacjach wspomnianych autorów przedstawię teraz modele błądzenia przypadkowego i dyfuzyjny jako konieczne z teoretycznego punktu widzenia uogólnienia podstawowego modelu teorii detekcji.

W opisanym wcześniej, podstawowym modelu teorii detekcji, obserwator podejmuje decyzję na podstawie wartości pewnego wewnętrznego sygnału. Pojawienie się tego sygnału nie jest oczywiście prostym, bezpośrednim następstwem stymulacji bodźcowej, tylko rezultatem bliżej nieokreślonego, przypuszczalnie złożonego procesu przetwarzania informacji. Nie sposób zgodzić się z założeniem, że taki proces dostarcza za każdym razem jedynie pojedynczą próbkę informacji na temat kategorii bodźca. Niezależnie od tego, czy osoba badana ma za zadanie wykrywać trudno identyfikowalne sygnały w sytuacji umiarkowanej presji czasowej, czy w sytuacji braku takiej presji, wartość detektora będzie zależeć od czasu, jaki dzieli pojawienie się bodźca od momentu podjęcia decyzji. Nawet w przypadku występowania presji czasowej osoba badana może po prostu zdecydować, że podejmie decyzję szybciej lub wolniej. Wewnętrzny sygnał teorii detekcji można zinterpretować jako zmieniającą się w czasie wartość skumulowanego świadectwa, przemawiającego za wyborem jednej lub drugiej alternatywy. Jeżeli osoba badana pozwoli, aby proces kumulacji trwał dłużej, decyzja dotycząca klasy bodźca będzie oparta na większej liczbie próbek z rozkładu sygnału, wobec czego prawdopodobieństwo poprawnej klasyfikacji wzrośnie.

Wnioski oparte na oszacowanych wartościach parametrów modelu teorii detekcji będą poprawne tylko pod warunkiem, że sposób rozstrzygnięcia przetargu czas-poprawność nie będzie zmienną zakłócającą. Ostrożność czy staranność podejmowania decyzji nie

[1] Teoria detekcji bodźców wielowymiarowych, wyboru z dowolnej skończonej liczby alternatyw i inne, klarownie i szczegółowo opisane między innymi przez Macmillana i Creelmana (2002) i Luce'a (1963).

może więc systematycznie zależeć od rodzaju bodźców i ich rozróżnialności, przyjętego kryterium, spostrzeganej trudności zadania, wymiarów różnic indywidualnych i wartości innych zmiennych, które akurat interesują badacza. W ogólnym przypadku warunek ten z pewnością nie będzie spełniony. Podobnie trudno też zakładać, że czas do podjęcia decyzji nie będzie zależny od początkowej motywacji, subiektywnej wartości przypisywanej wymaganiom czasowym i poprawnościowym, długości trwania zadania, spostrzeganej trudności poszczególnych warunków, na przykład spostrzeganej trudności przeszukiwania pamięci w zależności od wielkości zestawu, i tak dalej.

Wyobraźmy sobie dwie osoby, dla których obserwowane proporcje reakcji poprawnych i błędnych są takie, że przy założeniu podstawowego modelu teorii detekcji rozróżnialności są równe, a na podstawie obserwowanych krzywych ROC można wywnioskować, że model jest w przybliżeniu trafny. Wyobraźmy sobie dalej, że dysponujemy informacją na temat czasu reakcji dla każdej podjętej w trakcie badania decyzji. W jaki sposób należy interpretować podobieństwo w oszacowanych wartościach parametru d', jeżeli okazuje się, że średni czas reakcji pierwszej osoby jest dłuższy niż średni czas reakcji osoby drugiej?

Rozsądne mogłoby się wydawać przeprowadzenie kolejnego eksperymentu, w którym podjęta zostałaby próba kontrolowania czasu podejmowania decyzji za pomocą instrukcji, albo zmian w konstrukcji zadania, które miałyby zwiększyć podobieństwo czasu podejmowania decyzji u różnych osób i u tej samej osoby w różnych warunkach [2]. Niezależnie od tego, czy wyniki dodatkowego eksperymentu byłyby zgodne, czy nie z początkowymi wnioskami (podobne lub różne wartości d'), takie rozwiązanie byłoby niezadowalające.

Zagwarantowanie podobnego czasu podejmowania decyzji u wszystkich osób i we wszystkich warunkach eksperymentalnych może być w wielu wypadkach trudne, albo po prostu niewykonalne. Co więcej, zjawisko przetargu między czasem i poprawnością jest samo w sobie godne uwagi. Problem jest bardzo podobny do tego, który ma być rozwiązywany przez klasyczną teorię detekcji. Tak jak w teorii detekcji dla oszacowania rozróżnialności kluczowe jest uwzględnienie wpływu kryterium, tak teraz równie ważne okazuje się uwzględnienie sposobu rozstrzygnięcia przetargu między czasem i poprawnością. Klasyczna teoria detekcji nie dostarcza tutaj żadnego rozwiązania.

Żeby wyjaśnić, w jakim znaczeniu modele błądzenia przypadkowego i dyfuzyjny stanowią uogólnienie teorii detekcji, warto najpierw przyjrzeć się bliżej klasycznemu modelowi detekcji przy założeniu, że liczba próbek kumulowanej informacji jest ustalona i taka sama dla każdej podejmowanej w zadaniu decyzji. Jeżeli x_i to i-ta próbka sygnału, którego rozkład zależy od klasy bodźca, to $L(n) = \sum_{i=1}^{n} x_i$ będzie skumulowanym świadectwem uzyskanym w n krokach. Podstawowy model klasycznej teorii detekcji uzyskujemy zakładając, że $n = 1$, x_1 ma rozkład normalny o średniej zależnej od kla-

[2] Wskazówki na temat tego, jak można próbować to zagwarantować, znajdują się między innymi w cytowanej wcześniej pracy Macmillana i Cleermana, w suplemencie III

sy bodźca i odchyleniu standardowym równym 1, a osoba badana podejmuje decyzję A, jeżeli $L(n) < k$ i decyzję B, jeżeli $L(n) > k$.

Gdy $n > 1$, ale liczba próbek jest ustalona i nie jest zmienną zakłócającą, klasyczny model nadal będzie niezłym rozwiązaniem. W takim wypadku, jeżeli dla każdej próby zadania wartości świadectw pochodzą z rozkładu normalnego o średniej zależnej od klasy bodźca i stałej wariancji ($x_{ij} \sim N(\delta_j, \sigma),\ j = A, B$), przyjmując punkt dokładnie pomiędzy δ_A i δ_B za środek skali, oczekiwana wartość każdego świadectwa w obecności bodźca A (B) będzie równa $E(x_{iA}) = \delta$ ($E(x_{iB}) = -\delta$). Różnica między średnimi, reprezentująca w modelu rozróżnialność, będzie dana przez $d = 2\delta$, a oczekiwana wartość skumulowanego świadectwa pochodzącego z n próbek będzie równa $n\delta$ dla bodźca A i $-n\delta$ dla bodźca B (średnia sumy jest równa sumie średnich). Gdy liczba próbek nie jest zmienną zakłócającą, średnia rozróżnialność dana przez $n2\delta$ będzie równie dobra jak d', ponieważ n decyduje wtedy tylko o skali. W wielu wypadkach liczba próbek będzie jednak zmienną zakłócającą.

Jeżeli wariancja poszczególnych próbek jest taka sama dla bodźca A i B, to wariancja skumulowanego świadectwa wynosi $\sigma^2(n) = n\sigma^2$. Ze względu na wzrost wariancji rozkładu skumulowanego świadectwa w czasie, właściwą miarą rozróżnialności jest różnica między średnimi rozkładów podzielona przez wspólne dla obu rozkładów odchylenie standardowe, zależne od liczby zebranych próbek:

$$\begin{aligned} d'(n) =& \frac{n2\delta}{\sqrt{n\sigma^2}} = \frac{n2\delta}{\sqrt{n}\sigma} = \sqrt{n}2\frac{\delta}{\sigma} \\ =& \sqrt{n}d \end{aligned}$$

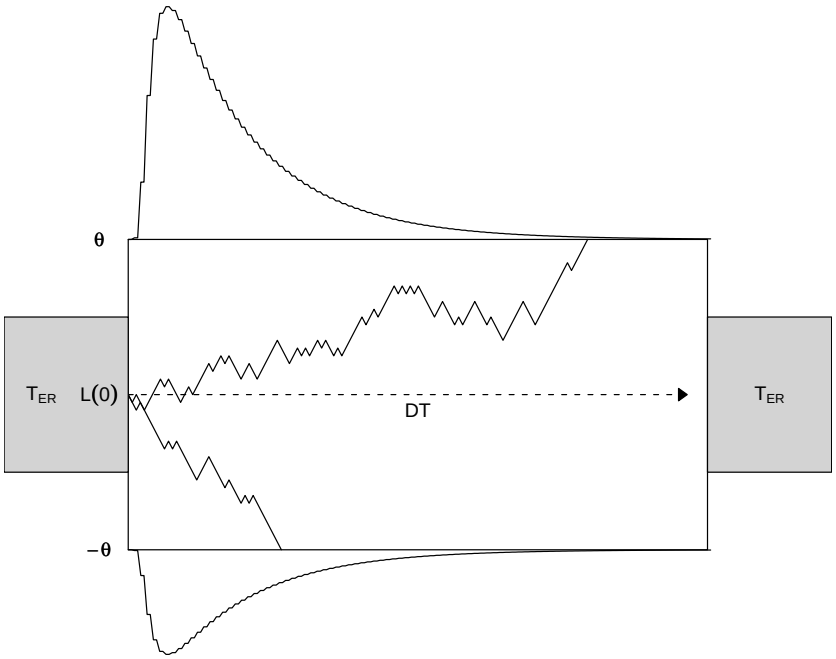
gdzie $d = 2(\delta/\sigma)$ to rozróżnialność *dla pojedynczej próbki*. Decyzja jest podejmowana na podstawie wartości świadectwa skumulowanego i zarówno $d'(n)$ jak i prawdopodobieństwa reakcji poprawnych i błędnych każdego rodzaju zależą od tego świadectwa, a nie od samej rozróżnialności dla pojedynczej próbki.

Dla ustalonej wartości n, $d'_1(n)/d'_2(n) = \sqrt{n}d_1/\sqrt{n}d_2 = d_1/d_2$, a więc stosunek dwóch wartości parametru d' jest taki sam, jak stosunek rozróżnialności dla pojedynczej próbki. Wartości n nie da się jednak oczywiście oszacować na podstawie obserwowanych proporcji reakcji poprawnych i błędnych. Gdyby dwie osoby różniły się liczbą próbek, ale nie rozróżnialnością dla pojedynczej próbki, tak, że dla pierwszej osoby $n_1 = 4$, a dla drugiej $n_2 = 16$ i $d_1 = d_2 = d$, wtedy rozróżnialności dla skumulowanego świadectwa wynosiłyby odpowiednio $d'_1(n_1) = \sqrt{n_1}d = 2d$ i $d'_2(n_2) = \sqrt{n_2}d = 4d$, wobec czego, stosując klasyczny model teorii detekcji uznalibyśmy, że bodźce są dwa razy lepiej rozróżnialne dla osoby drugiej, podczas gdy różnica wynikałaby wyłącznie z różnic w liczbie próbek, czyli w ilości czasu poświęconego na podjęcie decyzji. Również wpływ kryterium k będzie zależał od liczby próbek, a więc różnice w oszacowanej tendencyjności będą interpretowalne tylko w sytuacji, gdy liczba próbek jest znana, albo nie jest zmienną zakłócającą. Taki sam problem pojawia się dla wszystkich pozostałych miar opartych na teorii detekcji, takich jak obszar pod krzywą ROC.

Gdy tylko porzucimy założenie o stałym n, teoria detekcji może być rozwinięta do wersji, która pozwala modelować prawdopodobieństwa poszczególnych reakcji i całe rozkłady czasów reakcji poprawnych i błędnych. Dzięki temu, że na skutek takiego uogólnienia liczba wolnych parametrów rośnie nieznacznie i parametry te można czytelnie interpretować w kategoriach psychologicznych, znacząco wzrasta wrażliwość teorii na wyniki testów empirycznych. Dla uproszczenia omówię nieco dokładniej tylko jeden dyskretny model kumulacji świadectw (model błądzenia przypadkowego), a wersję ciągłą (model dyfuzyjny) przedstawię w zarysie.

2.2.1 Model kumulacji świadectw

Proces decyzyjny według dyskretnego modelu kumulacji świadectw polega na skokowym gromadzeniu informacji, będącej podstawą do klasyfikacji bodźca. Łączny czas poprzedzającego proces decyzyjny kodowania bodźca i następującego po tym procesie generowania reakcji motorycznej jest reprezentowany przez wolny parametr T_{ER} (ang. *Time of Encoding and Response generation*). Zakłada się, że całkowity czas reakcji, mierzony od momentu pojawienia się bodźca, jest sumą czasu kodowania i generowania reakcji (T_{ER}) i czasu poświęconego na podjęcie decyzji DT, czyli $RT = T_{ER} + DT$. Proces decyzyjny polega na sumowaniu kolejnych porcji informacji, które mogą być na przykład wydobywane z pamięci, albo powstawać na skutek innych, mniej lub bardziej złożonych procesów przetwarzania informacji w kolejnych, dyskretnych odstępach czasu. Przebieg tego procesu ilustruje poniższy diagram:



Rysunek 2.4: Schemat przebiegu dyskretnego procesu kumulacji świadectw

Linią łamaną zaznaczyłem dwie przykładowe trajektorie, z których jedna (tutaj akurat dłuższa) prowadzi do poprawnej, a druga do błędnej klasyfikacji. Początkowa wartość wsparcia $L(0)$ odpowiada tendencji do klasyfikowania bodźca do jednej z dwóch kategorii. Gdy $L(0) = 0$ proces decyzyjny przebiega nietendencyjnie, gdy $L(0) > 0$ faworyzowana jest alternatywa A, a gdy $L(0) < 0$ alternatywa B. W omawianej tutaj wersji modelu kolejne porcje informacji x_t w poszczególnych krokach czasowych t pochodzą z rozkładu normalnego o stałej wariancji i średniej δ, zależnej od prezentowanego bodźca. Arbitralnie można przyjąć, że jeśli prezentowany jest bodziec typu A, średnia będzie dodatnia ($\delta_A > 0$), jeżeli B, średnia będzie ujemna ($\delta_B < 0$).

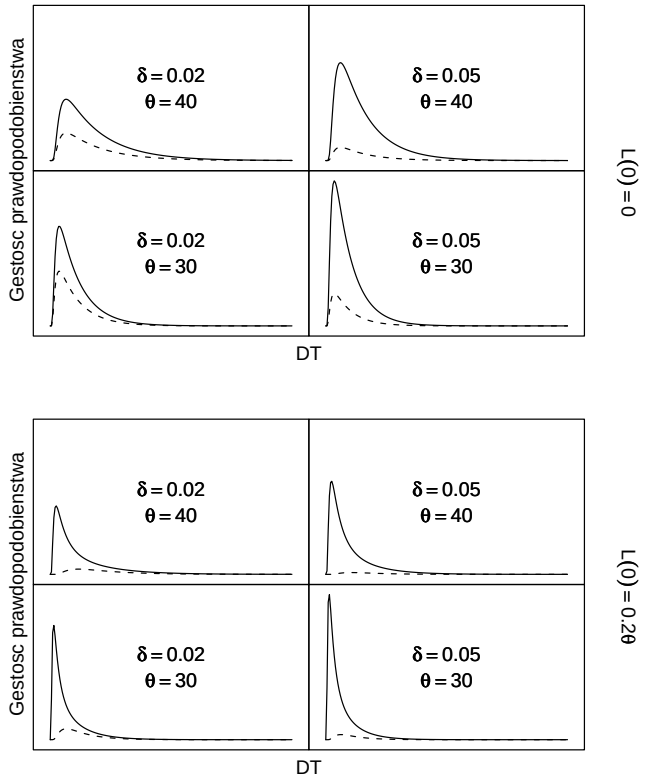
Kolejne próbki informacji sumowane są do momentu, gdy wartość skumulowana przekroczy jedną z dwóch wartości granicznych, θ lub $-\theta$, co spowoduje podjęcie decyzji o zaklasyfikowaniu bodźca do kategorii A lub B. Gdy $\delta = 0$, proces kumulacji świadectw nie faworyzuje żadnej z alternatyw i decyzja jest przypadkowa. Im bardziej δ rożni się od zera, tym szybciej kumulowane jest świadectwo dla określonej alternatywy. Jako że wariancja próbek jest niezerowa, proces zawsze może „zabłądzić" w niewłaściwym kierunku do tego stopnia, że przekroczony zostanie próg klasyfikacji błędnej. Prawdopo-

dobieństwo błędu będzie zależało jednak nie tylko od δ, ale również od θ. Im mniejszy będzie odstęp między progami klasyfikacji, tym mniejsza liczba kroków w niewłaściwą stronę wystarczy do przekroczenia progu odpowiadającego decyzji błędnej.

Parametr δ odpowiada tutaj parametrowi d' w klasycznej teorii detekcji, jest więc miarą rozróżnialności dla dynamicznie scharakteryzowanego procesu decyzyjnego. Teraz jednak obserwowalne reakcje zależą od łącznego wpływu rozróżnialności i dwóch, a nie jednego (parametr k w teorii detekcji), czynników podlegających kontroli ze strony osoby badanej. Pierwszy to początkowa wartość skumulowanego świadectwa $L(0)$, która powinna teoretycznie zależeć od subiektywnego prawdopodobieństwa każdej z alternatyw i subiektywnej użyteczności czterech możliwych rodzajów reakcji (dwóch możliwych reakcji poprawnych i dwóch możliwych reakcji błędnych). Drugi parametr decyzyjny to θ, określający sposób rozstrzygnięcia przetargu czas-poprawność. Im θ jest większe, tym bardziej rozsunięte są oba progi klasyfikacji. Wartość θ powinna więc zależeć od tego, czy osoba badana stara się wykonywać zadanie raczej szybko, czy raczej poprawnie. Wydaje się, że próg decyzyjny powinien też zależeć od rozróżnialności. Jeżeli bodźce są bardzo dobrze rozróżnialne, nie warto zbyt długo czekać z podjęciem decyzji, a o rozróżnialności bodźców osoba badana może przecież wnioskować na podstawie przebiegu wcześniejszych prób[3].

Jak widać na rysunku 2.4, rozkłady czasów pierwszego kontaktu z progami klasyfikacji są wyraźnie prawoskośne, przypominając pod tym względem rozkłady czasów reakcji. Czas reakcji jest jednak zmienną ciągłą, a w przedstawionej tutaj wersji modelu czas pierwszego kontaktu z progiem klasyfikacji jest zmienną dyskretną, obie zmienne różnią się zatem jakościowo. Nawet gdyby model dyskretny okazał się dobrym przybliżeniem, wyniki dopasowania byłyby zależne od zakładanej wielkości odstępu czasowego, co wprowadzałoby problematyczny element arbitralny. Ciągły model procesu dyfuzyjnego powstaje z dyskretnego modelu kumulacji świadectw w granicy, gdy czas trwania pojedynczego kroku dąży do zera. Poniżej przedstawiłem rozkłady czasów reakcji poprawnych (linia ciągła) i błędnych (linia przerywana) dla kilku przykładowych kombinacji wartości parametrów:

[3]Czasami można się spotkać z zarzutem, że parametry modeli kumulacji świadectw są zwykle skorelowane, a skoro reprezentują logicznie niezależne wymiary procesu decyzyjnego, ich wartości powinny być niezależne. Być może istnieją powody, dla których te parametry faktycznie nie powinny być skorelowane, ale takim powodem w żadnym razie nie jest niezależność logiczna.



Rysunek 2.5: Przykładowe rozkłady czasów reakcji poprawnych i błędnych dla modelu dyfuzyjnego

Większa rozróżnialność dla próbek δ (określana czasem jako tempo dryfu), tak samo jak niższa wartość progu decyzyjnego θ sprawiają, że przeciętnie proces kończy działanie szybciej. Dla ustalonej wartości δ wzrost θ lub tendencji do klasyfikowania bodźców jako należących do odpowiedniej klasy ($L(0) = 0{,}2\theta$) zmniejszają prawdopodobieństwo błędu (pole pod rozkładem zaznaczonym linią przerywaną).

Należy zaznaczyć, że wyróżnione teoretycznie czynniki decyzyjne, to jest rozróżnialność, przetarg czas-poprawność (próg decyzyjny) i początkowa wartość świadectwa (tendencyjność), które odgrywają istotną rolę w sytuacji wykonywania dowolnego zadania wyboru ze skończonej liczby alternatyw, w podstawowym modelu reprezentowane są przez minimalną potrzebną do tego celu liczbę wolnych parametrów. Rozkład normalny jest rozkładem granicznym sumy wielu zmiennych losowych, w miarę jak liczba tych zmiennych dąży do nieskończoności (centralne twierdzenie graniczne), dlatego założenie, że próbkowane informacje pochodzą z rozkładu normalnego jest teoretycznie uzasadnione. Niekontrowersyjne wydaje się też założenie, że nawet stosunkowo proste decyzje podejmowane są na podstawie wyników bardzo wielu elementarnych procesów prze-

twarzania informacji. Wobec braku dostatecznie wiarygodnej teorii, dotyczącej przebiegu tych zachodzących w tle procesów, założenie o normalności i stacjonarności rozkładu wypada uznać nie tylko za niewinne, ale wręcz konieczne uproszczenie.

2.2.2 Wsparcie empiryczne dla modelu dyfuzyjnego

Bogacz i in. (2006) wykazali niedawno, że niektóre ważniejsze modele kumulacji świadectw, takie jak model Ornsteina-Uhlenbecka (Busemeyer i Townsend, 1993) i cztery biologicznie motywowane modele koneksjonistyczne redukują się do modelu dyfuzyjnego przy założeniu rozsądnych granic dla wartości wolnych parametrów. Istnieje wiele modeli kumulacji świadectw, jednak jak dotąd model dyfuzyjny wydaje się najlepiej pasować do wyników bogatego zbioru zróżnicowanych eksperymentów (Wagenmakers, 2009). Warto w tym miejscu przytoczyć kilka takich przykładów.

W przypadku wielu zadań obserwuje się wolniejsze reakcje u osób starszych niż u osób młodszych. Do popularnych wyjaśnień tego zjawiska należy hipoteza, zgodnie z którą wraz z wiekiem mniej więcej takiemu samemu spowolnieniu ulegają wszystkie procesy poznawcze (Carella, 1985). Głównego wsparcia dla tej hipotezy mają dostarczać zależności ujawniające się na tak zwanych wykresach Brinley'a, na których zestawiony jest średni czas reakcji osób młodszych i starszych dla różnych warunków eksperymentalnych. Obserwowane na tych wykresach zależności są często w przybliżeniu liniowe, co miałoby oznaczać, że spowolnienie związane z wiekiem następuje niezależnie od warunku i zadania. Średni czas reakcji nie jest jednak miarą czasu trwania procesów poznawczych, tylko miarą czasu wykonania zadania. Żeby wywnioskować cokolwiek na temat czasu trwania lub innych własności procesów poznawczych trzeba dysponować teorią tych procesów. Model dyfuzyjny pozwala określić, jakie przypuszczalnie własności procesu zmieniają się z wiekiem. Rezultaty zastosowania tego modelu (Ratcliff, Thapar i McKoon, 2001, 2003; Ratcliff, Thapar, Gomez i McKoon, 2004; Ratcliff, Thapar i McKoon, 2006) zdają się świadczyć, że z wiekiem rośnie przede wszystkim czas procesów niedecyzyjnych (być może generowanie reakcji motorycznych lub kodowanie bodźców) i ostrożność reagowania. W zależności od zadania spowolnienie może, ale nie musi dotyczyć tempa kumulacji. Bez czegoś przypominającego model dyfuzyjny nie można się obejść interpretując wyniki tego rodzaju zadań w kategoriach przebiegu nieobserwowalnych procesów, a wyjaśnienie oparte na zastosowaniu modelu dyfuzyjnego wydaje się stosunkowo intuicyjne.

W przypadku wielu zadań osoby o wyższym IQ reagują szybciej (Jensen, 2006). Co ciekawe, zależność wydaje się silniejsza dla wolniejszych niż dla szybszych reakcji. Larson i Alderton (1990) nazwali to „efektem najgorszego wykonania". Ratcliff, Schmiedek i McKoon (2008) wykazali, że efekt najgorszego wykonania wynika z modelu dyfuzyjnego, jeżeli tylko założyć, że czas niedecyzyjny zmienia się przypadkowo między osobami, a IQ jest związane z różnicami w tempie kumulacji lub ostrożności. Założenie o występujących między osobami badanymi różnicach w czasie niedecyzyjnym nie jest ani trochę

kontrowersyjne, podobnie jak przypuszczenie, że osoby różniące się inteligencją będą się również różniły tempem kumulacji lub ostrożnością reagowania. Model dyfuzyjny dostarcza stosunkowo prostego wyjaśnienia tego efektu. Zdarza się również, że nie obserwuje się związku między IQ a średnim czasem reakcji lub proporcją błędów. Na przykład, Weeda, Wagenmakers i Huizenga (2007) nie zaobserwowali takich (statystycznie istotnych) zależności w zadaniu na dyskryminację. Dopiero dopasowanie do tych samych danych modelu dyfuzyjnego pozwoliło stwierdzić, że osoby o wyższym IQ charakteryzują się większym tempem kumulacji i mniej ostrożnym sposobem reagowania.

Model zastosowano także do analizy danych pochodzących z eksperymentów dotyczących efektów uczenia się (tak zwane „potęgowe prawo zapominania", Dutilh, Wagenmakers, Vandekerckhove i Tuerlinckx, 2008), utajonych postaw (test utajonych postaw, Greenwald, McGhee i Schwartz, 1998; Bones i Johnson, 2007; Klauer, Voss, Schmitz i Teige-Mocigemba, 2007), związków między poprzedzaniem a uczeniem się utajonym (Ratcliff i McKoon, 1997), przeszukiwania pamięci krótkoterminowej (Ratcliff, 1978), przeszukiwania wzrokowego i wielu innych zjawisk. Wyniki przeprowadzonych w tym celu przez Voss, Rothermunda i Voss (2004) eksperymentów zdają się świadczyć o zgodnej z teorią, względnie selektywnej modyfikowalności parametrów modelu przez odpowiednie czynniki eksperymentalne, takie jak macierz wypłat, akcentowanie w instrukcji czasu lub poprawności i trudność zadania.

Wszystkie wymienione wyniki i wiele innych, których nie ma potrzeby tutaj przytaczać, świadczą dobitnie, że model dyfuzyjny stosunkowo dobrze oddaje pewne ważne własności procesu przetwarzania informacji odpowiedzialnego za wykonywanie wielu zadań wyboru z dwóch alternatyw[4]. Wsparcie dla tego modelu ma zasadniczo inny charakter, niż wsparcie dla hipotez architektonicznych, które można uzyskać posługując się teorią identyfikacji architektury Townsenda. Warto tutaj podkreślić, że model dyfuzyjny został początkowo zaproponowany przez Ratcliffa właśnie jako teoria procesu przeszukiwania pamięci krótkoterminowej (Ratcliff, 1978). Nie ma w tym nic dziwnego, skoro przeszukiwanie pamięci krótkoterminowej bada się zwykle za pomocą zadań wyboru z dwóch alternatyw (bodziec testowy jest, albo nie jest obecny w zestawie do zapamiętania). Część wniosków dotyczących możliwych strategii badawczych w psychologii poznawczej, na których zamierzam oprzeć uzasadnienie ostatecznych tez pracy, można zilustrować odwołując się tylko do tych dwóch, jak się jeszcze okaże dosyć szczególnych propozycji teoretycznych.

[4]Model dyfuzyjny dopiero niedawno zaczął wyraźnie zyskiwać na popularności, głównie dzięki odkryciu efektywnych metod przybliżania jego funkcji wiarygodności. Chciałbym w tym miejscu podziękować Michaelowi Lee i Francisowi Tuerlinckxowi, którzy informowali mnie na bieżąco o postępach prac nad rozwiązaniem tego problemu i dostarczali mi cennych wskazówek, bez których nie miałbym szansy na szersze zastosowanie modelu dyfuzyjnego.

2.3 Podsumowanie

Teoria Townsenda i model kumulacji świadectw są moim zdaniem wzorcowymi przykładami zastosowania dwóch różnych, wzajemnie niewykluczających się strategii badania procesów poznawczych. Jedna z tych strategii polega na poszukiwaniu warunków identyfikowalności modelu ze względu na ogólnie zdefiniowaną, względnie wyczerpującą przestrzeń teoretycznie dopuszczalnych alternatyw. Pozornie, druga strategia zdaje się polegać na maksymalnym wykorzystaniu informacji zawartej w danych (rozkład prawdopodobieństwa jest wyczerpującą charakterystyką zmiennej losowej), w czym odlegle przypominałaby strategię stosowania modeli zintegrowanych. Czasem właśnie takie uzasadnienie dla stosowania modeli kumulacji świadectw podają autorzy, którzy je stosują.

Model dyfuzyjny jest ewidentnie lokalny i abstrahuje od wielu procesów, które muszą zachodzić, aby w ogóle możliwe było wykonywanie zadań wyboru z dwóch alternatyw, takich jak percepcja i kodowanie bodźców, selekcja reakcji motorycznej, interpretacja instrukcji, kontrola progu decyzyjnego, początkowego odchylenia, i tym podobne. Wymienione badania, w których model dostarczył interesujących i stosunkowo przekonujących wyjaśnień, nie są przykładem zastosowania strategii 20 pytań - przestrzeń alternatywnych hipotez nie jest nawet w przybliżeniu zarysowana. Mimo to, wsparcie empiryczne wydaje się silne, a model niezwykle użyteczny.

Wartość modeli kumulacji świadectw nie sprowadza się do tego, że predykcje tych modeli pasują do obserwowanych rozkładów prawdopodobieństwa czasu reakcji i poprawności. Stosujący je badacze starają się zawsze sprawdzić, na ile przewidywane rozkłady pasują do rozkładów obserwowanych, ale w istocie dopasowanie wcale nie powinno być w tym przypadku imponujące, a wręcz powinno być co najwyżej akceptowalne. Rozkład procesów resztowych nie jest znany, a obserwowane czasy reakcji są sumą łącznego czasu trwania tych procesów i procesu decyzyjnego, obserwowany rozkład czasów reakcji jest więc konwolucją rozkładu czasu trwania procesu decyzyjnego i bliżej nieokreślonych procesów resztowych. Dopóki rozkład procesów resztowych nie będzie lepiej poznany, predykcje modelu dyfuzyjnego będą musiały, czasem nawet znacząco, odbiegać od rozkładów obserwowanych, nawet jeżeli model dyfuzyjny jest bliski prawdy. Model Ratcliffa ma w sobie coś szczególnego, czego brakuje systemom przetwarzania skończonej liczby elementów Townsenda, a czego nie da się sprowadzić do trafności predykcji. Żeby ustalić, co to właściwie jest, w dwóch następnych rozdziałach zajmę się kwestią oceny wartości poznawczej hipotez.

Rozdział 3

Ilościowa ocena hipotez i jej ograniczenia

Tematem tego rozdziału jest przedstawione z perspektywy modelowania matematycznego zagadnienie ilościowej oceny hipotez, przez co rozumiem ocenę trafności wynikających z hipotez predykcji. Ujmując rzecz w znacznym uproszczeniu, na tym etapie rozważań przyjmuję następujący schemat procesu badawczego. Dysponując wiedzą z zakresu danej dziedziny (psychologii poznawczej), badacz formułuje w języku właściwym dla tej dziedziny pytanie, dotyczące interesującego go zjawiska (w jaki sposób przebiega przeszukiwanie pamięci krótkoterminowej). Następnie dochodzi do sformułowania pewnej hipotezy (przeszukiwanie pamięci krótkoterminowej przebiega szeregowo), albo zbioru alternatywnych hipotez, które wydają mu się wystarczająco rozsądne, aby poddać je ocenie empirycznej (proces przebiega wyczerpująco lub samowygaszająco). Rozważane hipotezy są możliwymi odpowiedziami na interesujące go pytanie. Aby ocenić, czy dana hipoteza może być uznana za bliską prawdy odpowiedź na to pytanie, badacz przeprowadza testy empiryczne. Wymaga to wprowadzenia dodatkowych założeń, dotyczących związku między hipotezą a przebiegiem testów empirycznych. W szczególności, założenia te pozwalają na wyprowadzenie obserwowalnych konsekwencji hipotez (operacjonalizacja). W końcu, uzyskane wyniki interpretowane są w taki sposób, aby wywnioskować coś na temat wiarygodności hipotez. W rozdziale czwartym przedstawię argumenty świadczące o tym, że w zarysowanym właśnie schemacie procesu badawczego najważniejszy element, to jest ocena jakościowa, został w zasadzie pominięty, najpierw jednak zajmę się zagadnieniem oceny ilościowej i granicami jej użyteczności.

3.1 Modelowanie matematyczne

Modelowanie matematyczne to eksplikacja w formalnym języku matematyki wyrażonych w postaci werbalnej propozycji teoretycznych. Ilościowa ocena propozycji teoretycznych w psychologii prawie zawsze wymaga jako koniecznego kroku przeprowadzenia wnioskowania statystycznego, które nie jest możliwe bez choćby częściowej forma-

lizacji teorii werbalnej, zwykle jednak terminem „modelowanie matematyczne" określa się procedury znacznie wykraczające poza wyrażenie niektórych założeń lub konsekwencji teorii w postaci hipotezy statystycznej dla potrzeb testu istotności (Luce, 1995; Myung i Pitt, 2002). Pomijając przypadki, które trudno jednoznacznie zaklasyfikować jako modelowanie dla potrzeb testu istotności, albo modelowanie jako formalizacja całości lub znacznej części teorii, takie jak modelowanie równań strukturalnych albo modelowanie graficzne, różnice między tymi dwoma sposobami wykorzystania modelowania wydają się względnie wyraźne. Problem złożoności modelu zwykle albo nie pojawia się wcale, albo jest znacznie mniej dotkliwy w przypadku formalizacji dla potrzeb testowania istotności (na przykład, jako szczególnego przypadku modelu liniowego) wybranych konsekwencji lub założeń teorii (dajmy na to, zgodnie z teorią dwie zmienne powinny być skorelowane). Niemniej, z powodu ważnych podobieństw, zarówno logikę testowania stosunkowo prostych hipotez jak i bardziej złożonych lub pełniej reprezentujących teorię modeli omówię z ogólnej perspektywy ilościowej oceny modeli matematycznych.

3.1.1 Model matematyczny jako rodzina rozkładów prawdopodobieństwa

W najogólniejszej rozważanej w tej pracy postaci, model matematyczny jest jedno lub wielowymiarową rodziną rozkładów prawdopodobieństwa, określoną na wektorze zmiennych obserwowalnych, indeksowaną przez wektor parametrów modelu:

$$M = \{f(x|\theta) : \theta \in \Theta\}$$

gdzie x to wektor zmiennych obserwowalnych, θ to wektor należący do przestrzeni parametrów modelu Θ, a f fo funkcja (jedno lub wielowymiarowego) parametrycznego rozkładu prawdopodobieństwa. Tak rozumiany model jest pewną rodziną rozkładów prawdopodobieństwa.

Każda konkretna wartość wektora θ wyznacza dokładnie jeden rozkład należący do M. Zgodnie z konwencją często stosowaną w literaturze statystycznej, elementy zbioru M będę czasem określał jako „hipotezy" lub „hipotezy punktowe". Jeżeli M to model liniowy z nieznaną średnią i wariancją, elementami tego zbioru będą wszystkie rozkłady normalne. W takim wypadku wektor θ będzie się składał z dwóch parametrów, to jest średniej i wariancji, a każda możliwa wartość tego wektora będzie wyznaczała unikalnie jeden rozkład normalny. W tym kontekście rozkład prawdopodobieństwa, hipoteza punktowa i wartość wektora parametrów oznaczają to samo. W praktyce prawie zawsze niektóre zmienne wyróżnia się jako zmienne zależne, a inne jako niezależne, co można przedstawić w równoważnej, choć może bardziej czytelnej postaci jako $f(y|\theta, x)$. Wtedy y oznacza obserwowalne zmienne zależne, a x obserwowalne zmienne niezależne, które są traktowane jako wartości stałe.

Niektóre lub wszystkie parametry składające się na wektor θ mogą być *ustalone*, zaś pozostałe parametry będą parametrami *wolnymi*. Wartość parametrów ustalonych jest

znana badaczowi i określona z góry, przed przeprowadzeniem badania. Ocena ilościowa modelu z wolnymi parametrami jako koniecznego kroku wymaga dopasowania takiego modelu do danych. Dopasowanie polega na poszukiwaniu takich wartości wolnych parametrów, aby były one w jakimś statystycznym sensie najlepsze, co oznacza maksymalizację tak lub inaczej rozumianej zgodności danych z opisem tych danych z perspektywy modelu. W przypadku modelu liniowego zwykle osiąga się to przez minimalizację resztowej sumy kwadratów, która jest jedną z wielu możliwych miar niedopasowania.

Konkretny model będzie wyrażał zależności między parametrami i zmiennymi w formie relacji matematycznych albo algorytmicznych. To właśnie te relacje wyznaczają rodzinę rozkładów prawdopodobieństwa, określoną na zmiennych obserwowalnych. Z punktu widzenia testowania adekwatności deskryptywnej (dopasowania), uogólnialności i złożoności modelu, charakter zmiennych (na przykład to, czy zmienna jest ilościowa czy jakościowa) jak i typ samego modelu (czy jest to model algorytmiczny, algebraiczny, koneksjonistyczny, czy jakikolwiek inny) ma jedynie znaczenie praktyczne w tym znaczeniu, że różne rodzaje modeli mogą wymagać lub umożliwiać zastosowanie różnych procedur dopasowywania i testowania, ogólna logika pozostaje jednak ta sama.

Jako że przeprowadzenie wnioskowania statystycznego wymaga uwzględnienia komponentu losowego, rozważana tutaj ogólna postać modelu matematycznego w sposób użyteczny charakteryzuje przeważającą większość modeli stosowanych w psychologii. Testowanie modeli, dla których funkcji rozkładu prawdopodobieństwa nie da się ustalić analitycznie, polega w istocie na jawnym bądź niejawnym przybliżaniu tej funkcji innymi metodami, na przykład za pomocą symulacji. Określenie „model matematyczny" jest więc adekwatne dla wszystkich modeli wyrażonych w postaci procedury symulacji (takich jak algorytmiczne), nawet jeżeli na pierwszy rzut oka może się wydawać, że wyprowadzanie predykcji nie ma w przypadku tych modeli charakteru dedukcyjnego. Symulację stosuje się między innymi wtedy, gdy ustalenie funkcji rozkładu prawdopodobieństwa reprezentującego dany model jest zbyt kosztowne lub nie wiadomo jak ją wyznaczyć. Nie jest mi znany żaden przykład potraktowania wyników symulacji jako obserwacji pozwalających na testowanie hipotez *empirycznych* dotyczących samego modelu, dlatego zakładam dalej, że w kontekście oceny ilościowej symulacja modelu jest jedynie użytecznym rozwiązaniem technicznych trudności związanych z dedukcyjnym wyprowadzeniem predykcji.

3.2 Test istotności hipotezy zerowej

Najczęściej stosowaną w psychologii procedurą wnioskowania statystycznego jest test istotności hipotezy zerowej, odtąd określany skrótem NHST. W większości przypadków opiera się on na zastosowaniu jakiejś postaci modelu liniowego (ogólnego, uogól-

nionego, lub innej)[1]. Aby zastosować NHST badacz musi najpierw zinterpretować wybrane założenia lub konsekwencje teorii w kategoriach hipotez statystycznych, zerowej i alternatywnej. Zwykle hipoteza zerowa odpowiada założeniu o braku związku między zmiennymi, a hipoteza alternatywna ma reprezentować pewne konsekwencje testowanej teorii, dotyczące hipotetycznych zależności między zmiennymi. Na przykład, badacz może przypuszczać, że w pewnych granicach, wraz ze wzrostem dawki kofeiny będzie rósł poziom pobudzenia, mierzony za pomocą skonstruowanego w tym celu kwestionariusza. Ocenie ilościowej może być wtedy poddana następująca hipoteza zerowa:

$$y = \theta_0 + 0x + \epsilon,\ \epsilon \sim N(0, \sigma)$$

gdzie y to wektor wyników uzyskanych w kwestionariuszu, θ_0 to średni wynik w kwestionariuszu, x to wektor kodujący zastosowane dawki kofeiny, a ϵ to wektor wyników resztowych o rozkładzie normalnym, z zerową średnią i nieznaną wariancją, reprezentujący błąd pomiaru i wpływ pozostałych czynników niekontrolowanych w eksperymencie[2]. Jak widać, hipoteza zerowa nie musi być i zwykle nie jest hipotezą punktową[3]. Na podstawie zebranych obserwacji obliczana jest odpowiednia statystyka - jeżeli zastosowano dwie dawki kofeiny, może to być statystyka t dla dwóch grup niezależnych. Posługując się standardową teorią wnioskowania statystycznego można wtedy ustalić, jakie jest prawdopodobieństwo p uzyskania wartości t większej lub równej od tej, jaka została zaobserwowana, przy założeniu hipotezy zerowej. Jeżeli wartość p jest odpowiednio mała stwierdza się, że hipotezę zerową należy odrzucić na rzecz hipotezy alternatywnej. Jeżeli wartość p jest większa niż pewna konwencjonalnie przyjęta wartość graniczna (zwykle $0{,}05$), uznaje się, że hipoteza zerowa nie może być odrzucona.

Do niewątpliwych zalet NHST należy jej prostota. Omówienie mniej lub bardziej poważnych wad i ograniczeń tej procedury zajęłoby tutaj zdecydowanie zbyt wiele miejsca, na szczęście nie muszę w tej pracy zajmować stanowiska w sporze o szerzej rozumianą użyteczność testu istotności hipotezy zerowej jako narzędzia ilościowej oceny hipotez. Odniosę się tylko do tych problemów, których częściowe rozwiązanie staje się możliwe dopiero dzięki zastosowaniu modelowania matematycznego i porzuceniu standardowych metod oceny dobroci dopasowania na rzecz metod uwzględniających w sposób teoretycznie uzasadniony statystyczną złożoność modelu.

[1] Można przeprowadzić test istotności dla innych modeli niż liniowy, istnieją też metody wnioskowania statystycznego na modelach liniowych, które nie polegają na testowaniu istotności hipotezy zerowej.

[2] Gdyby zależność między dawką kofeiny a poziomem pobudzenia miała być nieliniowa, model wyglądałby nieco inaczej. Zamiast wektora x model mógłby wtedy zawierać macierz kodującą poziomy czynnika, jednak w żaden sposób nie zmieniałoby to sensu procedury wnioskowania.

[3] Zgodnie z najogólniejszą znaną mi definicją hipoteza zerowa jest dowolnym modelem zagnieżdżonym w modelu reprezentującym hipotezę alternatywną. Odrzucając hipotezę zerową stwierdza się, że model ogólniejszy lepiej oddaje zaobserwowane regularności.

3.2.1 Kłopotliwe rozwiązania niektórych problemów związanych z testowaniem istotności hipotezy zerowej

Jednym z bardziej dotkliwych, powszechnie znanych ograniczeń NHST jest redukcja procesu wnioskowania statystycznego do rozstrzygnięcia na ogół niezbyt interesującej kwestii, jaką jest odrzucenie bądź nieodrzucenie hipotezy zerowej. Z reguły badacz jest zainteresowany uzyskaniem odpowiedzi na szereg pytań ilościowych, które nie dają się sprowadzić do uzyskanego poziomu istotności. Nawet jeżeli poziom istotności jest wysoki, sama ta informacja nie mówi nic ani o sile związku, ani o prawdopodobieństwie danych ze względu na interesującą badacza hipotezę alternatywną. Poziom istotności będzie wysoki nawet wtedy, gdy siła związku jest mała, jeżeli tylko zbiór obserwacji będzie wystarczająco duży i odwrotnie, nawet gdy siła związku jest stosunkowo duża, poziom istotności będzie niski dla mniejszych prób. Hipoteza alternatywna, tak jak jest prawie zawsze reprezentowana w NHST, ma niewiele wspólnego z hipotezą badawczą, oznacza bowiem jedynie negację hipotezy zerowej, a więc wyraża przypuszczenie, że jakieś bliżej nieokreślone związki, dające się wyrazić dopiero za pomocą ogólniejszego niż hipoteza zerowa modelu, faktycznie zachodzą. W większości przypadków nie tylko wiadomo z góry, że hipoteza zerowa jest bardzo daleka od prawdy, ale da się także wiele powiedzieć na temat wiarygodnych wartości, jakie teoretycznie powinna przyjmować hipoteza alternatywna, na przykład, często z góry wiadomo, że korelacja powinna być dodatnia, albo że powinna być ujemna.

Obecnie trudno znaleźć podręcznik metodologii, w którym NHST byłaby potraktowana bezkrytycznie. W odpowiedzi na zmasowany atak, zastrzeżenia dotyczące ograniczeń NHST i innych standardowych metod wnioskowania statystycznego, a także sugestie co do możliwych sposób ominięcia tych ograniczeń pojawiły się w opublikowanym w roku 1999 raporcie specjalnie do tego celu powołanego przez APA ciała (ang. *Task Force on Statistical Inference,* Wilkinson, 1999), zawierającym „zalecenia i wyjaśnienia dotyczące wykorzystania metod statystycznych w czasopismach psychologicznych". Cytując za autorami, celem raportu była:

> ... klaryfikacja niektórych kontrowersyjnych kwestii dotyczących zastosowań statystyki, włączając w to testowanie istotności i alternatywy; alternatywne modele i metody przekształcania danych, a także nowe metody, które stały się możliwe dzięki wzroście mocy obliczeniowych komputerów.
>
> (s. 594, tamże)

W praktyce formalne procedury statystyczne odgrywają niezwykle ważną rolę w procesie oceny teorii psychologicznych, dlatego znaczenie tego raportu jest trudne do przecenienia. Propozycje autorów stanowią jak sądzę reprezentatywną próbkę tego, co można odnaleźć we współczesnej literaturze dotyczącej metodologii badań psychologicznych, stąd też do innych publikacji będę się odwoływał jedynie sporadycznie.

3.2.2 Dwie interpretacje oszacowań przedziałowych

Wspomniany raport zawiera szereg części, poświęconych rozmaitym zagadnieniom związanym z planowaniem i przeprowadzaniem badań, analizą wyników i oceną hipotez. W szczególności, podpunkt zatytułowany „Testowanie hipotez" rozpoczyna się takim oto fragmentem, któremu towarzyszy sugestia, aby Czytelnik zapoznał się z pewnym wpływowym artykułem autorstwa Cohena (1994):

> Trudno wyobrazić sobie sytuację, w której dychotomiczna decyzja o akceptacji lub odrzuceniu [hipotezy alternatywnej] jest lepsza niż przedstawienie faktycznie uzyskanej wartości p lub, jeszcze lepiej, przedziałów ufności. Nigdy nie używaj niefortunnego zwrotu „akceptuję hipotezę zerową". Zawsze dostarczaj oszacowania wielkości efektu wraz z wartością p.
>
> (s. 599, tamże)

Przypuszczalnie najbardziej powszechnie akceptowanym rozwiązaniem problemów związanych z NHST jest właśnie wykorzystanie w procesie wnioskowania przedziałów ufności, oszacowań (przedziałowych) wielkości efektów, a także (o czym akurat z raportu dowiedzieć się nie można) porównań planowanych (analiza kontrastów jako alternatywa dla luźno związanego z hipotezą badawczą ogólnego modelu liniowego). Aby wyjaśnić, dlaczego wymienione rozwiązania są kłopotliwe, omówię dokładniej standardową logikę testowania hipotez dotyczących modelu liniowego. Większość argumentów, jakie przedstawię w tej części pracy, wynika wprost z częstościowej i subiektywistycznej bayesowskiej definicji prawdopodobieństwa i byłbym naprawdę zdumiony, gdyby choć jeden z nich miał się okazać oryginalny.

W przypadku modelu liniowego najprostszą postać hipoteza zerowa przyjmuje w sytuacji, gdy badacz zainteresowany jest nieznaną wartością średniej w populacji. Zakłada się wtedy (hipoteza alternatywna), że dane pochodzą z rozkładu normalnego o nieznanej średniej (μ) i odchyleniu standardowym (σ) i wartości tych parametrów szacuje się na podstawie danych metodą najmniejszych kwadratów. Hipoteza zerowa przyjmuje postać $\mu = 0$, a hipoteza alternatywna $\mu \neq 0$ (test bezkierunkowy). Odpowiednia teoria pozwala ustalić, z jakim prawdopodobieństwem średnia z próby może się różnić co najmniej o daną wielkość od średniej w populacji (zgodnie z modelem liniowym średnia z próby powinna mieć w przybliżeniu rozkład t z $n - 1$ stopniami swobody, gdzie n to liczba obserwacji). W szczególności, można wyznaczyć tak zwaną wartość krytyczną, czyli taką wielkość różnicy między zakładaną wartością średniej w populacji a średnią zaobserwowaną, że różnica między średnią z próby a średnią zakładaną będzie większa lub równa niż wartość krytyczna z określonym z góry prawdopodobieństwem. Dokładnie, standardowa teoria statystyczna pozwala obliczyć:

$$\text{Poziom istotności} = p(s(x) \geqslant c | h_0 \subset M)$$

gdzie p to poziom istotności, $s(x)$ to pewna statystyka (statystyka t, średnia, albo inna), a c to dowolna wartość krytyczna tej statystyki. Jeżeli prawdopodobieństwo związane z wartością krytyczną jest odpowiednio małe (nie większe niż $0{,}05$), a statystyka z próby ma wartość nie mniejszą niż ta wartość krytyczna, hipotezę zerową uznaje się za odrzuconą. W innym wypadku wyniki nie pozwalają na podjęcie żadnej decyzji. Można wykazać, że przy założeniu hipotezy zerowej postępowanie to prowadzi asymptotycznie (w miarę jak wielkość próby dąży do nieskończoności) do zakładanego procenta błędnych odrzuceń, o ile tylko model jest prawdziwy, to znaczy zawiera rozkład, z którego faktycznie pochodzą dane.

Cały ten proces można przedstawić graficznie, posługując się tak zwanymi przedziałami ufności. Odrzucenie hipotezy zerowej jest wtedy równoznaczne ze stwierdzeniem, że wartość reprezentowana przez tą hipotezę znajduje się poza granicami odpowiednich przedziałów. Oszacowania przedziałowe rzeczywiście dostarczają więcej informacji niż prosta decyzja o odrzuceniu bądź nieodrzuceniu hipotezy zerowej. Wielkość tych przedziałów związana jest z mocą statystyczną, czyli prawdopodobieństwem poprawnego odrzucenia hipotezy zerowej (wykrycia efektu) w sytuacji, gdy jest ona fałszywa (efekt istnieje). Im większy będzie efekt w populacji przy ustalonej wielkości próby, tym mniejsze będą zwykle przedziały ufności. Również, im większa próba przy ustalonej wielkości efektu, tym mniejsze przedziały ufności. Stosunkowo często wyprowadza się stąd wniosek, że przedziały ufności informują badacza o pewności, z jaką średnia w populacji znajduje się pomiędzy określonymi wartościami granicznymi. Zgodnie z tą logiką, jeżeli przedziały obliczono przy założeniu prawdopodobieństwa błędnego odrzucenia hipotezy zerowej $\alpha = 0{,}05$, interpretacja oszacowania przedziałowego miałaby brzmieć:

> (...) badacz może wyprowadzić prosty wniosek, że z prawdopodobieństwem 95 procent przedziały ufności zawierają μ. (3.2.2)
>
> (s. 353, Loftus, 2002)

Czytelnik może się łatwo przekonać o popularności tej interpretacji przeglądając zawartość licznych publikacji traktujących o zastosowaniach oszacowań przedziałowych. Gdy hipoteza zerowa dotyczy różnic między średnimi, nachylenia linii regresji lub interakcji, NHST można równoważnie przedstawić jako wnioskowanie oparte na przedziałowych oszacowaniach tych efektów.

Oszacowanie przedziałowe jest bardziej pouczające niż sam poziom istotności, ponieważ informuje o mocy statystycznej. Założenia na których oparte są zarówno przedziały ufności jak i wartość p nie pozwalają jednak na interpretację (3.2.2), ze względu na przyjmowaną definicję prawdopodobieństwa. Podstawowe dla wszystkich standardowych procedur wnioskowania statystycznego jest „obiektywne" rozumienie prawdopodobieństwa, jako relatywnej częstości występowania zdarzeń w nieskończonej liczbie

powtórzeń tego samego eksperymentu (Feller, 2006). Zakłada się przy tym, że dane pochodzą z dokładnie jednego, nieznanego rozkładu prawdopodobieństwa, należącego do modelu. Przedziały ufności zawierają wartość prawdziwą, albo jej nie zawierają. W takim ujęciu prawdopodobieństwo dotyczy zawsze *wyników badania*, a nigdy *wartości wolnych parametrów*.

Sformułowanie twierdzenia o charakterze probabilistycznym na temat parametrów modelu w paradygmacie częstościowym nie jest możliwe, chociaż można powiedzieć, z jakim prawdopodobieństwem może wystąpić różnica o określonej wielkości między oszacowaniem z próby a prawdziwą wartością w populacji. Na pierwszy rzut oka kwestia może sprawiać wrażenie „werbalnej" czy „filozoficznej" (w złym tego słowa znaczeniu), jeżeli jednak badacz jest skłonny zaakceptować interpretację (3.2.2), a ze zrozumiałych względów wielu badaczy właśnie tego chce, zmuszony jest jednocześnie porzucić częstościową definicję prawdopodobieństwa. Każda konkretna wartość wolnego parametru jest albo poprawna, albo błędna, ale nie jest rezultatem żadnego eksperymentu. W ramach częstościowego rozumienia prawdopodobieństwa hipoteza jest prawdziwa albo fałszywa, a nie mniej lub bardziej prawdopodobna. Można określić rozkład prawdopodobieństwa na wartościach parametrów, ale tylko jeżeli zaakceptuje się inną interpretację prawdopodobieństwa, na przykład subiektywistyczną. W interpretacji subiektywistycznej prawdopodobieństwo wyraża *siłę przekonania* co do możliwych wartości parametru w populacji, a nie relatywną częstość wyników tego samego eksperymentu w nieskończonej liczbie prób. W paradygmacie częstościowym przedziały ufności wyrażają niepewność co do przyszłych wyników tego samego eksperymentu.

Zważywszy na rolę, jaką wnioskowaniu opartemu na przedziałach ufności i interpretacji (3.2.2) przypisuje się często w kontekście krytyki NHST, można by dojść do wniosku, że psychologowie powinni być może zostać bayesianistami. Nie da się jednak konsekwentnie rozumieć prawdopodobieństwa częstościowo na początku przeprowadzania wnioskowania, by na końcu zinterpretować je subiektywistycznie. Parametr modelu albo jest, albo nie jest zmienną losową. Doskonały przykład powszechnego braku zrozumienia tej trudności stanowi zarzut formułowany czasem pod adresem NHST. Wielu autorów (np. Meehl, 1967; Cohen, 1994; Harlow, Mulaik i Steiger, 1997; Krantz, 1999; Loftus, 2002) słusznie zwraca uwagę, że hipoteza zerowa jest zwykle a priori niewiarygodna. Argument ten przybiera czasem postać podobną do tej, jaką sformułował Meehl:

> (...) μ_j można potraktować jako mierzalne wartości na osi liczb rzeczywistych. Identyczność dowolnych dwóch takich wartości oznacza, że różnica między nimi (również mierzalna wartość na osi liczb rzeczywistych) wynosi dokładnie zero, co jest [zdarzeniem] o zerowym prawdopodobieństwie.
>
> (s. 104, Meehl, 1967)

Gdyby na zbiorze wartości parametru ciągłego określić rozkład prawdopodobieństwa, na przykład rozkład normalny z niezerową wariancją, prawdopodobieństwo dowolnego zdarzenia byłoby równe powierzchni pod krzywą tego rozkładu. Ponieważ w przypadku hipotezy zerowej granice całkowania będą często równe (nie będą równe dla kierunkowej hipotezy zerowej), powierzchnia ta musi być równa zero. Zarzut ma jakoby obnażać zasadniczą wadę logiki testowania istotności hipotezy zerowej, wymaga jednak zignorowania jednego z centralnych założeń leżących u podstaw wszystkich standardowych metod wnioskowania statystycznego - parametr modelu nie jest zmienną losową.

W wielu wypadkach da się zrekonstruować logikę NHST w kategoriach subiektywistycznych w taki sposób, że częstościowe przedziały ufności będą się dosyć dokładnie pokrywały z tak zwanymi przedziałami wierności (ang. *credible intervals*) w ujęciu bayesowskim, które z kolei można już interpretować zgodnie z (3.2.2). Rekonstrukcja ta pozwala jednocześnie unaocznić głębsze problemy związane z wszystkimi standardowymi metodami wnioskowania statystycznego.

3.2.3 Bayesowska reinterpretacja testu istotności hipotezy zerowej

Aby uniknąć niepotrzebnych komplikacji, przebieg wnioskowania bayesowskiego zilustruję za pomocą możliwie prostego przykładu. Niech wyniki eksperymentu będą niezależnymi realizacjami zmiennej losowej o rozkładzie Bernoulliego z prawdopodobieństwem sukcesu θ. Badacz zainteresowany oceną hipotez na temat nieznanej wartości tego parametru może zastosować odpowiedni test istotności. Załóżmy, że w wyniku przeprowadzenia badania w zbiorze 20 obserwacji pojawiło się 11 sukcesów. Najczęściej stosowane w takim przypadku przedziały ufności, oparte na przybliżeniu rozkładu dwumianowego za pomocą rozkładu normalnego dla $\alpha = 0{,}05$ wynoszą w tym przykładzie $(0{,}32,\ 0{,}76)$. Badacz może więc odrzucić między innymi hipotezę zerową $\theta = 0$, albo $\theta = 0{,}1$ i nieskończenie wiele innych.

Aby dla tego samego problemu przeprowadzić wnioskowanie bayesowskie, należy określić rozkład aprioryczny na wartościach parametrów modelu (tutaj parametr jest tylko jeden). Rozkład ten ma wyrażać początkową niepewność co do wartości parametrów, jeszcze przed zebraniem wyników. Z aksjomatów rachunku prawdopodobieństwa wynika następujące równanie, znane jako reguła Bayesa:

$$f(\theta|y) = \frac{f(y|\theta)f(\theta)}{f(y)}$$

Równanie to staje się podstawą wnioskowania statystycznego, o ile zostanie przyjęta pozwalająca na to (na przykład subiektywistyczna albo tak zwana obiektywistyczna bayesowska) interpretacja prawdopodobieństwa. Funkcja $f(\theta)$ jest wtedy apriorycznym rozkładem prawdopodobieństwa, określonym na wektorze wolnych parametrów, $f(y)$ apriorycznym prawdopodobieństwem danych, a $f(\theta|y)$ rozkładem aposteriorycznym.

Funkcja wiarygodności $f(y|\theta)$ to rozkład prawdopodobieństwa danych ze względu na model. W momencie, gdy dane zostały już zebrane, y jest wielkością stałą i funkcja ta staje się funkcją wektora wolnych parametrów. Aprioryczny rozkład prawdopodobieństwa $f(\theta)$ wyraża początkową (przed zebraniem danych) niepewność co do wartości wektora wolnych parametrów, a $f(y) = \int_{\theta\in\Theta} f(y|\theta)f(\theta)\, d\theta$ to aprioryczne prawdopodobieństwo danych. Wymagające całkowania (lub sumowania, gdy parametry są dyskretne), występujące w mianowniku aprioryczne prawdopodobieństwo danych spełnia jedynie rolę stałej normalizacyjnej, gwarantującej, że obszar pod rozkładem aposteriorycznym będzie się sumował do jedności i można je zwykle pominąć, co pozwala zapisać regułę Bayesa w prostszej postaci:

$$f(\theta|y) \propto f(y|\theta)f(\theta)$$

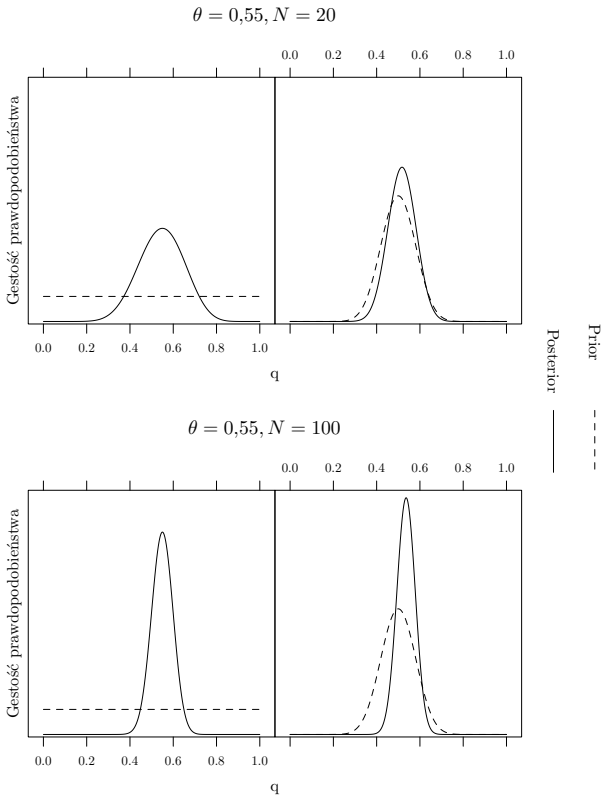
gdzie $\propto$ oznacza proporcjonalne do. Aposterioryczny rozkład prawdopodobieństwa $f(\theta|y)$ wyraża niepewność co do możliwych wartości parametru po tym, jak dane zostały zaobserwowane. Wnioskowanie bayesowskie polega zwykle właśnie na uaktualnianiu niepewności dotyczącej wartości wolnych parametrów na podstawie uzyskanych wyników.

Rozkład $f(\theta|y)$ jest dokładnie tym, co wymaga ustalenia, aby możliwa była czytelna interpretacja w duchu (3.2.2). Zgodna na potraktowanie parametrów modelu jako zmiennych losowych, a przedziałów ufności jako wyrażających niepewność co do wartości parametrów po zebraniu wyników prowadzi do rozmaitych interesujących konsekwencji. Zgoda na istnienie rozkładu aposteriorycznego na parametrach implikuje, na mocy aksjomatów rachunku prawdopodobieństwa, zgodę na istnienie rozkładu apriorycznego. Interpretacja (3.2.2) wymaga więc zgody na istnienie zarówno rozkładu aposteriorycznego jak i apriorycznego. Wyjaśnię teraz na przykładzie, w jaki sposób rozkład aposteriori zależy od rozkładu apriorycznego i danych.

Poniżej przedstawiłem cztery rozkłady aposterioryczne (linia ciągła), obliczone dla dwóch przykładowych zbiorów danych ($n = 20$ i $n = 100$, obserwowana proporcja sukcesów jest wszędzie równa $0{,}55$), odpowiadające dwóm różnym rozkładom apriorycznym (linia przerywana), określonym na nieznanym prawdopodobieństwie sukcesu θ. Rozkłady aprioryczne po lewej stronie (rozkłady jednostajne) odpowiadają całkowitej niepewności co do możliwych wartości θ (każda możliwa wartość jest apriorycznie jednakowo prawdopodobna). Rozkłady aprioryczne po stronie prawej (rozkład Beta z parametrami $\alpha = 20$ i $\beta = 20$) odpowiadają silnemu przekonaniu, że prawdopodobieństwo sukcesu wynosi około $0{,}5$. Ze zrozumiałych względów rozkłady takie jak te po lewej nazywane są nieinformacyjnymi lub płaskimi. Z rozkładu aposteriorycznego 95 procentowe przedziały *wierności*[4] można uzyskać odcinając krańce tego rozkładu, odpowiadające z każdej strony obszarowi pod krzywą równemu $0{,}025$. Przykład pozwala

[4]Tego rodzaju przedziały można określić na różne sposoby na rozkładzie aposteriorycznym, w zależności od pytania, na które mają udzielać odpowiedź. Wyczerpujące omówienie tego zagadnienia znajdzie Czytelnik między innymi w doskonałej pracy Gelmana, Carlina, Sterna i Rubina (1995).

zademonstrować kilka charakterystycznych cech wnioskowana bayesowskiego. Dla wygody będę się odtąd zamiennie posługiwał określeniami „rozkład aprioryczny" („aposterioryczny") i „prior" („posterior").



Rysunek 3.1: Przykładowe rozkłady aprioryczne i aposterioryczne dla kilku możliwych wyników, pochodzących z rozkładu Bernoulliego.

Posteriory mają wyrażać całą wiedzę, jaką badacz posiada po zebraniu obserwacji, zgodnie z aksjomatami rachunku prawdopodobieństwa w interpretacji subiektywistycznej. Niepewność reprezentowana przez posterior jest tym większa, im mniej informacji zawartej jest w priorze (im bardziej jest on płaski) i im mniej informacji na temat parametru dostarczają dane. Rozkład aposterioryczny jest więc zawsze wypadkową ilości informacji jaką niosą ze sobą dane i początkowej niepewności co do parametrów modelu. Im większa jest początkowa niepewność, tym bardziej „dane mówią same za siebie". Jeżeli początkowa niepewność jest mała (oparta na wynikach wielu poprzednich eksperymentów lub też na odpowiedniej teorii), minimalizowany będzie wpływ błędu próby. Wreszcie, w miarę jak rośnie wielkość próby, funkcja wiarygodności zaczyna dominować nad rozkładem apriorycznym, co widać na wykresach obrazujących rozkłady apo-

sterioryczne dla $n = 100$. Posługując się posteriorem, badacz może z łatwością udzielić odpowiedź na pytanie, w jakim zakresie z określonym prawdopodobieństwem znajduje się prawdziwa wartość parametru, albo z jakim prawdopodobieństwem wartość znajduje się w wybranym zakresie. W rzeczy samej, ponieważ posterior z założenia zawiera całą wiedzę na temat parametru uzyskaną po zebraniu danych, możliwe jest łatwe udzielenie odpowiedzi na dowolne pytanie probabilistyczne.

Krytykując NHST niektórzy autorzy (np. Cohen, 1994; Roberts i Pashler, 2000; Loftus, 2002, i wielu innych) zwracają uwagę na fakt, że poziom istotności związany jest z prawdopodobieństwem danych ze względu na hipotezę zerową ($f(y|h_0)$)[5], podczas gdy, jak twierdzą ci autorzy, badacza interesuje przede wszystkim prawdopodobieństwo hipotezy alternatywnej ze względu na dane ($f(h_1|y)$). Cytowany przez członków TFSI Cohen (1994) ujmuje to następująco:

> Co jest nie tak z NHST? Cóż, pomijając wiele innych kwestii, [NHST] nie mówi nam tego, co chcemy wiedzieć, a tak bardzo chcemy wiedzieć to, co chcemy wiedzieć, że z desperacji i tak wierzymy, że nam to mówi. Tym, co chcemy wiedzieć jest „W oparciu o te dane, jakie jest prawdopodobieństwo, że h_0 jest prawdziwa?".
>
> (s. 997, tamże)

Obliczenie tej wielkości staje się możliwe dzięki określeniu wszystkich elementów koniecznych do przeprowadzenia wnioskowania bayesowskiego. Z reguły Bayesa wynika wtedy natychmiast, że $f(h_0|d)$ zależy nie tylko od $f(d|h_0)$, ale także od aprioryczne-go rozkładu prawdopodobieństwa określonego na parametrach modelu. W kontekście wnioskowania statystycznego zgoda na przypisywanie hipotezom prawdopodobieństw oznacza jednak zgodę na paradygmat bayesowski. W konsekwencji standardowe procedury wnioskowania tracą sens. Z perspektywy bayesowskiej w większości sytuacji nie ma powodów, aby zgodnie z zaleceniami TFSI wprowadzać poprawki dla wielokrotnych porównań, stosować przybliżone metody wnioskowania oparte na standardowej teorii asymptotycznej zamiast dokładnych, zgadzać się na komplikacje związane ze standardowymi metodami metaanalizy, czy w szczególny sposób traktować rezultaty eksperymentu, który został przerwany przez badacza na podstawie uzyskanych wyników (Berger, 1980; Gelman i in., 1995; Robert, 2001). Być może właśnie ta ostatnia z wymienionych konsekwencji stanowiska bayesowskiego jest najtrudniejsza do przełknięcia dla zwolenników metod częstościowych. Rozważane podejścia do wnioskowania statystycznego opierają się na innej filozofii i często prowadzą w praktyce do różnych wyników, chyba że wnioskowanie bayesowkie będzie przeprowadzone w sposób, który ma je upodobnić do standardowego.

[5]Poziom istotności jest *związany* z tą wielkością, ale nie jest z nią tożsamy. Istotność mówi o prawdopodobieństwie uzyskania wartości odpowiedniej statystyki takiej jak zaobserwowana *lub większej*. Można więc powiedzieć, że poziom istotności zależy częściowo od prawdopodobieństwa danych, które nie zostały faktycznie zaobserwowane (Berger, 1980; Robert, 2001).

W wielu wypadkach, na przykład przy założeniu ogólnego modelu liniowego, przedziały wierności, zwłaszcza dla większych prób, będą się dość dokładnie pokrywały z częstościowymi przedziałami ufności, *o ile priory będą nieinformacyjne* (Gelman i in., 1995). Przedziały wierności dla parametru θ w przykładzie z 20 obserwacjami i płaskim priorem wynoszą $(0{,}34,\ 0{,}74)$, a więc są bardzo podobne[6] do przedziałów ufności $(0{,}32,\ 0{,}76)$. W przypadku małych prób różnice (tutaj nieznaczne) wynikają między innymi stąd, że podobnie jak wiele innych częstościowych reguł decyzyjnych, przedziały ufności oparte są na teorii asymptotycznej i dla małych prób, a więc takich, które często występują w praktyce, nie są dokładne, w przeciwieństwie do przedziałów wierności, które są dokładne niezależnie od wielkości próby.

Przybliżona równoważność wielu typowych przypadków zastosowania NHST i wnioskowania bayesowskiego z nieinformacyjnymi priorami pozwala na reinterpretację częstościowych przedziałów ufności w kategoriach subiektywistycznych. Ta zalecana przez wielu autorów wrażliwych na ograniczenia NHST i dość powszechnie, nawet jeśli zwykle „nieformalnie" stosowana interpretacja wymaga zgody na dwa założenia.

Po pierwsze, prawdziwa hipoteza należy do zakładanego modelu. Po drugie, nieinformacyjny prior poprawnie wyraża początkową niepewność badacza co do możliwych wartości parametrów. Jeżeli nie jest spełnione pierwsze założenie, zarówno w przypadku metod częstościowych jak i bayesowskich, nie można zagwarantować nieobciążoności wnioskowania. Inaczej mówiąc, nie można zagwarantować, że wraz ze wzrostem wielkości próby prawdopodobieństwo, że stosowanie metody wnioskowania doprowadzi do wyboru prawdziwej hipotezy będzie dążyło do jedności. Nieobciążoność jest tą własnością, bez której trudno uznać jakąkolwiek procedurę wnioskowania za użyteczną. Jeżeli drugie założenie nie jest spełnione, nie jest jasne, jak należałoby nadać sensowną interpretację posteriorowi. W praktyce prawie zawsze z góry wiadomo, że oba te założenia nie są spełnione w większym lub mniejszym stopniu. Stosowane modele są tylko przybliżeniami prawdziwego stanu rzeczy, a subiektywna niepewność rzadko, jeżeli w ogóle, daje się jednoznacznie wyrazić za pomocą rozkładu prawdopodobieństwa. Badaczowi decydującemu się na zastosowanie wnioskowania bayesowskiego pozostaje mieć nadzieję, że odstępstwa od założeń modelu nie są zbyt duże i być może również sprawdzić, w jakim stopniu wyniki zależą od przyjętych priorów, na przykład stosując przynajmniej kilka różnych rozsądnych rozkładów.

Pierwsze założenie (prawdziwość modelu) jest tak samo kłopotliwe dla obu stron sporu. Jak trafnie zauważył Rissanen (2003):

> Najczęściej nie dysponujemy wystarczającą wiedzą na temat maszynerii z której pochodzą dane, żeby móc ją wyrazić za pomocą rozkładu prawdopodobieństwa, którego próbki byłyby statystycznie podobne do zaobser-

[6]Gdy zastosować nieinformacyjny prior Jeffreya, czyli rozkład Beta z $\alpha = \beta = 1/2$, podobieństwo dodatkowo wzrasta.

> wowanych danych, przez co ostatecznie zmagamy się z niemożliwym do rozwiązania zadaniem, polegającym na szacowaniu tego, co nie istnieje.
>
> (s. 4, tamże)

Można jednak argumentować, że drugie założenie stanowi wyzwanie przede wszystkim dla (niekonsekwentnego) zwolennika metod częstościowych. Jeżeli zgodzić się na interpretację przedziałów ufności w kategoriach rozkładu prawdopodobieństwa określonego na wartościach parametrów, trzeba się jednocześnie zgodzić na stosowanie za każdym razem priorów nieinformacyjnych. Stanowisko to czasami trudno obronić nawet wtedy, gdy dany eksperyment przeprowadzany jest po raz pierwszy.

Niech jednorazowo przeprowadzony eksperyment polega na sprawdzeniu, czy pewna osoba posiada zdolność jasnowidzenia. W tym celu osoba ta jest proszona o odgadywanie kolejnych 10 wyników rzutu względnie rzetelną monetą[7]. Załóżmy, że w wyniku tego eksperymentu osobie badanej udało się odgadnąć 9 z 10 wyników. Rezultat pozwala odrzucić hipotezę zerową mówiącą o przypadkowości wyników zgadywania na poziomie $p = 0{,}02$, a przecież badacz, tak jak ewentualni odbiorcy jego wyników, będzie w tej sytuacji nadal przekonany, że zdolność jasnowidzenia jest fikcją. Wiadomo z góry, że taki rezultat prędzej czy później musi się pojawić. Podobnych niedorzeczności nie da się uniknąć polegając na standardowej logice testowania istotności, natomiast w wielu sytuacjach można ich uniknąć przeprowadzając wnioskowanie bayesowskie, dzięki możliwości uwzględnienia początkowej niepewności. W przykładzie z odgadywaniem wyników rzutu monetą badacz mógłby między innymi zastosować priory wyrażające silne i uzasadnione początkowe przekonanie, że wyniki zgadywania są przypadkowe, przez co wyniki eksperymentu musiałyby być o wiele bardziej przekonujące, żeby wniosek o ewentualnych niezwykłych zdolnościach osoby badanej mógł się okazać „statystycznie istotny".

Badania nigdy nie są realizowane w próżni teoretycznej i zgadzając się na określoną teorię i jej formalizację badacz jednocześnie musi przyznać, że żywi pewne przekonania na temat parametrów modelu. Jeżeli uważa, że między pewnymi zmiennymi występuje dodatnia korelacja, wydaje się, że przekonanie to powinno znaleźć odzwierciedlenie w wyborze priorów. Praktyka (niejawnego) stosowania priorów nieinformacyjnych staje się jeszcze bardziej niezrozumiała, gdy określony eksperyment był już przeprowadzony w przeszłości i dotychczasowe wyniki są badaczowi znane. Uzbrojony w możliwość wyboru priorów bayesianista zawsze może sprawdzić, do jakich wniosków prowadzą różne punkty widzenia reprezentowane przez różne priory, uzyskując tym samym, być może paradoksalnie, rodzaj intersubiektywnej obiektywności niedostępnej w ujęciu częstościowym.

Jedną z możliwych linii obrony NHST zreinterpretowanej jako wnioskowanie bayesowskie z priorami nieinformacyjnymi jest konserwatywność takiego postępowania.

[7] Przykład zaczerpnąłem z Bergera (1980), zmieniając go nieznacznie.

Można powiedzieć, że badacz wyprowadza wnioski na podstawie maksymalnie „niezobowiązujących" założeń co do prawdopodobnych wartości parametrów. Skoro jednak przedziały ufności mają wyrażać niepewność co do wartości parametrów, nie mogą jednocześnie wyrażać niepewności co do przyszłych obserwacji. Z perspektywy bayesowskiej ta ostatnia wielkość znajduje poprawny wyraz w aposteriorycznym rozkładzie predykcyjnym, czyli rozkładzie określonym na przyszłych obserwacjach, przy uwzględnieniu aposteriorycznej niepewności co do wartości parametrów:

$$\int_{\theta \in \Theta} f(y|\theta) f(\theta)\, d\theta$$

gdzie $f(\theta)$ to tym razem rozkład aposterioryczny. Wyznaczone w ten sposób przedziały predykcyjne są z konieczności szersze niż przedziały wierności (w przykładzie ze zmienną Bernoulliego i $n = 20$ wynoszą $(0{,}25,\ 0{,}85)$). Niepewność związana z wartością parametrów jest dodatkowo spotęgowana przez niepewność wynikającą z niedeterministycznego charakteru zmiennej obserwowalnej. Z perspektywy bayesowskiej NHST (stosowana niekonsekwentnie) polega na przecenianiu apriorycznej niepewności co do parametrów i niedocenianiu aposteriorycznej niepewności co do przyszłych obserwacji. Jak okaże się w dalszej części tego rozdziału, z ważnych względów należy uznać aposteriorycznyrozkład predykcyjny, a dokładniej ściśle z nim związaną wiarygodność brzegową, za użyteczną miarę uogólnialności modelu, czego nie można powiedzieć o częstościowych przedziałach ufności. Trafna ocena statystycznej uogólnialności modelu jest natomiast jedynym sposobem ilościowej oceny modelu skutecznie uwzględniającym jego złożoność.

Eleganckie rozwiązanie w paradygmacie bayesowskim znajduje również problem kumulacji wsparcia uzyskanego przez wielokrotne przeprowadzenie tego samego lub podobnych eksperymentów. Dla każdego takiego eksperymentu trzeba jedynie ustalić posterior i zastosować go jako rozkład aprioryczny wobec danych z kolejnego eksperymentu, poza tym wnioskowanie przebiega bez zmian. W ten sposób informacja uzyskana w jednym eksperymencie jest uwzględniana w kolejnym badaniu, przez co metoda wnioskowania staje się czymś w rodzaju metody uczenia się. Rozwiązanie oferowane przez paradygmat częstościowy polega z kolei na przeprowadzeniu złożonej technicznie metaanalizy, czyli na zastosowaniu osobnej, zaprojektowanej specyficznie dla tego celu procedury, której wyniki znowu interpretowane są zwykle, zgodnie z zaleceniami wielu autorów, w kategoriach subiektywistycznych.

Być może najczęściej formułowanym zastrzeżeniem wobec metod bayesowskich jest „subiektywny" charakter wnioskowania. Trudno jednak przytoczyć choćby jeden przekonujący powód, dla którego metody częstościowe należałoby uznać za bardziej obiektywne. W tych przypadkach, w których wnioskowanie bayesowskie z nieinformacyjnymi priorami daje wyniki takie same lub bardzo podobne jak wnioskowanie częstościowe wypada uznać, że element subiektywny jest tak samo nieunikniony, tyle że standardowe procedury wnioskowania opierają się na niejawnym przyjęciu określonego, dosyć szcze-

gólnego rozkładu apriorycznego, a wnioskowanie bayesowskie umożliwia wybór tego rozkładu. Element subiektywny jest również nieusuwalnie związany z wyborem, będącego przecież zawsze jedynie przybliżeniem, modelu probabilistycznego. Należy jednak pamiętać, że wnioskowanie bayesowskie to nie to samo co wnioskowanie w oparciu o subiektywistyczne rozumienie prawdopodobieństwa. Istnieją alternatywne dla subiektywistycznej interpretacje prawdopodobieństwa, na przykład obiektywistyczne bayesowskie (Press, 2003), pozwalające w podobny sposób korzystać z reguły Bayesa.

Jednym z powodów utrzymującej się długo niskiej popularności metod bayesowskich były trudności techniczne związane z obliczaniem aposteriorycznych rozkładów brzegowych. Nietrywialne modele zawierają zawsze co najmniej kilka wolnych parametrów. Aby ustalić analitycznie przedziały wierności dla określonego parametru takiego modelu trzeba zwykle obliczyć wielowymiarowy rozkład aposterioryczny, określony na całym wektorze parametrów, a następnie scałkować ten rozkład po wszystkich pozostałych parametrach. Do niedawna głównym sposobem ominięcia tych wymagań było stosowanie tak zwanych priorów sprzężonych (Fink, 1995), co znacznie ograniczało zakres praktycznie stosowalnych modeli[8]. Większość tych problemów została rozwiązana w latach 80-tych i 90-tych, między innymi dzięki zastosowaniu metod symulacji Monte Carlo z łańcuchami Markowa, a odpowiednie, również darmowe oprogramowanie jest od dawna powszechnie dostępne. Nadal jednak przeprowadzenie wnioskowania bayesowskiego wymaga ustalenia funkcji wiarygodności, co jest albo bardzo trudne, albo praktycznie niemożliwe w przypadku większości modeli wyrażonych w postaci procedury symulacji, na przykład złożonych modeli algorytmicznych, czyli między innymi wszystkich ważniejszych współczesnych „zintegrowanych modeli umysłu".

3.2.4 Uwagi na temat warunków użyteczności testu istotności hipotezy zerowej

Czytelnika przywiązanego do standardowych metod wnioskowania pragnę zapewnić, że daleki jestem od twierdzenia, jakoby przeważająca większość prezentowanych w literaturze naukowej wyników i wniosków opartych na tych wynikach była błędna. Jak sądzę, można stosunkowo dokładnie określić zakres sytuacji, w których NHST nie tylko nadmiernie nie przeszkadza, ale wręcz wyraźnie pomaga w ocenie wartości poznawczej hipotez. Jeżeli z teorii wynikają predykcje, które trudno inaczej *wyjaśnić*, a wyniki badania są wyraźnie zgodne z tymi predykcjami, istotność będzie wysoka, przedziały ufności wąskie, a hipoteza badacza uzyska silne wsparcie.

Jeżeli jednak obserwowany wzorzec można wyjaśnić na wiele sposobów, co wydaje

[8]Priory sprzężone to takie parametryczne rozkłady aprioryczne, że rozkład aposterioryczny należy do tej samej rodziny, a wartości parametrów posteriora można obliczyć na podstawie parametrów rozkładu apriorycznego i odpowiednich statystyk z próby. Unika się w ten sposób kosztownego całkowania, jednak ceną jest ograniczenie się do wykładniczej rodziny rozkładów prawdopodobieństwa, a więc stosunkowo wąskiego zakresu dopuszczalnych modeli.

się stanowić normę w badaniach psychologicznych, istotność testu statystycznego będzie miała niewiele wspólnego z rozstrzygnięciem między wiarygodnymi alternatywnymi wyjaśnieniami, ponieważ wszystko o czym badacz może się dowiedzieć na podstawie takiego wyniku, to niskie prawdopodobieństwo danych ze względu na hipotezę zerową. Skoro zaś obserwowany, niezgodny z hipotezą zerową wzorzec daje się wyjaśnić na wiele sposobów, hipoteza zerowa musi być a priori mało wiarygodna. Wartość informacyjna takiego rezultatu nie jest szczególnie imponująca. Być może jednym z powodów, dla których NHST do pewnego stopnia może „sprawdzać się w praktyce" jest nagminne, zarazem słuszne i niekonsekwentne naginanie zasad, na których procedura ta jest oparta. Wypada więc przyznać rację autorom raportu TFSI, kiedy stwierdzają, że „Dobre teorie i błyskotliwe interpretacje posuwają na przód dziedzinę bardziej niż sztywna, metodologiczna ortodoksja" (s. 603).

Niezależnie od tego, czy jest się bayesianistą czy nie, ważnym sposobem na wydatne zwiększenie konkluzywności wniosków płynących z wyników badań jest testowanie modelu reprezentującego wiernie możliwie wiele kluczowych założeń teorii. Między innymi dlatego prędzej czy później badacz jest zwykle zmuszony do skorzystania z modelowania matematycznego. Aby jednak modelowanie matematyczne doprowadziło do wzrostu konkluzywności wniosków, trzeba spełnić szeregu warunków, z których jeden, to jest uwzględnienie złożoności modelu, jest głównym tematem tego rozdziału.

3.3 Znaczenie modelowania matematycznego w procesie badawczym

Wiele matematycznych modeli mechanizmów poznawczych, określanych zwykle jako modele obliczeniowe, wyróżnia się złożonością. Może się wydawać, że szczegółowość tych modeli jest ich zaletą, wymuszającą ujawnienie ukrytych założeń i eksplikację pojęć, które w innym wypadku mogłyby być traktowane jako same przez się zrozumiałe. Proces modelowania matematycznego opiera się zawsze na pewnych (niekoniecznie jawnych) założeniach metateoretycznych, wymaga (niekoniecznie jawnie przeprowadzanej) metateoretycznej analizy i może prowadzić do ważnych metateoretycznych wniosków, na przykład dotyczących wieloznaczności pojęć, wewnętrznej niespójności teorii w postaci werbalnej, czy wreszcie wymagań, jakie musi spełniać badanie, aby jego wyniki można było uznać za wsparcie dla poddawanych ocenie hipotez. Do podstawowych zastosowań modelowania matematycznego w naukach empirycznych należy ilościowa ocena hipotez.

W moim odczuciu terminem „model matematyczny" określa się w psychologii zwykle modele nie dające się przedstawić jako szczególne przypadki modelu liniowego. Z tą (celowo nieostrą) charakterystyką nie zgodziliby się prawdopodobnie autorzy niektórych tekstów wprowadzających do modelowania matematycznego w psychologii. Na przykład, Myung i Pitt (2002) twierdzą:

> Modele matematyczne (...) stanowią próbę określenia wzorców zachowania poprzez zadawanie bezpośrednich pytań na temat mechanizmu leżącego u podłoża tego zachowania. (...) Dokonując modelowania matematycznego badacze formułują hipotezy na temat mechanizmu używając formuł matematycznych, algorytmów lub innych procedur symulacji, wymuszając w ten sposób jasność i precyzję. W rezultacie modelowanie matematyczne umożliwia a nawet wymaga wyprowadzenia precyzyjnych predykcji na podstawie przyjmowanych założeń, co zwiększa szanse na rozstrzygnięcie między hipotezami i modelami. Zalety tego podejścia są widoczne szczególnie wtedy, gdy predykcje i wyniki nie są oczywiste.
>
> Jak ujmuje to Luce (1995), modelowanie matematyczne jest podejściem do badania psychologicznego typu „otwartej czarnej skrzynki", w przeciwieństwie do [podejścia typu] „zamkniętej czarnej skrzynki" [właściwego dla] modelowania werbalnego.
>
> (s. 431, tamże)

Powyższa charakterystyka nie opisuje trafnie bodaj pierwszego psychologicznego nieliniowego modelu matematycznego, to jest tak zwanej krzywej zapominania, odkrytej pod koniec dziewiętnastego wieku przez Ebbinghausa (1885/1987). Autor ten zebrał wyniki imponującej liczby przeprowadzonych na sobie eksperymentów, polegających na uczeniu się i późniejszym odtwarzaniu list złożonych z ciągów liter („bezsensownych sylab"). Na podstawie tych danych udało mu się stwierdzić, że poziom wykonania spada najpierw gwałtownie, a potem coraz wolniej w funkcji odroczenia. Zależność sprawiała wrażenie na tyle systematycznej, że Ebbinghaus podjął próbę opisania jej za pomocą odpowiedniej funkcji matematycznej. Ostatecznie okazało się, że zaobserwowany wzorzec można w przybliżeniu opisać przez funkcję wykładniczą $R = e^{-t/s}$, gdzie t to czas odroczenia, s to „siła zapisu", a R to „ilość zapamiętanej informacji".

Przy całym szacunku dla przełomowych dokonań tego autora, krzywej zapominania nie sposób uznać ani za „sformułowanie hipotezy na temat mechanizmu", ani za „wyprowadzenie precyzyjnych predykcji na podstawie przyjmowanych założeń". Praca Ebbinghausa zawiera wiele błyskotliwych spostrzeżeń, eleganckich rozumowań i rzetelnych analiz[9], jednak ewentualna interpretacja parametrów i struktury tego modelu jest możliwa dopiero po jego sformułowaniu, ponieważ, tak jak wiele innych modeli w psychologii poznawczej, model ten został odkryty przypadkiem. Poszukując odpowiedniej funkcji Ebbinghaus nie korzystał z żadnych założeń dotyczących mechanizmu odpowiedzialnego za proces uczenia się.

[9]Tłumaczenie całości na język angielski wraz z tabelami zawierającymi wyniki eksperymentów znajduje się pod adresem $http://psychclassics.yorku.ca/Ebbinghaus$.

Krzywa zapominania, szczególnie gdy zostanie uzupełniona o dodatkowe wolne parametry[10], pozwala natomiast przypuszczalnie wyprowadzić bardziej precyzyjne predykcje dotyczące przyszłych obserwacji, niż to umożliwia werbalny opis obserwowanej zależności (stwierdzenie, że „poziom wykonania spada najpierw szybko, a potem coraz wolniej" nie jest ani zbyt odkrywcze, ani zbyt dokładne). Istnieje wiele modeli, którym przypisuje się rolę znacznie ważniejszą niż niekoniecznie pouczający, zwięzły *opis* obserwowanych regularności. Modele te mają być *wyjaśnieniami* odpowiedniego zachowania. Próbę ustalenia kryteriów podziału na modele opisowe i wyjaśniające podejmę w następnym rozdziale, teraz natomiast zajmę się zagadnieniem dobroci dopasowania rozumianego w sposób, który z wyjaśnianiem nie ma za wiele wspólnego. Krzywa zapominania nadaje się do tego celu znakomicie.

W cytowanej wcześniej pracy Myunga i Pitta nie udało mi się znaleźć ani jednej wzmianki o modelach konstruowanych metodą prób i błędów, bez żadnej określonej motywacji teoretycznej. Obraz jaki się stąd wyłania znacząco odbiega od współczesnej praktyki modelowania w psychologii. Na skutek gwałtownego rozwoju Sztucznej Inteligencji i metod statystycznych poszukiwanie modelu „pasującego do danych" stało się zajęciem w wielu wypadkach trywialnym (Roberts i Pashler, 2000), a ewentualne trudności nie muszą wcale wynikać z niezrozumienia modelowanego zjawiska, tylko z braku technicznych kompetencji lub niedostatecznego uporu badacza. Pomimo to, zagadnienie oceny ilościowej nadal nie jest zbyt dobrze rozumiane przez wielu użytkowników takich gotowych rozwiązań.

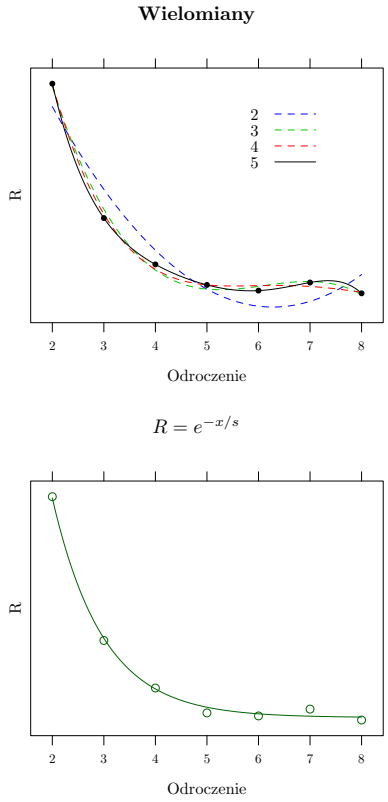
3.3.1 Standardowe metody oceny dobroci dopasowania

Zanim zastosuje się jakąkolwiek metodę oceny ilościowej trzeba zagwarantować, że model jest identyfikowalny i falsyfikowalny (Myung i Pitt, 2002). Identyfikowalność oznacza, że oszacowania wolnych parametrów są unikalne. Jeżeli badacz dysponuje dwoma punktami danych, zarówno model liniowy z wyrazem wolnym i parametrem nachylenia, jak i modele w nim zagnieżdżone są identyfikowalne, ponieważ do dwóch punktów danych pasuje dokładnie jedna hipoteza punktowa z każdego takiego modelu. Gdyby jednak model był wielomianem stopnia większego niż 1, zawsze istniałoby nieskończenie

[10]Współcześnie wykładniczy model zapominania uzupełnia się zwykle do postaci $R = (a-c)e^{-tb}+c$, gdzie c to asymptota, a to „współczynnik zapamiętywania", a b to „współczynnik zapominania" (Rubin, Hinton i Wenzel, 1999). Nie wiadomo, czy wszystkie te wolne parametry są rzeczywiście potrzebne. Jeżeli model z wolnym parametrem dla asymptoty będzie lepiej pasował do danych, najprostszym wyjaśnieniem tego faktu będzie stwierdzenie, że niezależnie od odroczenia istnieje niezerowe prawdopodobieństwo przypadkowego zgadnięcia, a tego prawdopodobieństwa nie da się kontrolować stosując na etapie testu elementy wcześniej prezentowane i nieprezentowane. Przypisanie parametrom psychologicznych etykiet nie zmienia faktu, że model wykładniczy nie opisuje przebiegu *zapominania*, tylko przebieg *wykonania zadań* pamięciowych. Nie są mi znane żadne przekonujące powody teoretyczne, przemawiające za tym, aby krzywą wykładniczą, potęgową, albo wiele innych podobnych uznać za coś więcej niż przybliżony *opis regularności nieznanego bliżej pochodzenia*, ujawniającej się w eksperymentach dotyczących pamięci.

wiele wartości wolnych parametrów, pozwalających idealnie odwzorować obserwowaną zależność. W takiej sytuacji parametrom nie sposób nadać sensowną interpretację. W kontekście oceny ilościowej falsyfikowalność modelu oznacza z kolei, że ze względu na charakter planu badawczego istnieją możliwe wzorce danych, których model nie będzie w stanie idealnie odwzorować. Na przykład, model liniowy z wyrazem wolnym i nachyleniem zawsze będzie w stanie odwzorować idealnie dwa punkty danych i będzie wtedy identyfikowalny. Model taki nie wnosi żadnej dodatkowej informacji poza dokładnym opisem tego, co w ogóle może zostać zaobserwowane w podobnym badaniu. Nie dostarcza ani lepszych predykcji niż opis werbalny, ani tym bardziej jakichkolwiek wyjaśnień.

Nadal najczęściej stosowane miary dobroci dopasowania i oparte na nich metody testowania promują silnie takie modele falsyfikowalne i identyfikowalne, które maksymalizują procent wariancji wyjaśnionej (Grünwald, 2007). Oczekiwane wartości danych ze względu na zidentyfikowany model różnią się wtedy jak najmniej od wartości faktycznie zaobserwowanych i w tym sensie „dobrze pasują" do uzyskanych wyników. Poniżej przedstawiłem dopasowanie krzywej zapominania i wielomianów stopnia od 2 do 5 do hipotetycznych wyników pomiaru poziomu odtworzenia w funkcji odroczenia.



Rysunek 3.2: Dopasowanie krzywej wykładniczej i wielomianów do danych pochodzących z modelu Ebbinghausa

Krzywa zapominania wydaje się całkiem dobrze, choć niedoskonale charakteryzować zależność między zmiennymi. Fakt ten znajduje odzwierciedlenie w wysokiej wartości współczynnika wariancji wyjaśnionej (tutaj $R^2 = 0{,}98$). Wielomiany stopnia 3 i 4 też pasują bardzo dobrze, a wielomian stopnia 4 nawet lepiej niż model prawdziwy (odpowiednio $R^2 = 0{,}96$ i $R^2 = 0{,}99$). Wielomian stopnia 5 pasuje idealnie ($R^2 = 1$), ale nie jest falsyfikowalny. Upraszczając można powiedzieć, że wielomiany stopnia od 3 do 5 odwzorowują zarówno systematyczną zależność między zmiennymi jak i przypadkową wariancję resztową[11]. Z dużym prawdopodobieństwem wyniki *przyszłego* eksperymentu, przeprowadzonego w podobnych warunkach, będą znacznie lepiej przewidywane przez oszacowaną *wcześniej* krzywą zapominania, niż przez oszacowany *wcześniej* wielomian. O modelach, które dopasowują się do zmienności przypadkowej mówi się, że są nadmiernie dopasowane. Model nadmiernie dopasowany jest więc z definicji kiep-

[11]Prawdziwy model nie jest w tym przypadku znany, dlatego zmienność przypadkową można określić tylko ze względu na zakładany model i nigdy nie będzie wiadomo, czy ta zmienność jest „rzeczywiście przypadkowa".

skim predyktorem przyszłych wyników, albo równoważnie, dysponując miarą zdolności do przewidywania przyszłych wyników można rozwiązać problem nadmiernego dopasowania.

Może się wydawać, że należy poszukiwać takiego modelu, ze względu na który procent wariancji wyjaśnionej będzie możliwie duży, ale jednocześnie model ten będzie falsyfikowalny. Konsekwentne stosowanie tego kryterium zawsze jednak doprowadzi do wyboru złego modelu, ponieważ współczynnik R^2 nie odróżnia zmienności systematycznej od przypadkowej i dla każdego nieidealnie pasującego modelu zawsze można znaleźć model pasujący również nieidealnie, ale trochę lepiej, czego przykładem jest dopasowany wcześniej wielomian 4 stopnia. W efekcie otrzymamy model, który „pochłania" część szumu i nie jest ani prawdziwy, ani szczególnie interesujący.

Jako miara dobroci dopasowania R^2 jest tylko statystyką opisową. To, co należy ustalić, to niepewność związana z zaobserwowanym dopasowaniem. Najczęściej stosowanymi metodami oceny tej niepewności są test χ^2 dla dyskretnych zmiennych obserwowalnych i test stosunku wiarygodności G^2 dla zmiennych dyskretnych lub ciągłych. Oba te testy są oparte na stosunku wiarygodności, a w praktyce różnice między nimi są zwykle niewielkie, dlatego w dalszej części ograniczę się do analizy ograniczeń związanych z ogólniejszym i zalecanym jako lepszy (Read i Cressie, 1988) testem G^2.

Do często cytowanych we współczesnej literaturze psychologicznej krytyków standardowych miar dobroci dopasowania należą Roberts i Pashler. W artykule pod znamiennym tytułem „How persuasive is a good fit? A comment on theory testing" (Roberts i Pashler, 2000) autorzy ci wymieniają szereg powodów, dla których typowe metody ilościowej oceny modeli matematycznych mają być zwykle pozbawione wartości. Za najsensowniejszą alternatywę należy ich zdaniem uznać metody oparte na idei indukcji eliminacyjnej. Zamierzam teraz przedstawić i przeanalizować argumenty Pashlera i Robertsa przeciwko standardowym metodom oceny ilościowej modeli matematycznych. Tok rozumowania tych autorów i wnioski jakie wyprowadzają zasługują moim zdaniem na uwagę, są bowiem wyraźnie i dobitnie sformułowane, a sam tekst jest często przychylnie cytowany[12].

3.3.2 Robertsa i Pashlera krytyka dobroci dopasowania

Opierając się na przeglądzie zawartości bazy Psychological Abstracts z lat 1987-1999 Roberts i Pashler stwierdzili, że w literaturze można odnaleźć „prawdopodobnie tysiące" przykładów zastosowania miar dobroci dopasowania jako głównego wsparcia dla modeli matematycznych z wolnymi parametrami. Według autorów, przypuszczalnie nie istnieją jednak takie teorie, które uzyskały wsparcie empiryczne głównie albo wyłącznie na podstawie dobrego dopasowania, a które uzyskałyby później wsparcie z innych źródeł,

[12]W samej bazie PsycArticles od momentu publikacji w 2000 roku tekst był cytowany 50 razy, a w bazie Google Scholar 314 razy.

takich jak potwierdzenie nowych predykcji. Problem wynika zdaniem autorów nie tyle ze specyficznych własności poszczególnych teorii, ale raczej ze stosowanej metody oceny. W szczególności, Roberts i Pashler zauważają, że dobroć dopasowania:

1. Nie mówi nic o elastyczności modelu, to jest o tym, do jakich możliwych wzorców model daje się dopasować.

2. Nie pozwala odróżnić modeli nadmiernie dopasowanych od dopasowanych tylko do zmienności systematycznej.

3. Nie mówi nic o tym, czy do tych samych danych nie pasują jakieś inne modele, oparte na innych założeniach.

4. Nie dostarcza informacji na temat zmienności obecnej w danych ze względu na zgodność z predykcjami teorii.

5. Nie informuje o wiarygodności możliwych wyników, do których teoria daje się dopasować.

6. Nie znajduje wsparcia w filozofii nauki.

7. Nie znajduje wsparcia w historii psychologii.

Propozycje autorów TFSI dotyczące testowania istotności można uznać za rozsądne w kontekście zastosowania modelu liniowego. Tego samego nie można powiedzieć o propozycjach dotyczących rozwiązania problemu złożoności modeli nieliniowych, ponieważ raport ten nie zawiera żadnych uwag na temat tego zagadnienia.

Rozwiązaniem ogólnego problemu oceny teorii psychologicznych ma być zdaniem Robertsa i Pashlera stosowanie pewnej wersji indukcjonizmu eliminatywistycznego, który w przeciwieństwie do dobroci dopasowania ma znajdować wsparcie w filozofii nauki. Do indukcjonizmu eliminatywistycznego i propozycji pokrewnych wrócę w następnym rozdziale, teraz natomiast zamierzam omówić dwa reprezentatywne rozwiązania problemów związanych z miarami dobroci dopasowania.

Wymienione w punktach 1-4 zastrzeżenia są słuszne, ale wszystkie te usterki da się z podobnym skutkiem usunąć albo za pomocą znanych od dawna metod bayesowskich, albo za pomocą nowszych metod opartych na teorii złożoności algorytmicznej (Li i Vitányi, 1997), takich jak metody oparte na zasadzie minimalnej długości kodu (Rissanen, 1986; Grünwald, 2005, 2007). Analiza własności tych metod pozwoli mi zidentyfikować pewne niepokojące konsekwencje złożoności modelu, które pozostają niepokojące nawet wtedy, gdy standardowe miary dobroci dopasowania zostają zastąpione przez miary skutecznie i w sposób teoretycznie uzasadniony uwzględniające złożoność.

3.4 Dwie alternatywy dla standardowych metod oceny dobroci dopasowania

Ocena zgodności między danymi a modelem, którą można określić jako standardową, przebiega nieco inaczej, gdy tylko opuścimy obszar ogólnego modelu liniowego. Mimo technicznych różnic, jej sens jest jednak ten sam. Decyzja o odrzuceniu bądź nieodrzuceniu hipotezy opiera się na wyniku rozumowania, w którym jedną z dwóch hipotez (hipotezę zerową) uznaje się za prawdziwą, oblicza się odpowiednią statystykę i porównuje ją z wartością krytyczną. Teoria leżąca u podstaw tego rozumowania pozwala ustalić coś na temat rozkładu z próby tej statystyki, w szczególności pozwala ustalić jakie jest prawdopodobieństwo, że statystyka przyjmie wartość większą lub równą od pewnej wartości krytycznej, związanej z przyjętym poziomem istotności. Dla ilustracji logiki standardowych metod oceny ilościowej modeli nieliniowych posłużę się symulowanymi obserwacjami, pochodzącymi ze wspomnianego wcześniej, wykładniczego modelu zapominania Ebbinghausa, uzupełnionego o składnik losowy ($R = e^{-t/0,2} + \epsilon$, $\epsilon \sim N(0, 0{,}02)$):

Test G^2 opiera się na stosunku największej wiarygodności ze względu na zakładany model i model w nim zagnieżdżony, reprezentujący hipotezę zerową. O modelu A mówi się, że jest zagnieżdżony w modelu B, jeżeli ma taką samą strukturę, a przestrzeń parametrów modelu A jest podzbiorem właściwym przestrzeni parametrów modelu B. Najczęściej taki model zagnieżdżony powstaje przez ustalenie wartości niektórych lub wszystkich wolnych parametrów na zero. W przypadku krzywej zapominania takim modelem zagnieżdżonym może być $R = e^{-t/1}$, można też porównać model Ebbinghausa z modelem bardziej złożonym, na przykład $R = ae^{-t/s}$.

Parametry każdego z dwóch testowanych modeli szacuje się metodą największej wiarygodności, to znaczy poszukuje się wektora parametrów $\hat{\theta} = \max_{\theta \in \Theta} f(x|\theta)$. Stosunek $LR = f(x|\hat{\theta}_B)/f(x|\hat{\theta}_A)$ informuje, o ile bardziej wiarygodne są dane ze względu na dopasowany model B (hipoteza alternatywna) niż A (hipoteza zerowa). Używa się terminu „wiarygodność" (ang. *likelihood*) zamiast terminu „prawdopodobieństwo" dlatego, że kiedy dane są znane, funkcja wiarygodności jest funkcją parametrów modelu, a w ujęciu częstościowym wartości parametrów nie mogą być mniej lub bardziej prawdopodobne[13].

Kiedy stosunek wiarygodności zostanie już obliczony, trzeba ustalić, czy uzyskana wartość wystarcza do odrzucenia hipotezy zerowej. W ogólnym przypadku dla modeli nieliniowych rozkład tego stosunku nie jest znany, ale wiadomo, że asymptotycznie, w miarę jak liczba obserwacji dąży do nieskończoności, przy założeniu hipotezy zerowej $G^2 = -2 \ln LR$ ma rozkład χ^2 z liczbą stopni swobody równą różnicy w liczbie wolnych parametrów. Im większa różnica w liczbie wolnych parametrów, tym większa wartość G^2 będzie potrzebna do odrzucenia hipotezy zerowej. Dla modelu Ebbinghausa i przytoczonego wcześniej modelu bardziej złożonego $G^2 = -0{,}69$, co okazuje się

[13] Takie uzasadnienie podał autor tego terminu, Fisher (Fisher, 1922).

wynikiem istotnym na poziomie 1 (test χ^2 z jednym stopniem swobody), a więc model prawdziwy zostaje utrzymany.

Wiarygodnymi alternatywami dla modelu takiego jak krzywa zapominania będą jednak zwykle modele o innej strukturze. Badacz będzie zainteresowany przede wszystkim oceną relatywnego dopasowania takich alternatywnych modeli, a nie modeli zagnieżdżonych w modelu testowanym albo modeli bardziej złożonych, w których model testowany jest zagnieżdżony. Testów takich jak G^2 czy χ^2 nie da się wtedy zastosować. Zasadniczą wadą tych testów jest również zredukowanie złożoności modelu do liczby wolnych parametrów, podczas gdy modele o tej samej liczbie wolnych parametrów mogą się bardzo różnić zdolnością do nadmiernego dopasowywania. Znacząco różnią się pod tym względem między innymi dwa jednoparametrowe klasyczne modele psychofizyczne, to jest Fechnera ($k \ln(x)$) i Stevensa (x^k), na co zwrócił uwagę już Townsend (1975).

3.4.1 Bayesowska selekcja modelu

Wnioskowanie na temat wartości wolnych parametrów na podstawie rozkładu aposteriorycznego jest tym samym co ocena hipotez punktowych należących do modelu na podstawie zebranych danych. Rozwiązanie problemu selekcji modelu w paradygmacie bayesowskim jest naturalnym uogólnieniem metody oceny hipotez punktowych na ocenę modeli. O ile, zakładając dany model, wnioskowanie na temat parametrów wymaga określenia priorów dla tych parametrów, o tyle ustalenie aposteriorycznego prawdopodobieństwa dla każdego z rozważanych modeli wymaga dodatkowo określenia rozkładu apriorycznego na zbiorze modeli. Gdy ten rozkład jest określony, regułę Bayesa stosuje się w taki sam sposób, jak w przypadku problemu oceny parametrów/hipotez punktowych:

$$p(M_k|y) = \frac{p(y|M_k)p(M_k)}{p(y)}$$

gdzie k to indeks modelu, $p(M_k)$ to aprioryczne prawdopodobieństwo k-tego modelu, $p(y) = \sum_k p(y|M_k)p(M_k)$ to niezależne od modelu, aprioryczne prawdopodobieństwo danych. Wielkość $p(y|M_k)$ to tak zwana wiarygodność brzegowa, dana wzorem $\int f(y|\theta_k) f(\theta_k)\, d\theta_k$, gdzie θ_k to wektor parametrów k-tego modelu. Gdy porównywane są dwa modele, można dla nich obliczyć stosunek prawdopodobieństw aposteriorycznych:

$$\frac{p(M_1|y)}{p(M_2|y)} = \frac{p(y|M_1)p(M_1)}{p(y)} \Big/ \frac{p(y|M_2)p(M_2)}{p(y)} = \frac{p(y|M_1)p(M_1)}{p(y|M_2)p(M_2)}$$

Wzór ten stosowany jest najczęściej przy założeniu jednakowych prawdopodobieństw apriorycznych dla obu modeli, to jest $p(M_1) = p(M_2) = 0{,}5$. Ulega wtedy dalszemu uproszczeniu do $K = p(y|M_1)/p(y|M_2)$ i pod tą postacią znany jest jako czynnik

Bayesa (Kass i Raftery, 1995). Wartość $K > 1$ oznacza, że dane silniej wspierają model występujący w liczniku, a $K < 1$ oznacza silniejsze wsparcie dla modelu w mianowniku. W przeciwieństwie do testu istotności hipotezy zerowej, którego szczególnym przypadkiem jest test G^2, żaden z modeli nie ma statusu uprzywilejowanego i nie pojawia się charakterystyczna dla NHST, kłopotliwa asymetria decyzyjna.

O ile można uznać zastosowane priory za poprawnie wyrażające początkową niepewność co do wartości parametrów/hipotez i modeli, aposterioryczne prawdopodobieństwo hipotez i aposterioryczne prawdopodobieństwo modeli jest dokładnie tym, co wielu autorów uważa najwyraźniej za właściwy cel wnioskowania statystycznego, a nawet procesu oceny wartości poznawczej hipotez. Trudno inaczej zrozumieć, dlaczego autorzy ci tak często przytaczają wielkość $p(h|y)$ jako główny przedmiot zainteresowania badacza.

W przeciwieństwie do wspomnianego wcześniej stosunku wiarygodności, czynnik Bayesa pozwala na porównywanie dowolnych, nie tylko zagnieżdżonych modeli parametrycznych. Ograniczenie do modeli zagnieżdżonych znacząco zmniejsza użyteczność metody oceny. W przeciwieństwie do skonstruowanego dla potrzeb testu G^2 modelu zagnieżdżonego lub ogólniejszego, zaproponowane przez innych autorów modele, a więc takie, które przede wszystkim należy uznać za alternatywne wyjaśnienia, zwykle będą miały zupełnie inną funkcję wiarygodności. W przeciwieństwie do miar takich jak AIC (ang. *Akaike Information Criterion;* Akaike, 1973) i BIC (ang. *Bayesian Information Criterion;* Schwarz, 1978), w opinii wielu użytkowników umożliwiających sensowne porównywanie modeli niezagnieżdżonych, czynnik Bayesa uwzględnia automatycznie i dokładnie, a nie tylko asymptotycznie, złożoność wynikającą nie tylko z liczby wolnych parametrów, ale również z postaci funkcyjnej modelu.

Aby zrozumieć, w jaki sposób złożoność uwzględniana jest we wnioskowaniu bayesowskim, należy zwrócić uwagę, co stanie się z wiarygodnością brzegową ($p(y|M)$) potrzebną do obliczenia czynnika Bayesa, gdy model zostanie uzupełniony o dodatkowy wolny parametr, a priory dla parametrów wyjściowych pozostaną bez zmian. Zakładając, że dane są niezależnymi obserwacjami pochodzącymi z rozkładu normalnego o zerowej średniej i nieznanej wariancji, a badacz jest zainteresowany porównaniem modelu liniowego ze średnią ustaloną na 0 (M_0) z modelem, w którym średnia jest wolnym parametrem (M_1), mamy:

$$p(y|M_0) = \int \prod_i f_N(y_i|\sigma)p(\sigma)\, d\sigma$$

$$p(y|M_1) = \int \left(\int \prod_i f_N(y_i - \mu|\sigma)p(\sigma)\, d\sigma \right) p(\mu)\, d\mu$$

gdzie i to indeks próby, a f_N to rozkład normalny o średniej 0. Jeżeli M_0 jest prawdziwym modelem, to o ile próba będzie wystarczająco reprezentatywna, wyrażenie $\int \prod_i f_N(y_i - \mu|\sigma)p(\sigma)\, d\sigma$ przyjmie maksymalną wartość dla średniej μ bliskiej zera i będzie wtedy w

przybliżeniu równe $p(y|M_0)$. Wymagane do obliczenia $p(y|M_1)$ dodatkowe całkowanie po średniej musi więc spowodować, że aprioryczne prawdopodobieństwo danych ze względu na M_1 będzie mniejsze.

Niepewność związana z dodatkowym wolnym parametrem zwiększa uwzględnianą przez czynnik Bayesa niepewność predykcyjną modelu. W podobny sposób można zademonstrować, jak wprowadzenie dodatkowego wolnego parametru może zwiększać niepewność aposteriorycznę co do wszystkich parametrów. Co szczególnie ważne, koszt związany ze złożonością zależy tutaj nie tylko od liczby wolnych parametrów, ale również od konkretnej postaci funkcji wiarygodności i od wielkości próby. Wpływ wielkości próby polega na tym, że wraz ze wzrostem n funkcja wiarygodności zaczyna dominować nad priorem.

Aby dostarczyć głębszego uzasadnienia dla bayesowskiej metody selekcji modelu i jednocześnie zwrócić uwagę na kilka problematycznych kwestii, związanych przypuszczalnie ze wszystkimi metodami oceny ilościowej uwzględniającymi złożoność, omówię teraz bodaj najbardziej reprezentatywną alternatywną metodę selekcji, opartą na zasadzie minimalnej długości kodu.

3.4.2 Selekcja modelu oparta na zasadzie minimalnej długości kodu

Dla Czytelnika, który nie miał wcześniej okazji do zapoznania się z podstawami wnioskowania bayesowskiego, sens tego podejścia może być początkowo trudny do uchwycenia, w każdym razie długo pozostawał taki dla mnie. Niemniej, mimo głębokich, częściowo niemożliwych do pogodzenia różnic, zarówno podejście częstościowe jak i bayesowskie łączy wspólny fundament w postaci rachunku prawdopodobieństwa. Metody selekcji modelu oparte na zasadzie minimalnej długości kodu, określane dalej za pomocą skrótu MDL, stanowią pod tym względem wyjątek. W wersji, którą zamierzam teraz omówić, metody te wywodzą się z połączenia rachunku prawdopodobieństwa z teorią informacji (Rissanen, 1986; Grünwald, 2005, 2007; Myung, Navarro i Pitt, 2006).

W obliczu konieczności uzasadnienia pewnych wniosków dotyczących tych metod uznałem za swój obowiązek przybliżyć niektóre ważne pojęcia, twierdzenia i dowody. Starałem się przy tym, aby treść pozostawała w miarę samowystarczalna, dlatego próbowałem unikać skrótów myślowych i te nieliczne dowody, które uznałem za warte przytoczenia, podaję pozostawiając możliwie mało domyślności Czytelnika. W stopniu, w jakim mi się to udało, podrozdział ten powinien być nieco bardziej zrozumiały, ale też mniej wyczerpujący niż opublikowane niedawno w literaturze psychologicznej, skądinąd bardzo dobre wprowadzenia Grünwalda (2005), Myunga, Navarro i Pitta (2006) i Schiffrina, Lee, Kima i Wagenmakersa (2008), na których zresztą ta część pracy była w znacznym stopniu wzorowana.

MDL jest formalizacją idei wnioskowania statystycznego, rozumianego jako wykrywanie występujących w danych regularności (Rissanen, 1986). Miarą sukcesu tak rozumianego wnioskowania jest osiągnięty stopnień kompresji danych, czyli zakodowania

danych za pomocą mniejszej liczby symboli, niż tego wymaga opis dosłowny. Spośród szeregu unikalnych cech tego podejścia dwa zasługują, ze względu na swój podstawowy charakter, na szczególną uwagę.

Pierwszą unikalną cechą jest wysłowienie i analiza problemu wnioskowania statystycznego w kategoriach kodowania, co pozwala między innymi na skonstruowanie osobnych miar złożoności modelu i danych. Drugą unikalną cechą jest porzucenie założenia o prawdziwości modelu, czyli założenia, że rozkład, z którego faktycznie pochodzą obserwacje, należy do modelu. Obie cechy wyraźnie odróżniają MDL i podobne metody od metod tradycyjnych i bayesowskich. Objaśnienie, w jaki sposób pojęcia kodowania i kompresji pozwalają na konstrukcję metod selekcji modelu uwzględniających w sposób teoretycznie uzasadniony złożoność wymaga przytoczenia kilku definicji i twierdzeń z teorii kodowania.

Dowolny zbiór danych można przedstawić jako ciąg symboli. Na przykład, wyniki pomiarów dwóch zmiennych da się przedstawić jako taki ciąg, że wartość zmiennej x dla i-tej jednostki eksperymentalnej (osoby lub grupy) występuje na miejscu $2i - 1$, a wartość zmiennej y dla tej samej jednostki znajduje się na miejscu $2i$. Dane zapisane są wtedy w postaci ciągu $x_1, y_1, x_2, y_2, \ldots, x_n, y_n$, gdzie n to liczba obserwacji. Stosując taki sam zabieg można „uszeregować" zapis dowolnego wielozmiennowego zbioru danych. Symbolami mogą być tutaj liczby zmiennoprzecinkowe o ustalonej precyzji, albo liczby zapisane w kodzie binarnym. Ta ostatnia reprezentacja znacząco ułatwia myślenie w kategoriach długości kodu. Ze względów technicznych będzie najlepiej, jeżeli od razu przejdę do ciągów binarnych, takich jak dwa przedstawione poniżej:

$$0101010101010101 \ldots 01010101010101\ (A)$$
$$1101101010000111 \ldots 00001010011110\ (B)$$

Na pierwszy rzut oka ciąg A wygląda na bardzo regularny. Znajduje to odzwierciedlenie w fakcie, że da się go opisać za pomocą wyrażenia „ciąg złożony z n wyrazów, w którym zero występuje na miejscach nieparzystych, a 1 na miejscach parzystych". Gdyby ciąg ten miał $n = 10^{10}$ wyrazów, przytoczony opis byłby znacznie bardziej zwięzły niż zapis dosłowny, o ile oczywiście wyrazy dłuższego ciągu wykazywałyby nadal tą samą regularność. Można powiedzieć, że ciąg A, a prawdopodobnie również kolejne symbole pochodzące z tego samego źródła, da się opisać przez proste prawo. Ciąg B wygląda z kolei na dość przypadkowy. B składa się z prób pobranych z rozkładu Bernoulliego z prawdopodobieństwem sukcesu równym $1/2$, wrażenie przypadkowości nie może być więc trafniejsze. Jak się wkrótce okaże, ciąg B jest w zasadzie niekompresowalny, to znaczy, z bardzo dużym prawdopodobieństwem długość najkrótszego opisu tego ciągu będzie niewiele mniejsza niż długość opisu dosłownego, niezależnie od długości ciągu.

Określenie kompresowalności ciągów wymaga zgody na metodę opisu, przez co należy rozumieć odwzorowanie różnowartościowe zbioru możliwych ciągów danych w pewien zbiór kodów (opisów), będących również skończonymi ciągami symboli. Gdyby

odwzorowanie nie było różnowartościowe, istniałyby kody nie dające się jednoznacznie interpretować. Pożądane jest również, aby metoda opisu pozwalała na przesyłanie skonkatenowanych („sklejonych") kodów w taki sposób, że odbiorca może je bezbłędnie odczytać. Warunek ten spełniają tak zwane kody prefiksowe, to jest takie, że żaden kod nie jest prefiksem innego kodu. Na przykład, opis przyporządkowujący ciągom danych binarnych 00, 01, 10 i 11 kody A, AB, BA i B odpowiednio nie jest prefiksowy, ponieważ kody A i B są prefiksami kodów AB i BA, przez co między innymi skonkatenowany ciąg danych 0011 nie jest jednoznacznie dekodowalny. Dla uproszczenia, zgodnie z konwencją przyjętą w literaturze, będę się czasem zamiennie posługiwał terminami „kod", „opis" i „metoda opisu".

Najbardziej ekspresyjną uniwersalną metodą opisu są języki programowania ogólnego przeznaczenia. Wybór ten prowadzi do przyjęcia definicji złożoności ciągu jako długości najkrótszego programu generującego dany ciąg i kończącego działanie, to jest tak zwanej złożoności Kołmogorowa. Niestety, tak zdefiniowana złożoność nie jest efektywnie obliczalna. Nie istnieje program, który dla każdego możliwego ciągu danych dostarcza najkrótszy program generujący ten ciąg i kończy działanie, co oznacza, że zawsze będą istniały możliwe regularności, których nie będzie się dało w ten sposób uchwycić.

Dodatkową trudność stanowi zależność stopnia kompresji od wyboru języka programowania, wprowadzająca do oceny złożoności element arbitralny. Okazuje się jednak, że dla dowolnych dwóch języków programowania i dowolnego ciągu danych x, długość najkrótszego kodu dla x wyrażonego w tych językach różni się co najwyżej o pewną stałą, niezależną od x. Jeżeli język A pozwala na wyrażenie danych z danego źródła za pomocą ciągów długości $\frac{1}{10}n$, a język B pozwala na kompresję w stopniu mniejszym lub równym niż $\frac{1}{10}n + 20$, gdzie n to długość ciągu danych, to gdy $n = 10000$ długość kodu w języku B jest większa co najwyżej o 2 procent. Wybór języka traci więc na znaczeniu wraz ze wzrostem długości ciągu danych. W zależności od języka i stopnia kompresowalności danych, wielkość próby potrzebna, aby różnica była praktycznie do pominięcia, może być jednak zbyt duża. Od użytecznej miary złożoności należy oczekiwać znikomego udziału elementu arbitralnego dla ciągów dowolnej długości, ponieważ wielkość próby jest zawsze mocno ograniczona względami praktycznymi.

Zanim przejdę do omówienia praktycznie użytecznych metod oceny złożoności, przytoczę za Grünwaldem (2005) dowód niskiej kompresowalności ciągów losowych, który pozwala zrozumieć związek między regularnością, losowością i kompresowalnością. Niech kodami będą dowolne skończone ciągi binarne, a zbiór danych obejmuje wszystkie n elementowe ciągi binarne, czyli opis jest funkcją odwzorowującą różnowartościowo zbiór n elementowych ciągów binarnych w zbiór ciągów binarnych skończonej długości. Niezależnie od tego, jaką metodę opisu zastosujemy, za pomocą ciągów binarnych jednoelementowych możemy zakodować co najwyżej dwie możliwe sekwencje, ponieważ istnieją tylko dwa ciągi binarne jednoelementowe, 0 i 1. Za pomocą ciągów dwuelementowych możemy zakodować co najwyżej cztery sekwencje, a ogólnie za pomocą ciągów k elementowych możemy zakodować co najwyżej 2^k sekwencji. Wobec te-

go, za pomocą ciągów maksymalnie k elementowych możemy zakodować co najwyżej $\sum_{l=1}^{k} 2^l = 2^{k+1} - 2$ sekwencji. Spośród wszystkich sekwencji n elementowych liczba tych kompresowalnych co najmniej o jeden bit, a więc kodowalnych za pomocą ciągów o długości co najwyżej $n-1$ wynosi nie więcej niż $2^{n-1+1} - 2 = 2^n - 2$. Jeżeli wszystkie możliwe n elementowe ciągi binarne są jednakowo prawdopodobne, to proporcja ciągów kompresowalnych co najmniej o k bitów jest równa prawdopodobieństwu kompresji co najmniej o k bitów i wynosi $(2^{n-k+1} - 2)/2^n < 2^{n-k+1}/2^n = 2^{1-k}$, maleje więc wykładniczo wraz z liczbą bitów kompresji. Na przykład, niezależnie od wybranej metody kodowania prawdopodobieństwo co najmniej 5 procentowej kompresji 1000 elementowego losowego ciągu binarnego jest mniejsze niż 2^{-49}. To właśnie należy odtąd rozumieć przez stwierdzenie, że ciągi losowe są w zasadzie niekompresowalne.

Całe to rozumowanie ma oczywiście sens tylko wtedy, gdy metoda kodowania jest wybrana przed uzyskaniem danych. Gdyby było inaczej, każdą możliwą sekwencję można by zakodować za pomocą ciągu jednoelementowego. Wiadomość taką dałoby się odczytać tylko wiedząc z góry, jaka jest jej treść, przez co metoda opisu nie nadawałaby się do komunikacji.

Gdy dane pochodzą z pewnego rozkładu prawdopodobieństwa, istnieje ścisły związek między kompresowalnością danych a tym rozkładem, wyrażony przez nierówność Krafta i twierdzenie o kodowaniu źródła Shannona. Nie ma potrzeby, abym w tym miejscu dokładnie przytaczał lub udowadniał te twierdzenia. Z nierówności Krafta wynika między innymi, że dla każdego efektywnie obliczalnego rozkładu prawdopodobieństwa $f(x)$, gdzie x to wektor danych o dowolnej ustalonej długości, istnieje kod prefiksowy C taki, że długość wektora x zapisanego za pomocą tego kodu wynosi $l_C(x) = -\lceil \log_2 f(x) \rceil$[14]. Dla niezależnych próbek z tego samego rozkładu prawdopodobieństwo ciągu n obserwacji wynosi $\prod_{i=1}^{n} f(x)$, czyli generalnie maleje wykładniczo wraz z n, a więc długość kodu $-\lceil \log_2 f(x) \rceil$ rośnie liniowo z n i błąd usuwany przez zaokrąglenie w górę traci na znaczeniu, co uzasadnia uproszczenie zapisu długości kodu do $\log_2 f(x)$. Twierdzenie Shannona o kodowaniu źródła mówi, że każdy taki odpowiadający rozkładowi prawdopodobieństwa kod prefiksowy jest optymalny w znaczeniu minimalizacji oczekiwanej długości kodu. Jak łatwo sprawdzić, taki kod będzie dawał opisy tym dłuższe, im mniejsze jest prawdopodobieństwo danych, lub równoważnie, będzie tym silniej kompresował ciągi danych, im częściej one występują. Dzięki temu, że kody te są w pożądanym znaczeniu optymalne, znika problem arbitralności metody opisu. Należy pamiętać, że przedmiotem zainteresowania nie jest tutaj nigdy konkretna postać kodu, tylko jego długość.

Na mocy nierówności Krafta i twierdzenia Shannona można utożsamić rozkłady prawdopodobieństwa z kodami prefiksowymi, czyli mówić o złożoności danych, rozumianej jako długość opisu, ze względu na jakiś rozkład prawdopodobieństwa. Pozostaje jeszcze wyjaśnić, jak można niearbitralnie określić złożoność (długość opisu) danych ze

[14] $\lceil z \rceil$ oznacza najmniejszą liczbę całkowitą większą lub równą niż z.

względu na model. Rozwiązanie tego problemu wymaga wprowadzenia pojęcia kodów i modeli uniwersalnych.

Kody i modele uniwersalne

Przypuśćmy, że nadawca i odbiorca zgodzili się wcześniej na używanie pewnego zbioru kodów L do przekazywania informacji na temat elementów zbioru możliwych ciągów danych D. Po zaobserwowaniu pewnego ciągu danych d, nadawca może znaleźć taki kod C należący do L, że C pozwala na najkrótszy opis d spośród wszystkich kodów należących do L. Takie postępowanie minimalizuje długość opisu dla wszystkich możliwych ciągów danych ze względu na zbiór L. Ponieważ jednak kod wybierany jest po zaobserwowaniu ciągu danych, a odbiorca nie wie z góry, jaki ciąg został zaobserwowany, metoda opisu nie może działać[15]. Taką niemożliwą metodę opisu, dającą zawsze najkrótszy możliwy opis ze względu na określony zbiór kodów, będę dalej nazywał „kodem jasnowidza"[16].

Dzięki ustalonej wcześniej możliwości utożsamienia rozkładów prawdopodobieństwa z kodami można teraz przeprowadzić analogiczne rozumowanie, traktując model jako zbiór rozkładów/kodów. Dane $x \in X$ mogą być ciągami n liczb binarnych, a model M rodziną rozkładów dwumianowych. Dla dowolnego x zawsze będzie istniała jakaś wartość parametru θ (prawdopodobieństwo sukcesu) maksymalizująca prawdopodobieństwo x. Ponieważ maksymalizacja prawdopodobieństwa oznacza minimalizację logarytmu prawdopodobieństwa, rozkład/kod wyznaczony przez oszacowanie θ metodą największej wiarygodności da najkrótszy możliwy opis x spośród wszystkich rozkładów/kodów należących do M. Taka idealnie zwięzła metoda opisu byłaby kodem jasnowidza ze względu na model, a więc nie może działać.

Okazuje się, że w wielu sytuacjach istnieją kody $\bar{L}$, które są prawie tak dobre jak kod jasnowidza, a dokładniej, takie kody, które dla każdego ciągu danych dają długości opisu różniące się od długości kodu jasnowidza co najwyżej o pewną stałą k, zależną tylko od modelu:

$$\forall x \left[\bar{L}(x) = \min_{C \in M} C(x) + k_M \right]$$

Takie „prawie idealne" kody określa się jako uniwersalne ze względu na model. Należy pamiętać, że kod uniwersalny nie musi należeć do M. Co ciekawe, bayesowska metoda selekcji modelu stanowi szczególny przypadek zastosowania kodów uniwersalnych. Przypominam, że czynnik Bayesa to stosunek brzegowej wiarygodności danych ze względ-

[15] Korzystając z nierówności Krafta można wykazać, że dla dowolnego zbioru kodów L nie istnieje kod, który dla każdego ciągu danych x daje opis długości nie większej niż opis x osiągalny za pomocą kodu dającego najkrótszy opis x spośród wszystkich kodów L.

[16] „Kod jasnowidza" nie jest technicznym terminem tej teorii. Wprowadziłem go dla zwiększenia czytelności wywodu.

du na testowane modele, w przypadku skończonej lub przeliczalnej przestrzeni parametrów Θ równy $\sum_\theta f(x|\theta)f(\theta)$. Brzegowa wiarygodność jest pewnym rozkładem prawdopodobieństwa, można ją więc utożsamić z kodem dającym długość opisu danych x równą $-\log_2 \sum_\theta f(x|\theta)f(\theta)$, a ponieważ $\sum_\theta f(x|\theta)f(\theta) \geqslant f(x|\theta_0)f(\theta_0)$ dla każdego $\theta_0 \in \Theta$ (suma wartości dodatnich jest nie mniejsza niż każdy z jej składników), to długość opisu danych ze względu na uniwersalny kod/brzegową wiarygodność wynosi:

$$-\log_2 \sum_\theta f(x|\theta)f(\theta) \leqslant -\log_2(f(x|\theta_0)f(\theta_0)) = -\log_2 f(x|\theta_0) - \log_2 f(\theta_0)$$

dla każdego $\theta_0 \in \Theta$. Długość ta różni się więc od długości ze względu na kod jasnowidza co najwyżej o pewną stałą. Gdy przestrzeń Θ nie jest ani skończona ani przeliczalna, sumowanie trzeba zastąpić przez całkowanie, poza tym wnioski pozostają bez zmian. Uniwersalność brzegowej wiarygodności to jeden z sygnałów intrygujących związków między metodami selekcji opartymi na minimalnej długości kodu a metodami bayesowskimi.

Dla dowolnego kodu uniwersalnego długość opisu będzie rosła wraz z n (prawdopodobieństwo ciągu danych maleje wraz ze wzrostem liczby obserwacji), wobec czego wpływ stałej będzie do pominięcia, o ile wielkość próby będzie odpowiednio duża. Pojawia się ten sam problem, co w przypadku złożoności Kołmogorowa. Niearbitralna miara długości danych ze względu na model powinna być niezależna od wyboru kodu uniwersalnego, trzeba więc poszukiwać takich kodów uniwersalnych, które byłyby w jakimś rozsądnym znaczeniu optymalne. Aktualnie jednym z popularniejszych, teoretycznie uzasadnionych rozwiązań tego problemu jest tak zwany rozkład znormalizowanej największej wiarygodności (NML, ang. *Normalized Maximum Likelihood*). Dla skończonego lub przeliczalnego zbioru X możliwych ciągów danych rozkład ten dany jest przez:

$$nml(x) = \frac{f(x|\hat{\theta}(x))}{\sum_{y\in X} f(y|\hat{\theta}(y))} \tag{3.1}$$

gdzie $\hat{\theta}(x)$ to oszacowanie największej wiarygodności. Rozkład NML można w uproszczeniu potraktować jako usunięcie pewnej wady kodu jasnowidza, która sprawia, że nie może on działać. Stosowanie (niemożliwego) kodu jasnowidza polega na kodowaniu każdego ciągu danych za pomocą rozkładu maksymalizującego prawdopodobieństwo tego ciągu. Wynik sumowania tych prawdopodobieństw po wszystkich możliwych ciągach danych będzie większy od 1. Z nierówności Krafta wynika jednak również i taka konsekwencja, że dla każdego kodu prefiksowego C określonego na dowolnym przeliczalnym lub skończonym zbiorze możliwych ciągów danych X, suma długości opisów po wszystkich elementach X ze względu na C musi być równa $\sum_{x\in X} -\log_2 Q(x)$, gdzie Q to pewna funkcja taka, że $\sum_{x\in X} Q(X) \leqslant 1$. Wada kodu jasnowidza, polegająca na tym, że odpowiednik funkcji Q (największa wiarygodność danych ze względu na model)

daje sumę przekraczającą 1 jest usuwana przez stała normalizacyjną występującą w mianowniku rozkładu NML. Co więcej, z nierówności Krafta wynika także, że kod NML jest kompletnym kodem prefiksowym, to znaczy, żaden inny kod prefiksowy nie osiąga długości mniejszych lub równych dla każdego x i mniejszych dla pewnego x[17].

Względnie wyczerpujące omówienie twierdzenia, które świadczy o tym, że rozkład NML jest w interesującym znaczeniu właściwą metodą opisu, pozwalającą niearbitralnie określić kompresowalność danych ze względu na model, wymaga wprowadzenia fundamentalnego dla MDL pojęcia złożoności modelu:

$$\mathrm{COMP}_n(M) = \log_2 \sum_{x \in X} p(x|\hat{\theta}(x))$$

gdzie M to model, p to rozkład należący do modelu, $x \in X$ to zbiór danych wielkości n, a $\hat{\theta}(x)$ to oszacowanie największej wiarygodności. To jest po prostu logarytm wyrażenia występującego w mianowniku równania (3.1). Tak jak $\log_2 p(x|\hat{\theta}(x))$ można interpretować jako długość opisu danych ze względu na model za pomocą kodu jasnowidza, tak $\mathrm{COMP}_n(M)$ jest sumaryczną miarą kompresowalności *jakichkolwiek* n elementowych ciągów danych za pomocą kodu jasnowidza dla modelu M. Wartość tej miary informuje o stopniu, w jakim model potrafi się maksymalnie dopasować do wszystkich możliwych wyników danego eksperymentu. Modele bardziej złożone będą pasowały do większej liczby możliwych ciągów danych, wobec czego w ogólnym przypadku będą w mniejszym stopniu ograniczały zbiór przyszłych obserwacji, czyli będą dostarczały mniej informacji na temat prawdopodobnych przyszłych wyników.

Teraz niech M będzie dowolnym modelem, a p rozkładem niekoniecznie należącym do M, określonym na tym samym zbiorze możliwych ciągów danych X co model. Dla każdego takiego p i $x \in X$ można obliczyć żal wynikający z zastosowania p ze względu na model M:

$$-\log_2 p(x) - \left[\min_{q \in M} -\log_2 q(x)\right]$$

co można interpretować jako dodatkową liczbę bitów, potrzebną do zakodowania danych x za pomocą rozkładu p, w porównaniu z minimalną liczbą bitów, gdy używany jest kod jasnowidza dla modelu M. Jeżeli dla każdego x istnieje unikalna wartość wektora parametrów modelu θ, maksymalizująca prawdopodobieństwo danych ($\hat{\theta}(x)$), żal wynosi:

$$-\log_2 p(x) - \left[-\log_2 q(x|\hat{\theta}(x))\right]$$

Jednym ze sposobów uniezależnienia wartości tego wyrażenia od x jest przyjęcie maksymalnego żalu po wszystkich możliwych ciągach danych jako miary jakości kompresji za

[17] Kod jest kompletnym kodem prefiksowym wtedy i tylko wtedy, gdy Q sumuje się do 1, tak jak to ma miejsce w przypadku rozkładu NML.

pomocą rozkładu p ze względu na model M. Okazuje się, że rozkład NML daje najmniejszy maksymalny żal ze względu na model, na którym jest określony, albo inaczej mówiąc, jest „najmniej ryzykowną" metodą kompresji, jeżeli tylko wartość $\mathrm{COMP}_n(M)$ jest skończona[18]. Żal jest wtedy stały dla każdego x i wynosi dokładnie $\mathrm{COMP}_n(M)$. Zgodnie z zasadą minimalnej długości kodu, spośród rozważanych modeli należy wybrać ten, który minimalizuje niearbitralnie określoną długość opisu danych ze względu na model, czyli należy wybrać model i minimalizujący wielkość:

$$-\log_2 nml(x|M^{(i)}) = -\log_2 p(x|\hat{\theta}^{(i)}(x)) + \mathrm{COMP}_n(M^{(i)}) \tag{3.2}$$

to jest tak zwaną stochastyczną złożoność danych ze względu na model.

Minimalizacja (3.2) jest równoważna z minimalizacją G^2 tylko asymptotycznie. Kara za złożoność odróżnia MDL od omawianego wcześniej uogólnionego stosunku wiarygodności G^2 i innych podobnych metod. W przeciwieństwie do G^2, (3.2) nie redukuje miary złożoności do liczby wolnych parametrów, co pozwala na sensowne porównywanie modeli zagnieżdżonych lub niezagnieżdżonych. Można powiedzieć, że test G^2 i inne podobne metody opierają się na zastosowaniu kodu jasnowidza, czyli nazbyt optymistycznym oszacowaniu przyszłego sukcesu predykcyjnego modelu. Warto przytoczyć kilka teoretycznych powodów, dla których opisaną wyżej miarę złożoności i opartą na niej metodę selekcji modelu wypada uznać za uzasadnioną.

3.4.3 Związki między metodą selekcji opartą na minimalnej długości kodu i niektórymi metodami alternatywnymi

Selekcja oparta na modelach uniwersalnych jest blisko związana z wnioskowaniem bayesowskim. Jak już wspomniałem wcześniej, brzegowa wiarygodność jest szczególnym przypadkiem zastosowania modelu uniwersalnego, dlatego że każdy rozkład apriorycznyokreślony na modelu jest modelem uniwersalnym. W BMS (bayesowskiej metodzie selekcji) powinno się stosować modele uniwersalne poprawnie wyrażające subiektywne prawdopodobieństwo, w MDL stosuje się model minimalizujący maksymalną możliwą stratę kompresji, czyli maksymalny możliwy błąd predykcji.

Ograniczenie swobody wyboru priorów w MDL wydaje się z jednej strony pożądane, priory mogą być przecież arbitralne, jednak z drugiej strony czasami priory, które nie są najlepsze według kryteriów MDL, będą dokładnie tym, czego chcemy użyć, na przykład dlatego, że wiemy coś więcej na temat rozkładu w populacji (rozkłady wyników kwestionariuszowych, priory uzasadnione teoretycznie, i tak dalej). O bliskich związkach między BMS i MDL przekonuje także fakt, że w przypadku modeli będących

[18]Wartość $\mathrm{COMP}_n(M)$ jest skończona, jeżeli X jest skończony. W innym wypadku $\mathrm{COMP}_n(M)$ może być nieskończona. Wtedy rozkład NML będzie niezdefiniowany i nie będzie istniał żaden model osiągający stały żal dla każdego x. Rozwiązaniem tego problemu jest przyjęcie słabszego kryterium optymalności.

rodzinami wykładniczymi zastosowanie MDL daje rezultat bardzo podobny, a asymptotycznie identyczny jak zastosowanie czynnika Bayesa z tak zwanymi priorami Jeffrey'a[19] (Grünwald, 2007).

Nie jest jasne, czy różnice w traktowaniu priorów w MDL i BMS są aż tak głębokie, jak sugeruje to między innymi Grünwald (2005, 2007). Model uniwersalny w MDL jest szczególnym rodzajem priora, ale nadal priorem. Grünwald twierdzi, o ile mi wiadomo słusznie, że tego rodzaju priory nie są opisane w żadnym podręczniku do wnioskowania bayesowskiego. Być może jednak dałoby się uzupełnić teorię wnioskowania bayesowskiego o priory „ostrożne", a nie tylko nieinformacyjne. Nie jest oczywiste, czy całkowita początkowa subiektywna niepewność ze względu na model powinna wyrażać niepewność co do możliwych wartości parametrów (zwykłe płaskie priory), możliwych wyników (priory Jeffreya), czy może wyrażać taki rodzaj niepewności, który minimalizuje największy możliwy błąd predykcji (NML), albo oczekiwany błąd predykcji. Minimalizacja tej ostatniej wielkości jest podstawą metod opartych na tak zwanej minimalnej długości wiadomości (ang. *Minimum Message Length*, Wallace, 2005)), której zwolennicy sugerują, że dopiero takie podejście do wnioskowania jest „prawdziwie bayesowkie". Na pewnym poziomie zaawansowania teorii racjonalnej coraz trudniej powiedzieć, co powinno być maksymalizowane (albo minimalizowane), co zresztą czyni te teorie jeszcze bardziej interesującymi.

Pewna reinterpretacja MDL pozwala dostrzec związki między tą metodą a metodami walidacji krzyżowej i pokrewnymi. Dla danego wektora danych x można ustalić, jak model radzi sobie z przewidywaniem kolejnych obserwacji na podstawie oszacowań opartych na wszystkich obserwacjach wcześniejszych. Dla wektora x składającego się z n niezależnych obserwacji i pewnego rozkładu f, przez x^i oznaczając wektor złożony z elementów od pierwszego do i-tego mamy:

$$\begin{aligned} f(x) &= \prod_{i=1}^{n} f(x_i) \\ &= \prod_{i=1}^{n} \frac{f(x^i)}{f(x^{i-1})} \\ &= \prod_{i=1}^{n} \frac{f(x_i) f(x^{i-1})}{f(x^{i-1})} \\ &= \prod_{i=1}^{n} f(x_i | x^{i-1}) \end{aligned}$$

[19]Priory Jeffrey'a (Jeffreys, 1946) są szczególnym rodzajem priorów nieinformacyjnych. Cechą charakterystyczną tych priorów jest skądinąd problematyczna we wnioskowaniu bayesowskim niezmienniczość rezultatów ze względu na reparametryzację modelu. Zwykłe priory nieinformacyjne mają to do siebie, że zmiana postaci funkcyjnej modelu na równoważną sprawia, że takie priory mogą nie być już więcej nieinformacyjne, mimo że model jest tak naprawdę ten sam, a zmienił się tylko sposób zapisu.

Jeżeli teraz w ostatnim wyrażeniu zastąpimy x^{i-1} przez oszacowanie $\hat{\theta}(x^{i-1})$, zlogarytmizujemy i zmienimy znak, uzyskamy skumulowany błąd predykcji, albo skumulowaną stratę zwięzłości kodu $\sum_{i=1}^{n} -\ln f(x_i|\hat{\theta}(x^{i-1}))$.

Kluczowe znaczenie odgrywa tutaj fakt, że ocena dobroci dopasowania dokonywana jest za każdym razem ze względu na zgodność predykcji z obserwacją, która nie była wykorzystywana do obliczenia tego oszacowania. Można powiedzieć, że rozkład lub model potraktowany jest jako strategia predykcyjna - idea sekwencyjnej predykcji leży u podstaw tak zwanego podejścia „prekwencyjnego" (Dawid, 1984). Przy odpowiednim doborze estymatora, zastosowanie MDL daje wyniki zbliżone do wyboru modelu minimalizującego taki skumulowany błąd predykcji (Dawid, 1984). Podobnie działają liczne metody oparte na idei walidacji krzyżowej (Browne, 2000), polegające na ocenie dobroci dopasowania w oparciu o zbiór danych różny od tego, który był wykorzystany do oszacowania parametrów (zbiory walidacyjny i kalibracyjny). Chociaż można powiedzieć, że selekcja prekwencyjna jest szczególnym przypadkiem zastosowania ogólnie rozumianej walidacji krzyżowej, wiele popularnych wersji tego ostatniego rozwiązania nie ma żadnego głębszego uzasadnienia teoretycznego.

Trudno przecenić rolę uniezależnienia się od założenia o prawdziwości modelu. Na tym, prawie zawsze jawnie fałszywym (nawet jeżeli „prawdziwym w przybliżeniu") założeniu opierają się między innymi metody wnioskowania bayesowskiego i standardowe metody wnioskowania statystycznego. Niemniej, jak dotąd nie spotkałem się z żadnymi przekonującymi argumentami, które świadczyłyby o zasadniczo większej odporności MDL, albo innych metod nie wymagających tego założenia.

Należy pamiętać, że ani MDL, ani wnioskowanie bayesowskie w praktyce nie mają wiele wspólnego z coraz bardziej popularnymi ostatnio wskaźnikami dopasowania takimi jak BIC albo AIC. Obie te miary redukują złożoność modelu do liczby wolnych parametrów, dzięki czemu rozwiązanie problemu selekcji modelu staje się obliczeniowo trywialne, ceną jest jednak błędna ocena elastyczności. Żadne miary redukujące złożoność do liczby wolnych parametrów nie będą dobrym rozwiązaniem, ponieważ modele o tej samej liczbie wolnych parametrów mogą się bardzo różnić złożonością.

Sprawdziłem[20], jak często skróty BIC lub AIC występują w artykułach dostępnych w bazie PsycArticles. W ciągu ostatnich dziesięciu lat liczba takich publikacji systematycznie rosła - w latach 1999-2004 liczba publikacji zawierających wyrażenie „Bayesian Information Criterion", „Akaike Information Criterion" lub odpowiednie skróty nie przekraczała 15 na rok, ale już w latach 2005-2009 liczba takich publikacji na rok wahała się od 22 do 47. Liczba dostępnych w bazie PsycArticles publikacji zawierających wyrażenie „Minimum Description Length" lub skrót MDL jest taka sama jak liczba publikacji zawierających wyrażenie „Bayes Factor" i wynosi 1. Obie te publikacje (Pitt, Myung i Zhang, 2002; Trafimow, 2003) są tekstami metodologicznymi, a nie przykładami zastosowania opisanych metod w praktyce. Jeden z tych artykułów (Trafimow, 2003), opu-

[20] 19 sierpnia, 2009.

blikowany w Psychological Review, zawiera jakoby „zaskakujące wglądy dotyczące testowania hipotez i oceny teorii" wynikające z reguły Bayesa. Trafimow popełnia niestety elementarne błędy, wykazując się przy tym ewidentnym brakiem zrozumienia podstawowych założeń wnioskowania bayesowskiego, na co zwrócili uwagę w opublikowanym później komentarzu Lee i Wagenmakers (2003).

Przypuszczalnie trzeba będzie jeszcze poczekać, zanim waga i sens problemu statystycznej złożoności modelu dotrze do ogólnej świadomości badaczy. Nawet niezwykle zasłużony dla psychologii matematycznej Estes pisząc o MDL stwierdził, że „w konsekwencji [zastosowania nowszych metod selekcji modelu] możliwe stało się porównywanie tylko takich modeli, które nie różnią się znacznie złożonością. (...) Gdy porównywane modele stają się bardziej złożone, staje się coraz bardziej prawdopodobne, że model wybrany za pomocą strategii relatywnego testowania będzie po prostu tym, który okazał się być lepiej dopasowany do zmienności przypadkowej. Wobec tego, można oczekiwać, że strategia relatywnego testowania będzie skuteczna tylko jeżeli uwzględniane modele będą stosunkowo proste (z grubsza, będą zawierały nie więcej niż trzy do czterech parametrów szacowanych na podstawie danych)" (s. 7, Estes, 2002). Jako całość zacytowany fragment nie pasuje ani do metod standardowych, ani tych nowszych, nie wiadomo też na jakiej podstawie Estes uznał, że akurat cztery wolne parametry stanowią górną granicę akceptowalnej złożoności. W tym samym artykule Estes broni własnej, dosyć osobliwej wersji falsyfikacjonizmu, jako ogólnej metody oceny wartości poznawczej hipotez, opowiadając się w ten sposób po stronie strategii 20 pytań. Za taką lub inną odmianą falsyfikacjonizmu opowiada się także wprost lub niewprost wielu krytyków NHST, między innymi Ci z nich (Cohen, Loftus, Roberts i Pashler), których zacytowałem w tej pracy.

3.5 Podsumowanie

Metody bayesowska i minimalnej długości kodu nie wyczerpują dostępnych badaczowi narzędzi oceny dobroci dopasowania. Istnieje cała rodzina metod bayesowskich i metod opartych na pojęciu złożoności algorytmicznej, istnieje też wiele metod ateoretycznych, takich jak wymieniona wcześniej walidacja krzyżowa, albo metody oparte na bootstrapie. W praktyce metody te czasami dają wyniki podobne, ale często wyraźnie różne (Grünwald, 2007; Myung, Forster i Browne, 200; Myung i in., 2006; Wagenmakers i Waldrop, 2006). Stosowanie metody, która na podstawie subiektywnej opinii użytkowników i rezultatów mniej lub bardziej pouczających symulacji wydaje się czasem „działać w praktyce" nie jest rozwiązaniem ani trochę zadowalającym.

Typowe różnice w dobroci dopasowania nieliniowych modeli są małe[21]. Nieuchronną konsekwencją zgody na dowolne, sprawiające dobre wrażenie metody selekcji jest

[21]Na podstawie skromnej znajomości literatury jestem skłonny zaryzykować stwierdzenie, że różnice te często nie przekraczają pięciu procent „wariancji wyjaśnionej", a dobroć dopasowania poszczególnych modeli jest zwykle wysoka.

wzrost trudności w porównywaniu wyników między badaniami. Rozwiązaniem nie jest oczywiście stosowanie wszystkich dostępnych „obiecujących" metod. Wyprowadzenie czytelnych wniosków teoretycznych na podstawie uzyskanego w ten sposób, zwykle niejednoznacznego wzorca, byłoby ogromnym i rzadkim osiągnięciem.

Biorąc pod uwagę wszystkie wymienione do tej pory argumenty wypada zgodzić się z tezą Robertsa i Pashlera o nikłej użyteczności tradycyjnych metod oceny dobroci dopasowania. Wkraczając w obszar modelowania matematycznego nie można uniknąć porównywania modeli, które najczęściej nie są ani liniowe, ani zagnieżdżone. Nie mam nic przeciwko stosowaniu prostych rozwiązań wobec prostych problemów, takich jak testowanie istotności hipotezy zerowej w kontekście modelu liniowego. Jeżeli badacz chce wiedzieć, czy jakiś bliżej nieokreślony efekt występuje, czy zależność jest raczej dodatnia, czy ujemna, albo czy jakieś czynniki oddziaływują interakcyjnie czy nie, w wielu wypadkach nie będzie powodów do zastosowania bardziej wyszukanych rozwiązań. Standardowe metody wnioskowania są jednak w ogólnym przypadku bezużyteczne jako metoda oceny nietrywialnych modeli matematycznych i nie wygląda na to, aby ten stan rzeczy miał kiedykolwiek ulec zmianie. Jak już wspomniałem, cytowany wcześniej raport TFSI nie zawiera niestety nawet najmniejszej wzmianki na temat omówionych w tym rozdziale zagadnień.

MDL i BMS są reprezentatywnymi przykładami teoretycznie uzasadnionych metod selekcji modelu. MDL dostarcza osobnej miary złożoności, podczas gdy w BMS złożoność jest uwzględniana niejawnie. Trudną do przecenienia zaletą MDL jest uniezależnienie się od założenia o prawdziwości modelu. Zarówno teorie jak i ich formalne eksplikacje w postaci modeli matematycznych są tylko przybliżeniami prawdziwego stanu rzeczy, dlatego prawie zawsze na pewno są mniej lub bardziej fałszywe. Od metody oceny ilościowej należy wymagać, aby była odporna, w znaczeniu możliwie niskiej wrażliwości na odstępstwa od założeń (Wilcox, 1998; Huber, 2004; Wilcox, 2009).

Metoda oceny ilościowej musi uwzględniać złożoność modelu. Zarówno w przypadku MDL jak i BMS, a także innych nowszych metod selekcji, problem złożoności jest zredukowany do kwestii uogólnialności statystycznej. Być może inaczej się nie da, jednak uogólnialność statystyczna niekoniecznie jest tym, co należy maksymalizować, gdy podstawowy cel stanowi ocena ogólnie rozumianej wartości poznawczej hipotez.

Nawet zakładając, że stosowana w MDL lub podobnych podejściach miara złożoności jest właściwa w tym znaczeniu, że umożliwia najlepszą możliwą ocenę statystycznej uogólnialności modelu, w praktyce stosowanie takiej metody będzie prowadziło do trudności związanych z nieuniknioną zależnością tej miary od wielkości próby. O ile złożoność testowanych modeli przekroczy pewne granice, bliższy prawdy ale bardziej złożony model może być względnie systematycznie odrzucany na rzecz modelu prostszego, ponieważ wielkość próby potrzebna do wykorzystania dodatkowej złożoności modelu bliższego prawdy będzie nieosiągalna. Ta bardzo niepokojąca własność wynika wprost z założonego celu, jaki metody typu MDL mają realizować, to jest minimalizacji błędu przewidywania przyszłych obserwacji. Głębokie związki między MDL a BMS wskazują

na to, że BMS rozwiązuje w istocie ten sam albo bardzo zbliżony problem. O tyle o ile na przykład MDL jest oparta na dobrej teorii, dostarcza dobrego kryterium oceny realizacji tego celu i niczego więcej.

Miara uogólnialności statystycznej jest miarą jakości predykcji, czego w żadnym razie nie można powiedzieć o dobroci dopasowania. W tym sensie udaje się częściowo rozwiązać problem oceny ilościowej w przypadku tych teorii i modeli, wobec których zastosowanie falsyfikacjonizmu (eksperymenty krzyżowe) jest albo bardzo trudne, albo niemożliwe. W miarę wzrostu złożoności proporcja takich „słabo testowalnych" teorii i modeli będzie oczywiście rosła. Nie może być inaczej, skoro wzrost złożoności modelu oznacza wzrost liczby wymiarów przestrzeni modeli alternatywnych. Dlatego, o ile celem jest maksymalizacja jakości predykcji, jak dotąd miary uogólnialności są i przypuszczalnie pozostaną jedynym dostatecznie uniwersalnym, teoretycznie uzasadnionym rozwiązaniem.

Wszystkie współczesne modele zintegrowane wyrażone są w postaci procedury symulacji a nie funkcji wiarygodności, zawierają wiele lub bardzo wiele wolnych parametrów i funkcja wiarygodności dla tych modeli najczęściej nie jest znana. W zdecydowanej większości publikacji[22] zawierających wyniki świadczące rzekomo o wysokiej zgodności wynikających z tych modeli predykcji z obserwacjami, ilościowa ocena modelu polega na zastosowaniu współczynnika R^2, $RMSE$, czasami testu χ^2. Żadna z tych miar nie informuje o trafności predykcji, ponieważ żadna z tych miar nie uwzględnia poprawnie elastyczności modelu. Oznacza to, że *pomimo przeprowadzenia licznych badań, aktualnie niewiele można powiedzieć na temat trafność predykcji modeli zintegrowanych*.

Wiele modeli matematycznych w psychologii nie ma w solidnego uzasadnienia teoretycznego. Dotyczy to przede wszystkim niektórych modeli, które mają być formalną reprezentacją *mechanizmu* albo *struktury poznawczej*. Nie ma żadnych teoretycznych powodów, aby z góry wykluczyć, że przeszukiwanie pamięci krótkoterminowej działa szeregowo, równolegle, dyskretnie, w sposób ciągły, i tak dalej. To samo dotyczy każdego innego nieobserwowalnego mechanizmu. Zakładając, że w psychologii poznawczej nadal będą powstawały przede wszystkim tego rodzaju modele, kryterium jakości predykcji będzie przypuszczalnie często odgrywało niezwykle ważną albo wręcz decydującą rolę w procesie oceny wartości poznawczej hipotez. Jakość predykcji, oceniana czy to za pomocą wyników eksperymentów krzyżowych, czy też metod wnioskowania uwzględniających złożoność, albo jakichkolwiek innych środków, nie jest jednak ani jedynym, ani nawet najważniejszy kryterium oceny *teorii*, chociaż jest najważniejszym kryterium oceny *modelu jako narzędzia służącego do przewidywania*.

Pozostałe kryteria, takie jak uogólnialność i rozwojowość teorii reprezentowanej przez model, interpretowalność parametrów i moc wyjaśniająca wydają się znacznie mniej wymierne, co nie oznacza, że nie da się stwierdzić, w jakich warunkach będą lepiej lub gorzej

[22]Czytelnik może to łatwo sprawdzić zapoznając się między innymi z licznymi publikacjami dostępnymi za darmo na stronie domowej modelu zintegrowanego ACT-R (http://act-r.psy.cmu.edu).

spełnione. Treścią następnego rozdziału jest właśnie ogólniejsza, to znaczy wykraczająca poza ocenę trafności predykcji problematyka oceny wartości poznawczej hipotez. Poświęcone temu rozważania rozpocznę od omówienia ograniczeń cieszącego się w psychologii niezasłużoną popularnością falsyfikacjonizmu.

Rozdział 4

Jakościowa ocena hipotez

Proponowana przez Robertsa i Pashlera wersja falsyfikacjonizmu, który będę czasem nazywał krótko „eliminatywizmem", jest udoskonaloną i rozwiniętą wersją metody „efektywnego wnioskowania" (ang. *Strong Inference*) Platta (1964). Omówione w tym rozdziale koncepcje Platta, Robertsa i Pashlera są szczególnymi wersjami indukcjonizmu eliminatywistycznego, zgodnie z którym ocena wartości poznawczej hipotez polega na odrzucaniu tych hipotez spośród ogółu rozważanych, których predykcje okazały się niezgodne z obserwacjami. Koncepcja ta jest z kolei szczególnym przypadkiem falsyfikacjonizmu, który jest szczególnym przypadkiem probabilizmu, czyli poglądu głoszącego, że ocena wartości poznawczej hipotez sprowadza się do oceny ich prawdopodobieństwa ze względu na dane (Grobler, 2006).

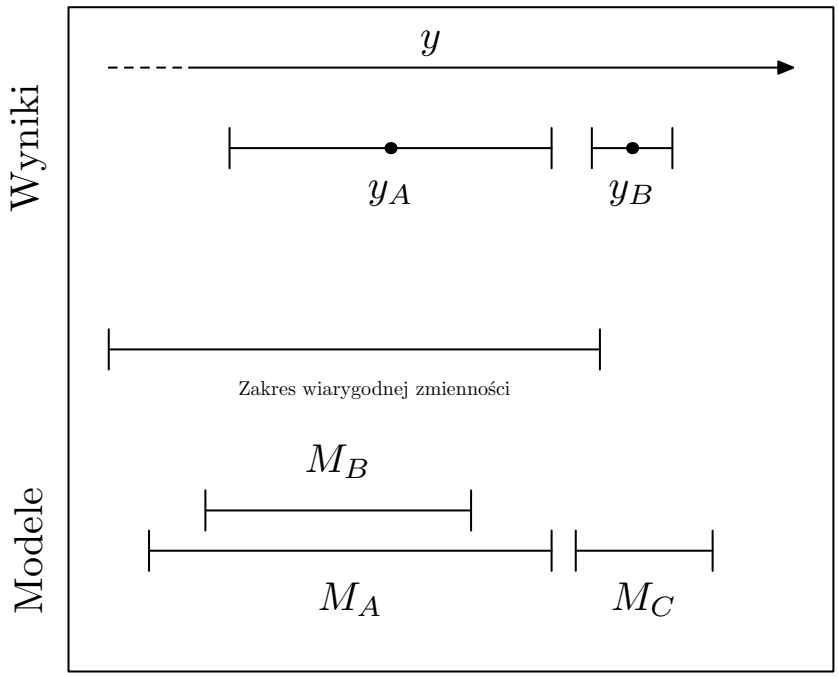
4.1 Indukcjonizm eliminatywistyczny

Zgodnie ze stanowiskiem eliminatywizmu, empiryczna ocena hipotez ma polegać na systematycznym odrzucaniu alternatywnych hipotez na podstawie niezgodności predykcji z wynikami badań. Zdaniem Robertsa i Pashlera stanowisko to znajduje wsparcie w filozofii nauki i oparte na nim metody są bodaj najskuteczniejszym sposobem oceny teorii ze względu na wyniki eksperymentu. Jak zauważają autorzy, efektywne wnioskowanie w praktyce stosowane jest niezwykle rzadko, co ma być jakoby jednym z ważniejszych powodów niezadowalającego postępu psychologii nie tylko poznawczej. W oryginalnej wersji Platta (1964) proces oceny hipotez ma przebiegać w czterech krokach:

1. Wyliczenie hipotez alternatywnych.
2. Zaprojektowanie jednego lub większej liczby eksperymentów krzyżowych, z których każdy powinien pozwolić, tak skutecznie jak to tylko możliwe, wykluczyć co najmniej jedną z alternatyw.
3. Przeprowadzenie eksperymentu w taki sposób, aby uzyskać czyste wyniki.

4. Powtórzenie procedury od początku, przez sformułowanie hipotez bardziej szczegółowych albo sekwencyjnych.

Robertsa i Pashlera wersja eliminatywizmu polega na ocenie relatywnego potwierdzenia predykcji (nie zredukowanego do dobroci dopasowania) danej teorii, w porównaniu do potwierdzenia predykcji alternatywnych, wyjaśniających to samo zjawisko teorii, z uwzględnieniem apriorycznej wiarygodności możliwych wyników i statystycznego charakteru wnioskowania. Eksplicitne uwzględnienie niepewności związanej z oszacowaniami, elastyczności teorii i apriorycznej wiarygodności wyników stanowi nowość w porównaniu do wersji Platta. Główna idea pozostaje jednak ta sama - rewizja oceny modeli (bo właśnie o modelach piszą Roberts i Pashler) polega na odrzucaniu tych modeli, których predykcje są niezgodne z wynikami badań. Z powodu wymienionych różnic, dla wygody wersję Platta będę dalej nazywał „eliminatywizmem prostym", a wersję Robertsa i Pashlera „eliminatywizmem statystycznym". Ten ostatni w zastosowaniu do dwóch modeli można przedstawić graficznie w postaci następującego diagramu:



Rysunek 4.1: Przykład zastosowania eliminatywizmu statystycznego Robertsa i Pashlera

Oś y reprezentuje dowolną miarę określoną na danych, y_A i y_B to przykładowe oszacowania przedziałowe wartości tej miary, a $M_i, i = A, B, C$ to przedziały predykcyjne trzech testowanych modeli. Miara ze względu na którą modele są testowane może być dowolnie skomplikowana, może to być efekt jakiegoś czynnika, miara interakcji, oszacowanie nachylenia linii regresji albo krzywizny i tym podobne. Nie musi być ona jednowy-

miarowa, ale dla uproszczenia, bez szkody dla dalszych wniosków, ograniczę się do przypadku jednowymiarowego. Kluczowe znaczenie dla zastosowania metody ma ustalenie predykcji teorii, rozumianej jako możliwe wzorce obserwacji (wartości miary), które teoria dopuszcza i które wyklucza (predykcje jako przedział). Eliminatywizm statystyczny uwzględnia zatem w pewien sposób złożoność modelu. Im bardziej model jest elastyczny, tym szersze będą przedziały predykcyjne.

Na zamieszczonym powyżej diagramie wynik y_A pozwala odrzucić model M_C, a wynik y_B pozwala odrzucić oba pozostałe modele. Gdyby uzyskano wynik y_B zgodnie z założeniami eliminatywizmu statystycznego wsparcie dla modelu M_C byłoby stosunkowo wysokie, ponieważ taki rezultat jest niespodziewany (wykracza poza przedział wiarygodnej zmienności), a predykcje modelu M_C i uzyskane wyniki są stosunkowo dokładne (wąskie przedziały).

W najprostszym przypadku, gdy rozważana jest tylko jedna teoria, wyniki mogą albo zgadzać się z jej predykcjami, albo nie. Jeżeli wyniki się zgadzają, ale są a priori dość prawdopodobne, albo oszacowanie miary jest obarczone dużą niepewnością, wsparcie dla teorii będzie słabsze. Rozumowanie to wydaje się zgodne z dosyć podstawową intuicją, mówiącą że potwierdzenie niespodziewanych konsekwencji teorii waży więcej, niż potwierdzenie konsekwencji znanych lub z góry prawdopodobnych. Jak doskonale ujął to Lakatos, cytowany zresztą przez wspomnianych autorów „Miarą sukcesu teorii Newtona nie jest to, że kamienie, gdy się je upuści, spadają w kierunku ziemi, niezależnie od tego jak często się to powtarza" (s. 6, Lakatos, 1978). Gdy testowana jest pojedyncza teoria, przyrost wiedzy w postaci rewizji oceny tej teorii nastąpi wtedy, gdy jej predykcje będą zarazem wyraźnie zgodne z obserwacjami (przedziały ufności pokrywają się z grubsza z przedziałami predykcyjnymi) i niespodziewane ze względu na wiedzę zastaną, albo wtedy, gdy wyniki pozwalają odrzucić tą teorię. Można powiedzieć, że w pierwszym przypadku odrzucana jest (trudna do określenia) część wiedzy zastanej.

Stąd, że teoria, która okazuje się niezgodna ze znanymi obserwacjami, może nie być zbyt dobra, nie wynika, że zgodność ze znanymi obserwacjami mówi coś interesującego o jej jakości. Zdaniem Robertsa i Pashlera (2000), w psychologii zgodność ze znanymi obserwacjami zwykle nie mówi nic interesującego, ponieważ „...dane psychologiczne często nie są zaskakujące. Dlatego zdolność do przewidywania takich danych nie może dostarczać zbyt wielkiego wsparcia dla jakiejkolwiek teorii" (s. 369, tamże). Przedziały aprioryczne (zakres wiarygodnej zmienności) mają pozwalać na uwzględnienie wiedzy zastanej, na którą mają się składać znane wyniki badań, zdrowy rozsądek i alternatywne teorie.

Jeżeli testowana jest więcej niż jedna teoria, niosącym najwięcej informacji i zarazem najbardziej przekonującym potwierdzeniem byłoby uzyskanie oszacowania miary z wąskimi przedziałami ufności, które mieści się w granicach możliwie wąskich predykcji tylko jednej z teorii. Oznaczałoby to, że wyniki są stosunkowo zaskakujące (przewiduje je tylko jedna teoria), wyraźnie wynikają z teorii (wąskie przedziały predykcyjne) i są wyraźnie zgodne z tymi predykcjami (wąskie przedziały ufności).

W kłopotliwej i zdaniem tych autorów częstej w psychologii sytuacji, gdy przedziały predykcyjne alternatywnych teorii znacznie się pokrywają, wyniki eksperymentu pozwalają dokonać rewizji oceny teorii tylko wtedy, gdy są niezgodne z co najmniej jedną teorią. Żeby tak się stało, wyniki jako przedział muszą wypadać poza przedziały predykcyjne. W innym wypadku wyniki uznaje się za tak samo zgodne ze wszystkimi rozważanymi teoriami, co oznacza brak przyrostu wiedzy w postaci rewizji ocen.

Omawiany w pierwszym rozdziale eksperyment Townsendta i Fifica jest szczególnym przypadkiem zastosowania tak rozumianego eliminatywizmu. Testowane są wprost wszystkie możliwe klasy modeli ze względu na zakładaną przestrzeń alternatyw. Uzyskane wyniki, dzięki odpowiednio zaplanowanemu eksperymentowi, wspierają dokładnie jedną z tych klas. Nawet w takiej sytuacji, wyniki testu nie są jednak konkluzywne z powodów przygodnych, to znaczy wynikających z nieusuwalnej niepewności co do spełnienia niezliczonych dodatkowych założeń, koniecznych do interpretacji danych, ale także z powodów zasadniczych.

Do zasadniczych źródeł niekonkluzywności wyników testu empirycznego należy niepewność co do spełnienia jawnych założeń wspólnych dla wszystkich alternatyw i niepewność co do znaczenia odstępstw od tych założeń dla predykcji, na przykład leżącego u podstaw teorii Townsenda założenia o dyskretności i bezbłędności hipotetycznych procesów, albo bezpośredniej selektywności wpływu zastosowanych czynników. W teorii Townsenda założenia te nie są testowane i przyjmuje się je, ponieważ umożliwiają pożądaną interpretację rezultatów eksperymentu. Innymi słowy, przestrzeń alternatyw może być źle zdefiniowana przez co teoria może nie być na tyle ogólna, aby zagwarantować poprawność wniosków. Nie wiadomo więc między innymi, czy kontrast funkcji przeżyciowej pozwala na identyfikację architektury i reguły stopu dla procesów ciągłych i niekoniecznie bezbłędnych[1]. Ponieważ zawsze trzeba coś, a właściwie bardzo wiele, założyć, tego rodzaju niekonkluzywność nigdy nie będzie do końca usuwalna.

O ile dobrze rozumiem Robertsa i Pashlera, ustalenie dobroci dopasowania miałoby odpowiadać z grubsza stwierdzeniu, czy punktowe oszacowanie miary opartej na wynikach badań znajduje się w bliżej nieokreślonym zakresie predykcji teorii. Przyjmując tą interpretację, dobroć dopasowania nie pozwala stwierdzić ani jak dokładne są te oszacowania, ani jak bardzo elastyczny jest model, ani tego, jak bardzo wyniki są zgodne z innymi teoriami. Porównywanie dopasowania między teoriami miałoby natomiast mówić o tym, czy oszacowanie punktowe miary znajduje się w zakresie predykcji rozważanych teorii, ale już nie o tym, jak bardzo ograniczony jest zakres predykcji każdej z teorii, ani jak bardzo prawdopodobne a priori jest to oszacowanie. Zgodnie z tym, co napisałem w poprzednim rozdziale, zastosowanie MDL, BMS lub innych podobnych metod selekcji usuwa te usterki.

[1]Townsend twierdzi, że nieopublikowane wyniki symulacji wskazują na zadowalającą niewrażliwość wniosków na odstępstwa od założenia o bezbłędności (Townsend 2009, korespondencja osobista).

W porównaniu do wersji prostej, eliminatywizm Robertsa i Pashlera wymaga bardziej skomplikowanej analizy statystycznej i uzasadnienia przyjętych przedziałów apriorycznych, na pierwszy rzut oka nie wydaje się jednak wymagać wyliczenia wszystkich alternatywnych hipotez. Ta ostatnia cecha jest szczególnie ważna w psychologii, ponieważ problemy badawcze w psychologii są często słabo zdefiniowane, dlatego zbioru alternatyw nie można określić nawet w przybliżeniu, co uniemożliwia zastosowanie eliminatywizmu prostego. Propozycja, aby znane wyniki badań, alternatywne teorie i „zdrowy rozsądek" uwzględnić w ramach przedziałów apriorycznych, chociaż wprowadza element subiektywny, sprawia zatem wrażenie uzasadnionej.

Eliminatywizm jest stosunkowo prostą odpowiedzią na fundamentalne i trudne pytanie o empiryczne kryteria oceny teorii. Jak już wspomniałem, od momentu publikacji w roku 2000 w samej tylko bazie Google Scholar tekst Robertsa i Pashlera cytowany był 314 razy, a popularność wspomnianej publikacji Platta wydaje się wręcz oszałamiająca - artykuł doczekał się 1261 publikacji cytujących w bazie Google Scholar, dla porównania angielskie wydanie „Wstępu do psychoanalizy" Freuda jest cytowane w tej bazie 1356 razy. Zdaniem O'Donohue i Buchanana (2001), „Strong Inference" należy do najczęściej zalecanych lektur w podręcznikach do metodologii w naukach społecznych i przyrodniczych. Oba teksty są wybitnie perswazyjne, na przykład artykuł Platta rozpoczyna się zdaniem „Pewne systematyczne metody myślenia w nauce mogą doprowadzić do znacznie bardziej gwałtownego postępu niż inne".

W obu omawianych tutaj wersjach eliminatywizm nawiązuje do ważnych idei w filozofii nauki, to jest indukcji eliminacyjnej i eksperymentów krzyżowych Bacona, metody alternatywnych wyjaśnień Chamberlina (1965), falsyfikacjonizmu Poppera (1977), a w przypadku wersji statystycznej, dodatkowo do bayesianizmu i cytowanego wcześniej spostrzeżenia Lakatosa. Za autorami można więc powiedzieć, że eliminatywizm znajduje *pewne* wsparcie w filozofii nauki. Stanowisko to nie znajduje jednak szczególnie silnego wsparcia ani w historii nauki, ani we *współczesnej* filozofii nauki, a już na pewno we współczesnej filozofii nauki nie ma czegoś takiego, jak powszechna zgoda co do użyteczności tej metody, tak jak by tego chcieli autorzy wersji statystycznej. Wręcz przeciwnie, o ile można się doszukać czegoś przypominającego konsensus na temat eliminatywizmu, zgoda dotyczy raczej jego znikomej użyteczności. Żeby się o tym przekonać, wystarczy zapoznać się z zawartością niemal dowolnej, względnie aktualnej monografii poświęconej tym zagadnieniom. Z każdym krokiem metody Platta i wszystkimi nowinkami Robertsa i Pashlera wiążą się poważne trudności natury zasadniczej i praktycznej.

4.1.1 Wady indukcjonizmu eliminatywistycznego

Argumentując na rzecz eliminatywizmu prostego, Platt powołuje się na kilka przykładów z historii nauk empirycznych, konkretnie biologii molekularnej i fizyki wysokich energii. Przykłady te mają ilustrować skuteczność efektywnego wnioskowania, które zdaniem tego autora jest „tą jedną właściwą metodą naukową i zawsze nią było". Platt nie

podaje niestety żadnych bliższych informacji na temat przebiegu procesu badawczego w przypadkach, na które się powołuje. Na podstawie rzetelnej analizy tekstów źródłowych, przytaczając przy tym wiele cytatów z odpowiednich prac, O'Donohue i Buchanan (2001) wykazali, że ani jeden z cytowanych przez Platta przykładów nie pasuje do schematu efektywnego wnioskowania. Autorzy ci przedstawili również, znowu opierając się na przyzwoitym materiale dowodowym, interesujące omówienie zupełnie innych strategii badawczych, stosowanych świadomie i z doskonałym skutkiem przez Galileusza, Darwina, Lyella i Newtona.

W ujęciu tych autorów podejście stosowane przez Galileusza wydaje się najbliższe dedukcyjnemu wyprowadzaniu wniosków z możliwie przekonujących, ostrożnych założeń. Dokonując oceny hipotez Galileusz chętnie korzystał z pomocy eksperymentów *myślowych*. Darwin zdawał się często świadomie poszukiwać świadectw empirycznych niezgodnych z własnymi teoriami, jednak trudno się u niego doszukać prób „zderzania ze sobą" alternatywnych hipotez, albo w ogóle określania zbioru dopuszczalnych alternatyw. Gdy wyniki okazywały się niezgodne z predykcjami, teoria była najczęściej *korygowana*, a nie odrzucana. Z analiz O'Donohue i Buchanana wynika, że także podejście Darwina wydaje się lepiej pasować do rozumowania dedukcyjnego, opartego na ogólnych, popartych licznymi obserwacjami przesłankach. Nawet gdyby tego chciał, Lyell nie mógłby zastosować efektywnego wnioskowania, był bowiem geologiem i ze względu na podejmowane przez niego problemy badawcze metoda eksperymentalna była z reguły bezużyteczna. Do schematu rzekomo jedynej właściwej metody naukowej zupełnie nie pasuje też podejście Newtona, będące zdaniem Platta przykładem „rzadkich i indywidualnych osiągnięć, wykraczających poza jakąkolwiek regułę i metodę" (s. 351, Platt, 1964). O'Donohue i Buchanan (2001) zauważają też, podobnie jak Roberts i Pashler, że w praktyce eliminatywizm jest stosowany niezwykle rzadko. W przeciwieństwie do Platta, Robertsa i Pashlera autorzy Ci sugerują, że kiedy eliminatywizm jest stosowany, często zdarza się, że prowadzi do błędnych wniosków. Wobec powyższego wypada stwierdzić, że historia nauki nie dostarcza wsparcia dla tezy, jakoby efektywne wnioskowanie, a ogólnie falsyfikacjonizm, był podstawową metodą oceny wartości poznawczej hipotez w naukach empirycznych.

Oto jak Roberts i Pashler wyjaśniają, w jaki sposób można ustalać predykcje:

> ... przypuśćmy, że teoria [tak jak jest reprezentowana przez model] ma dwa wolne parametry: a, który może się zmieniać w zakresie od 0 do 10 i b, który może się zmieniać w zakresie od 0 do 1. Aby ustalić, jakie są [przedziałowe] predykcje teorii na przykład dla liczby prób do kryterium, należałoby ... ustalić predykcje dla wszystkich kombinacji wartości parametrów. Predykcjami tej teorii dla tego wymiaru byłby cały zbiór możliwych liczb prób do kryterium, które teoria potrafi wygenerować.
>
> (s. 364, tamże)

Dalej opisany jest przykład faktycznego zastosowania tej metody. Dla konkretnego modelu *dwuparametrowego* udało się ustalić, że ogranicza on zakres dopuszczalnych wartości miary opartej na czasach reakcji i że te ograniczenia są związane z wartością pewnej innej miary. Wyniki (dwuwymiarowe oszacowania przedziałowe) okazały się w kilku przypadkach wypadać poza obszar przewidywany przez model. Zdaniem Robertsa i Pashlera taką teorię należy odrzucić.

W przypadku większości modeli matematycznych w psychologii znaczna część albo wszystkie parametry nie mają żadnych ustalonych granic, a ich liczba może się czasem wahać od kilkunastu do kilkudziesięciu. Nawet zakładając, że zgodnie z zaleceniami Robertsa i Pashlera dla niektórych z tych parametrów można przyjąć uzasadniony, wiarygodny zakres zmienności, zbiór rozkładów z których należy próbkować rośnie kombinatorycznie wraz ze wzrostem liczby parametrów, a zbiór możliwych miar nie ma żadnych określonych granic. W ogólnym przypadku przeszukiwanie takiej przestrzeni byłoby świadectwem determinacji graniczącej z obłędem. Gdyby badacz zechciał uwzględnić wszystkie aktualnie dostępne teorie dotyczące danego zjawiska, granice obłędu byłyby dawno przekroczone.

Nie wiadomo, jak ocenić przyrost wiedzy wynikający z zastosowania eliminatywizmu prostego lub statystycznego, jeżeli zbiór *wszystkich* alternatywnych teorii nie jest znany, a przecież nigdy nie jest znany. Załóżmy, że w wyniku zastosowania tej metody udało się względnie konkluzywnie odrzucić dwa z pięciu rozważanych alternatywnych modeli. Jak silne jest wsparcie dla pozostałych trzech? Jeżeli na przykład można sformułować dziesięć alternatywnych modeli, opartych na wzajemnie wykluczających się założeniach psychologicznych, które również zgadzają się z wynikami tego eksperymentu, bez dodatkowej informacji wyniki będą bezużyteczne jako narzędzie oceny wszystkich tych założeń. Wniosek ten dobrze ilustruje omówiona w rozdziale pierwszym historia badań dotyczących szeregowości i równoległości. Z perspektywy eliminatywizmu miarą przyrostu wiedzy jest stopień zawężenia zbioru alternatyw. Niestety, jak już wspomniałem, zdecydowana większość problemów badawczych w psychologii jest słabo zdefiniowana, przez co nawet zbiór względnie intuicyjnych hipotez alternatywnych jest trudny do określenia. Kiedy zbiór alternatyw nie jest znany, trudno cokolwiek powiedzieć na temat wartości zastosowania tej procedury. Metoda wydaje się opierać na całkiem dorzecznych intuicjach, ale ewidentnie czegoś jej brakuje.

Następny problem jest związany z nieusuwalną niekonkluzywnością jakiegokolwiek testu empirycznego. Zgodnie z powszechnie akceptowaną przez współczesnych filozofów nauki tezą Duhema-Quine'a, nie istnieje obserwacja nieuteoretyzowana. Obserwowalne konsekwencje wynikają z teorii empirycznej tylko na mocy nie dającego się nigdy ostatecznie określić zbioru dodatkowych założeń, takich jak założenia co do sprawnego działania sprzętu zastosowanego do przeprowadzenia eksperymentu, odpowiedniego zrozumienia instrukcji przez osoby badane, selektywności wpływu, dyskretności i bezbłędności procesów w przypadku teorii Townsenda, ale także bardziej podstawowych założeń metodologicznych, metateoretycznych, a nawet epistemologicznych i ontolo-

gicznych. Po uzyskaniu wyników ocena przynajmniej niektórych hipotez powinna zostać zrewidowana, ale które to mają być hipotezy i jak należy dokonać tej rewizji jest zawsze kwestią wykraczającą daleko poza ramy konkretnego badania. Tego typu zastrzeżenia sprawiają mniejsze trudności, gdy przed przeprowadzeniem testu niektóre teorie są wyraźnie lepsze niż inne ze względu na moc wyjaśniającą, stopień uzasadnienia przez ogólniejszą teorię, rozwojowość całej klasy propozycji teoretycznych do której dana hipoteza należy, i tym podobne, jednak w psychologii takie kryteria są niemal bezużyteczne.

Mimo licznych nawiązań do wnioskowania bayesowskiego, eliminatywizm statystyczny nie odbiega zanadto od standardowej logiki wnioskowania statystycznego. Zgodnie z zaleceniami, należy uwzględnić wiarygodne hipotezy, ale ich wiarygodność nie jest stopniowalna. Na podstawie uzyskanych wyników hipoteza może być odrzucona bądź utrzymana, ale ani eliminatywizm statystyczny ani prosty nie zawiera zaleceń dotyczących rewizji stopniowalnej oceny hipotez. Propozycja, aby uwzględnić przedziały predykcyjne, przedziały apriorycznej wiarygodności i przedziały ufności nie wymaga traktowania hipotez jako mniej lub bardziej prawdopodobnych i nie jest jasne, w jaki sposób należałoby rozwinąć metodę, aby ten rodzaj niekonkluzywności został skutecznie uwzględniony bez konieczności poprawnego zdefiniowania przestrzeni alternatyw.

Historia nauk przyrodniczych obfituje w przykłady rezultatów, które okazywały się niezgodne ze skądinąd bardzo użytecznymi teoriami, a mimo to teorie te nie zostały odrzucone, ponieważ oznaczałoby to zbyt wielkie koszty. W psychologii jednak takie kryteria rzadko znajdują zastosowanie. Hipotezy nie mogą mieć wyraźnego związku z ogólniejszą i względnie powszechnie akceptowaną teorią, skoro taka nie istnieje. Relatywna rozwojowość całego podejścia, reprezentowanego przez daną hipotezę, na przykład koneksjonizmu albo stosowania modeli zintegrowanych, jest trudna do określenia i nie wydaje się stanowić przedmiotu jakichkolwiek szerzej zakrojonych badań metateoretycznych.

Do ważnych źródeł niskiej informacyjności wyników oceny ilościowej modeli należy ich często słaby związek z teorią. Tak jak wielu innych autorów, Roberts i Pashler traktują terminy „teoria" i „model" zamiennie, przez co nie rozróżniają między predykcjami teorii i predykcjami modelu, nie mogą więc dostrzec tego problemu, a przecież bez interpretacji w postaci teorii werbalnej model matematyczny jest tylko pozbawioną znaczenia, abstrakcyjną strukturą formalną.

O ile związek logiczny między teorią i modelem nie jest szczególnie silny, niezgodny z obserwacjami model może zostać nawet dość radykalnie skorygowany z zachowaniem założeń teorii i nie będzie w tym nic złego. Gdyby można się było dowiedzieć, że *żaden* model sensownie reprezentujący teorię nie pasuje zbyt dobrze do wyników jednego lub większej liczby eksperymentów, eliminatywistyczna rewizja oceny teorii byłaby bardziej uzasadniona. Gdyby można się było dowiedzieć, że teoria dopuszcza tak wiele modeli, że praktycznie nie ogranicza dopuszczalnych wyników jakiegoś eksperymentu, wiadomo byłoby przynajmniej, w jaki sposób nie da się tej teorii testować. Z reguły ani tego

pierwszego, ani tego drugiego nie możemy się dowiedzieć nawet w przybliżeniu, ponieważ język i struktura logiczna wielu teorii w psychologii poznawczej czyni je wybitnie podatnymi na interpretację. W konsekwencji braku silnego związku między teorią werbalną a modelem badacz nie ma powodu, aby zanadto przejmować się niekorzystnymi dla ulubionej teorii wynikami eksperymentu przypominającego krzyżowy, albo kiepską statystyczną uogólnialnością reprezentującego teorię modelu.

4.1.2 Podstawowa wada probabilizmu

Wszystkie wymienione wyżej problemy związane z eliminatywizmem, to jest nieokreśloność przestrzeni alternatywnych hipotez, trudności związane z ustalaniem niezgodnych predykcji nawet dla stosunkowo mało złożonych modeli, nieuwzględnienie związku modelu z teorią werbalną i wynikająca z uteoretyzowania obserwacji niekonkluzywność jakiegokolwiek testu empirycznego, chociaż istotne, odwracają uwagę od najważniejszego błędu popełnianego nie tylko przez Robertsa i Pashlera. Podstawowym kryterium oceny teorii w naukach empirycznych nie jest (relatywna) *zgodność predykcji* z wynikami mniej (strategia dwudziestu pytań) lub bardziej (modele zintegrowane) bogatego i zróżnicowanego repertuaru badań. Cohen, Loftus i inni krytycy testowania istotności hipotezy zerowej kompletnie mijają się z prawdą, kiedy twierdzą, że tym, czego badacz najbardziej chce się dowiedzieć, jest prawdopodobieństwo hipotezy ze względu na dane.

Tym, czego badacz przede wszystkim chce, a w każdym razie powinien chcieć się dowiedzieć, jest odpowiedź na pytanie, jakie jest *najlepsze wyjaśnienie* określonych regularności. Wyjaśnianie nie jest jednak tym samym, co przewidywanie, o czym znowu można się dowiedzieć z dowolnej, względnie aktualnej monografii poświęconej problemom filozofii nauki. Ludzie potrafią przewidywać wiele zjawisk, których nie potrafią wyjaśnić. Korzystając ze współczesnych metod uczenia maszynowego można poszukiwać „nowych praw matematycznych", pozwalających całkiem nieźle przewidywać nawet złożone regularności w sytuacji, gdy sens tych regularności nie jest znany. Inaczej mówiąc, przewidywanie nawet stosunkowo złożonych regularności nie wymaga ich zrozumienia. To, że wiele metod uczenia maszynowego, z których psychologowie poznawczy tak chętnie korzystają, nie wymaga teorii wyjaśniającej modelowane regularności, jest jedną z najważniejszych praktycznych zalet tych metod. Można z powodzeniem posługiwać się modelami matematycznymi, korygować i rozwijać je w oparciu o wyniki kolejnych badań za pomocą nowszych metod wnioskowania, nie dowiadując się przy tym zbyt wiele na temat natury badanych zjawisk. Nie tylko historia koneksjonizmu obfituje w przykłady takich modeli.

Teoria może być dobrym wyjaśnieniem pewnych zjawisk nawet wtedy, gdy nie pozwala ich przewidzieć, albo gdy nie jest w ogóle testowalna eksperymentalnie. Dotyczy to między innymi wielu wyjaśnień ewolucyjnych, które są wyjaśnieniami typu historycznego, wyjaśnień dotyczących zdarzeń niereplikowalnych, albo wyjaśnień zjawisk replikowalnych, ale ze swej natury słabo przewidywalnych, na przykład zachowania się sys-

temów dynamicznych, albo systemów, których działanie w znacznym stopniu zależy od nieznanych warunków początkowych, a przecież zachowanie ludzi jest mocno związane z całą historią ich interakcji ze środowiskiem. Między wyjaśnianiem a przewidywaniem zachodzą ważne związki, ale nie można ich utożsamić.

4.2 Uwagi na temat wyjaśniania

Grobler (2000, 2006) jako podstawowe wyróżnia trzy klasyczne koncepcje oceny wartości poznawczej hipotez, to jest koncepcję probabilistyczną (albo indukcjonistyczną), zgodnie z którą poznawcza wartość hipotezy sprowadza się do jej prawdopodobieństwa określonego ze względu na dostępne świadectwa empiryczne, koncepcję falsyfikacjonistyczną (albo dedukcjonistyczną), zgodnie z którą rewizja oceny polega na odrzucaniu hipotez, których predykcje okazały się niezgodne z obserwacjami i koncepcję abdukcjonistyczną, zgodnie z którą rewizja oceny polega na wnioskowaniu do najlepszego wyjaśnienia. Za wyjątkiem ostatniej, wymienione koncepcje opierają się na założeniu, że wartość poznawczą hipotez empirycznych można w jakiś sposób sprowadzić do kwestii zgodności predykcji z obserwacjami. Zgoda na koncepcję abdukcjonistyczną, którą przyjmuję w tej pracy, wymaga ustalenia, czym właściwie jest wyjaśnienie i na czym polega wyższość pewnych wyjaśnień nad innymi.

Argumentując na rzecz własnej koncepcji wyjaśniania, Grobler zwraca uwagę między innymi na problemy związane z nomologiczno-dedukcyjnym modelem Hempla-Oppenheima (1948; 1966, dalej w skrócie N-D). Zgodnie z tym modelem, wyjaśnieniem jest poprawne logicznie rozumowanie, którego wnioskiem jest to, co stanowi przedmiot wyjaśniania (explanandum), a przesłankami jest pewien zbiór założeń (eksplanans). Przynajmniej jedno założenie należące do ekplanansu musi być prawem, a rozumowanie musi być takie, że pozbawione tego założenia nie byłoby nadal poprawne logicznie. Wszystkie założenia eksplanansa muszą być dodatkowo sprawdzalne empirycznie i prawdziwe. Jeżeli nie jest spełniony warunek prawdziwości założeń należących do eksplanansa, wyjaśnienie określane jest jako potencjalne. Wyjaśnienie byłoby zatem dedukcyjnym wyprowadzeniem predykcji na podstawie ogólnych praw i opisu warunków początkowych. Model nomologiczno-dedukcyjny został później przez Hempla rozwinięty do postaci uwzględniającej statystyczny charakter niektórych praw.

Liczne elementarne wady modelu N-D są już od dawna tak dobrze znane, że nie ma potrzeby abym je wszystkie w tym miejscu wyliczał. Wiele z tych wad ma swoje źródło w tym, że model N-D nie ogranicza w wystarczającym stopniu charakteru zależności między eksplanansem a eksplanandum, co czyni go bezradnym wobec rozmaitych kontrprzykładów. Jeżeli w danych warunkach zmienne A i B są skorelowane, zgodnie z modelem N-D, przy założeniu odpowiednich warunków początkowych, wartości przyjmowane przez zmienną A będą wyjaśniane przez wartości przyjmowane przez zmienną B i vice versa. Poziom inteligencji byłby więc (częściowym, bo statystycznym) wyjaśnie-

niem wzrostu, a wzrost byłby częściowym wyjaśnieniem poziomu inteligencji. Grobler przytacza jeszcze szereg innych koncepcji wyjaśniania, wyliczając znane przykłady ujawniające ich ograniczenia. Opisałem tylko model N-D, dlatego że jest on przywoływany jako mniej więcej trafny w niektórych znanych mi podręcznikach do metodologii badań psychologicznych.

Odpowiedzi na pytanie, czym jest wyjaśnienie, należy zdaniem Van Frassena (1980) szukać, poddając analizie pytanie, na które wyjaśnienie jest odpowiedzią. Wyjaśnienie jest zwykle odpowiedzią na pytanie rozpoczynające się od słowa „dlaczego". Opierając się na tym elementarnym spostrzeżeniu można przeformułować wiele klasycznych koncepcji wyjaśniania - zgodnie z modelem N-D wyjaśnienie to odpowiedź typu „Ponieważ *A* wynika z ...", zgodnie z koncepcją przyczynową będzie to odpowiedź typu „Ponieważ *A* zostało spowodowane przez ...", i tak dalej. Podobnie jak Van Frassen, Grobler twierdzi, że trudności związane z niektórymi koncepcjami wyjaśniania wynikają z błędnego założenia, że wyjaśnianie jest rodzajem rozumowania, w dodatku niezależnego od kontekstu. Wyprowadzenie zgodnie z regułami logiki wniosku o wystąpieniu eksplanandum na podstawie zakładanych praw i opisu warunków początkowych jest rozumowaniem. Udzielenie odpowiedzi na pytanie, dlaczego ktoś jest wysoki, rozumowaniem nie jest, chociaż rozumowanie może zostać przeprowadzone w celu uzasadnienia danego wyjaśnienia, może również odgrywać pewną rolę w procesie poszukiwania dobrej odpowiedzi.

Jak zauważa Grobler, poprawność wyjaśnienia zależy od kontekstu. Pytając o to, dlaczego efekt wielkości zestawu w zadaniu Sternberga jest w przybliżeniu liniowy, autorowi pytania może chodzić o czynniki biologiczne, określony na wyższym poziomie abstrakcji przebieg procesów przetwarzania informacji, cechy samego zadania, zastosowaną metodę analizy statystycznej, albo wiele innych rzeczy. Udzielenie odpowiedzi na wszystkie możliwe wersje tego pytania nie jest możliwe. To, o co chodzi autorowi pytania wynika, jak to ujmuje Grobler, z jego zainteresowań i potrzeb poznawczych.

Żeby odpowiedź w postaci wyjaśnienia była możliwa, pytanie musi mieć określoną, niekoniecznie jawną strukturę. W szczególności, na pytanie „dlaczego efekt wielkości zestawu jest w przybliżeniu liniowy?" nie ma odpowiedzi. Żeby odpowiedź istniała, pytanie musi określać pewien zbiór alternatyw i zawierać odpowiedni akcent zdaniowy. Na przykład, na pytanie „dlaczego efekt wielkości zestawu jest w przybliżeniu *liniowy*, a nie monotonicznie rosnący z przyspieszeniem ujemnym?" odpowiedź można udzielić, co oczywiście nie znaczy jeszcze, że będzie ona poprawna.

Ta „erotetyczna" koncepcja wyjaśniania wydaje się znacznie lepiej pasować do przebiegu rewizji oceny wartości poznawczej hipotez w naukach empirycznych niż koncepcje utożsamiające wyjaśnienie z predykcją. Pozwala uchwycić sens, najwyraźniej uważanych przez wielu naukowców za wartościowe, wyjaśnień teleologicznych (inaczej celowościowych: „ponieważ *A* służy do ..."), przyczynowych, historycznych i innych. Omawiany w rozdziale drugim model dyfuzyjny umożliwia między innymi odpowiedź na pytania „dlaczego osoby przekonane, że bodźce typu *A* występują częściej niż bodźce ty-

pu B reagują *szybciej* na bodźce typu A niż B, a nie tak samo szybko lub wolniej?", albo „dlaczego szybsze wykonywanie reakcji prowadzi do *wzrostu* proporcji reakcji błędnych, a nie do braku takich zmian?", albo „dlaczego rozkład czasów reakcji jest *prawo*, a nie lewoskośny lub symetryczny?", albo „dlaczego w przypadku prostych zadań czasem *nie obserwuje się* korelacji między szybkością reagowania a poziomem inteligencji, nawet gdy próba jest duża, a nie jest tak, że dla prób co najmniej średniej wielkości związek jest prawie zawsze wykrywalny?". Zgodność predykcji modelu dyfuzyjnego z wynikami badań jest czymś innym niż moc wyjaśniająca tego modelu. Sama informacja o akceptowalnej trafności predykcji świadczy jedynie o tym, że model pozwala uchwycić określone obserwowalne regularności. Tak naprawdę badacz chce się dowiedzieć nie tego, czy predykcje jednego modelu są lepsze niż innego, tylko tego, dlaczego obserwowane regularności są takie, a nie inne. Badacz chce przede wszystkim zrozumieć zjawiska.

Przytoczona koncepcja pozwala udzielić częściowej odpowiedzi na pytanie, dlaczego metodologia selektywnego wpływu jest *niezwykle cennym*, a nie tylko umiarkowanie cennym lub bezwartościowym narzędziem oceny wartości poznawczej hipotez w psychologii poznawczej. Gdy udaje się znaleźć czynniki wpływające selektywnie na parametry modelu reprezentującego teorię, może to czasem oznaczać, że wsparcie dla tej teorii jest stosunkowo silne. Siła tego wsparcia nie sprowadza się jednak do samej zgodności wyników z predykcjami modelu. Żeby zaobserwowanie względnie selektywnej modyfikowalności dostarczało ważnej informacji na temat jakości teorii jako wyjaśnienia, wpływ czynników eksperymentalnych musi mieć pewien uchwytny sens. Gdyby okazało się na przykład, że parametry standardowego modelu teorii detekcji dają się selektywnie modyfikować za pomocą czynników, które nie mają zrozumiałego związku z badanym procesem jako procesem podejmowania decyzji w warunkach niepewności, wpływ byłby selektywny, wyniki dałoby się „modelować", ale zaobserwowana selektywność nie znajdowałaby wyjaśnienia w teorii werbalnej, sformalizowanej przez konkretny model. Taki model pozwalałby przewidywać być może nieoczywiste związki między zmiennymi, ale sens zaobserwowanych związków pozostawałby tajemniczy, to jest niewyjaśniony.

W przypadku wielu teorii selektywna modyfikowalność nie świadczy wcale o modularności. Selektywnie modyfikowalne wolne parametry modelu mogą reprezentować funkcjonalnie rozróżnialne własności procesu, a nie odrębne moduły. Nie od rzeczy byłoby jak sądzę uogólnienie pojęcia selektywności wpływu do wpływu sensownego ze względu na teorię. Właśnie ta sensowność, to znaczy to, że takie a nie inne czynniki wpływają selektywnie w taki a nie inny sposób na parametry modelu, na przykład kontrast między bodźcem i tłem wpływa selektywnie na rozróżnialność w modelu teorii detekcji, decyduje o wsparciu dla teorii. Nawet jeżeli nie uda się znaleźć czynników selektywnie oddziałujących na własności pewnego procesu, może uda się znaleźć czynniki wpływające na te własności nieselektywnie, ale w sposób, który trudno inaczej wyjaśnić. W ostateczności, dlaczego na przykład parametry modelu teorii detekcji nie miałyby być skorelowane, tak jak teoretycznie dopuszczalne jest, aby skorelowane były wydajności pewnych przebiegających równolegle lub szeregowo procesów przetwarzania skończonego

zbioru elementów? Dlaczego w takim wypadku należałoby się upierać przy co najmniej bezpośrednio selektywnym wpływie? Tak naprawdę sednem sprawy jest wpływ zrozumiały ze względu na zakładaną teorię.

Za najważniejszy walor metodologii selektywnego wpływu uważam to, że udane jej zastosowanie może dostarczać wsparcia dla określonej psychologicznej interpretacji parametrów i struktury modelu. Żeby teoria reprezentowana przez dany model cokolwiek wyjaśniała, elementy modelu muszą być odpowiednio interpretowalne. Metodologia selektywnego wpływu służyłaby więc przede wszystkim do ustalenia, czy dana teoria nadaje się do formułowania odpowiedzi na pytania, na które powinna odpowiedzi udzielać - selektywna modyfikowalność tempa dryfu przez jasność bodźca w modelu dyfuzyjnym oznacza, że interpretacja odpowiedniego parametru w kategoriach rozróżnialności percepcyjnej jest uprawniona. Wniosek jest uzasadniony pomimo tego, że żadna przestrzeń alternatywnych modeli procesu wyboru z dwóch alternatyw nie była rozważana. Jeżeli wszystkie wolne parametry modelu okażą się odpowiednio, zgodnie z teorią modyfikowalne, najlepszym wyjaśnieniem tego faktu będzie stwierdzenie, że teoria ta jest bliska prawdy a model stanowi jej akceptowalną formalizację.

Porzucenie błędnego stanowiska, zgodnie z którym decydująca dla oceny teorii jest zgodność predykcji z obserwacjami, na rzecz stanowiska, zgodnie z którym podstawowym kryterium oceny teorii jest to, jak dobrze teoria pozwala wyjaśnić te zjawiska, które wyjaśnić powinna, prowadzi do kilku niezupełnie trywialnych konsekwencji. Predykcja nadal jest ważnym, ale już nie najważniejszym kryterium oceny. Żeby teoria była dobrym wyjaśnieniem jakichkolwiek zjawisk, musi sama być dostatecznie zrozumiała, co oznacza między innymi i to, że jej pojęcia muszą być w miarę jednoznacznie i jasno zdefiniowane. Z dwóch modeli takich, że jeden pozwala dosyć dokładnie przewidywać wybraną klasę regularności, ale trudno nadać parametrom tego modelu czytelną interpretację, natomiast drugi, dzięki nadającej mu interpretację teorii, pozwala te regularności dość dobrze zrozumieć, ale gorzej przewidzieć, ten drugi, a nie ten pierwszy będzie bardziej rozwojowy. Zdolność do formułowania nowych, obiecujących hipotez zależy przecież od wstępnego zrozumienia natury badanego zjawiska.

Zaakcentowanie wymienionych pożądanych własności teorii, w połączeniu z obserwacją, że zgodność teorii z wynikami badań jest tylko o tyle pouczająca, o ile związek tych wyników z teorią jest silny, prowadzi do wniosku, że badania teoretyczne są równie ważnym narzędziem postępu w naukach empirycznych co badania empiryczne. Wniosek ten jestem skłonny uznać za niekoniecznie trywialny ponieważ, jak postaram się wykazać w dalszej części pracy, psychologowie poznawczy w praktyce na ogół go nie respektują. Praca, którą Czytelnik ma w rękach zawiera wiele przykładów ilustrujących to stanowisko. Czasami można nawet stwierdzić, czy teoria jest wartościowa, nawet jeżeli nic bliżej nie wiadomo na temat trafności jej predykcji. Będzie tak wtedy, gdy teoria opiera się na przekonujących definicjach podstawowych terminów, należących do aparatu pojęciowego danej dziedziny. Jeżeli ostatecznie okaże się, że przyjęcie tych definicji prowadzi do błędnych predykcji i wadliwych wyjaśnień, rezultatem tej obserwacji będzie wyraźny

postęp. Dowiemy się dzięki temu, że w fundamentalny sposób źle rozumieliśmy badane zjawiska. Ten postęp nie będzie się jednak brał stąd, że *jakaś* teoria okazała się niezgodna z obserwacjami, tylko stąd, że błędna okazała się teoria zarazem wyraźnie wyartykułowana i mocno oparta na podstawowych, względnie narzucających się intuicjach. Żeby tego rodzaju postęp był w psychologii poznawczej możliwy, trzeba najpierw spróbować ustalić, skąd takie przekonujące i wyraźnie wyartykułowane teorie miałyby się brać.

Wbrew temu, co twierdzi Platt, a co sugerują Roberts i Pashler albo cytowany we wprowadzeniu Newell, nie ma jednej, dobrej, uniwersalnej metody naukowej. Eliminatywizm posiada na tyle poważne wady i jest tak rzadko stosowany, że nie sposób go uznać za podstawową metodę rewizji oceny wartości poznawczej teorii. Newell miał rację, kiedy twierdził, że nie można grać z naturą w 20 pytań i wygrać. Nie jest to jednak powód, aby uznać badanie stosunkowo wyizolowanych procesów za złe rozwiązanie, dlatego że w przeważającej większości przypadków nie polega to wcale na graniu w 20 pytań. Rzadkość z jaką eliminatywizm jest faktycznie stosowany nie jest dowodem rozpowszechnionej ignorancji czy lekceważenia dla uniwersalnych standardów metodologicznych, tylko wynika z poważnych ograniczeń tego stanowiska. Rzadkość eksplicitnego i konsekwentnego uwzględnienia alternatywnych wyjaśnień, łamanie lub omijanie zasad stosowanych procedur wnioskowania statystycznego, logiczna i pojęciowa swoboda i inne przypadłości prowadzą do negatywnych konsekwencji, ale zapewniają też elastyczność podejścia, niezbędną w obliczu nietrywialnych problemów. Żeby jednak skutecznie zbliżać się do nieustannie wymykającego się celu, jakim jest odkrywanie coraz lepszych teorii, jak ujął to w innym kontekście Rachlin (2004), trzeba być lekkim jak ptak, ale nie jak piórko. Trzeba unikać zarówno metodologicznej ortodoksji jak i metodologicznego i teoretycznego niedbalstwa.

Psychologowie poznawczy nie są moim zdaniem zasadniczo gorzej przygotowani do radzenia sobie z wyzwaniami własnej dziedziny niż przedstawiciele innych nauk empirycznych, każda taka grupa społeczna ma jednak pewne cechy szczególne, decydujące o mocnych i słabych stronach stosowanego zazwyczaj podejścia. Ze względu na ostateczne wnioski tej pracy, przynajmniej jedna cecha dominującego w psychologii nie tylko poznawczej stylu uprawiania nauki, to jest osobliwy stosunek do badań teoretycznych, zasługuje w tym miejscu na uwagę.

4.3 Badania teoretyczne w psychologii

Rzut oka na typowy podręcznik lub monografię z zakresu psychologii poznawczej wystarcza aby zauważyć, że najwięcej miejsca poświęca się na ogół opisom eksperymentów i ich rezultatów. Jednocześnie liczba alternatywnych, niekoniecznie wykluczających się teorii jest stosunkowo duża, a zakres zjawisk wyjaśnianych przez te teorie mały lub trudny do określenia. Teorie psychologiczne, chociaż liczne, zajmują tak mało miejsca, ponieważ da się je często streścić w kilku zdaniach, od czasu do czasu uciekając się do pomocy

diagramu, w wyjątkowych okolicznościach uzupełniając opis względnie nieskomplikowanym wzorem. Przypuszczam, że uzasadnieniem tego stanu rzeczy ma być złożoność i zagadkowość zjawisk badanych przez psychologię, a także młody wiek samej dziedziny.

Bardziej niepokojący jest poziom rozumowań i jakość stosowanych definicji. Pomijając już liczbę terminów i ich możliwych, luźno określonych znaczeń, co krok można się natknąć na usterki logiczne lub pojęciowe. Klasyfikacje są tylko czasami wyczerpujące i oparte na systematycznym stosowaniu tych samych kryteriów. Pewne wyniki zdają się świadczyć o wyższości niektórych teorii nad pozostałymi, ale trudno powiedzieć jak bardzo, ponieważ związki logiczne między teoriami rzadko są jasne, a jeszcze rzadziej stanowią przedmiot osobnych analiz. Ewidentnie zaniedbanym środkiem poprawy tego stanu rzeczy jest analiza metateoretyczna. Problemy związane z przedmiotem psychologii mają nie tylko charakter empiryczny, ale także, a w wielu wypadkach przede wszystkim, teoretyczny. Głównym źródłem trudności nie są problemy związane z przewidywaniem, tylko ze zrozumieniem przedmiotu psychologii.

Wpisanie w wyszukiwarce Google hasła „theoretical psychology" i odrobina wytrwałości pozwala odnaleźć strony dwóch (sic!) stowarzyszeń[2], których celem statutowym jest promowanie i prowadzenie badań teoretycznych w psychologii, to jest Society for Theoretical and Philosophical Psychology (oddział APA, $www.westga.edu/\ stpp$, dawniej znany jako Division for Theory and Philosophy of Psychology albo Division 24) i International Society for Theoretical Psychology ($psychology.ucalgary.ca/istp$).

ISTP działa od roku 1980-go i w tym czasie stowarzyszenie to zorganizowało 12 konferencji w Europie i Stanach Zjednoczonych. Ze strony domowej można się dowiedzieć, że ISTP jest „międzynarodowym forum dla teoretycznych, metateoretycznych i filozoficznych dyskusji w psychologii, ze szczególnym naciskiem na współczesne psychologiczne debaty".

Ze strony domowej STPP można się z kolei dowiedzieć, że „jest to bardzo eklektyczny oddział, którego członkowie pochodzą z różnych obszarów specjalizacji. Wspólnym przedmiotem zainteresowań tej zróżnicowanej grupy jest filozofia psychologii i nauk społecznych, a także społecznych podstaw psychologii". Oddział ten wydaje jedno czasopismo (Journal of Theoretical and Philosophical Psychology, w skrócie JTPP), którego pierwszy numer ukazał się w roku 1986 i pomijając przypadki wydań łączonych, czasopismo to ukazuje się dwa razy rocznie. Przejrzałem wszystkie abstrakty, we wszystkich opublikowanych dotychczas numerach tego czasopisma.

O ile przytoczone liczby wyglądają skromnie, rozpiętość tematyki poruszanej przez publikujących w JTPP autorów jest imponująca, czego należało się zresztą spodziewać po tak „zróżnicowanej grupie". Artykuły traktują o prawach zwierząt, psychologii „ja", reorganizacji APA (kilka publikacji), problematyce gender, świadomości transcendentalnej, wrodzonej refleksji fenomenologicznej, Heideggerze (m. in. związkach jego teo-

[2]Pomijam tutaj egzotyczne strony poświęcone psychologii teoretycznej prowadzone przez pojedynczych autorów.

rii z wizją świata według Einsteina), Kierkegaardzie, Habermasie, Gadamerze, Levinasie, Sartrze, Arystotelesie (w 1990 roku poświęcono mu cały numer), Lacanie, Humie, Husserlu, obrazie Boga w teorii Piageta, regresji do średniej, jeszcze wiele razy o Heideggerze i Levinasie, Spinozie, taoistycznym podejściu do rozwiązywania konfliktów, i tak dalej. Nie znający profilu tego czasopisma Czytelnik mógłby w pierwszej chwili pomyśleć, że traktuje ono o historii filozofii, przypuszczalnie z akcentem na klasycznych myślicieli. Żeby trafić na nazwisko znanego współczesnego filozofa, podejmującego wprost tematykę istotną dla współczesnej psychologii, trzeba mieć sporo szczęścia (cały numer poświęcony Dennettowi). Dosłownie kilka artykułów dotyczy behawioryzmu, psychologii poznawczej i koneksjonizmu, trochę lepiej (pod względem ilościowym) jest z psychoterapią.

Trudniej ustalić coś na temat problematyki podejmowanej na konferencjach ISTP. W odpowiedzi na wystosowaną do sekretarza stowarzyszenia prośbę o udostępnienie spisu tytułów dotychczasowych wystąpień zostałem poinformowany, że ISTP nie archiwizuje ani tytułów, ani abstraktów, udało mi się jednak uzyskać dostęp do przypuszczalnie reprezentatywnego zbioru wybranych publikacji. Dość powiedzieć, że w przypadku ISTP sytuacja nie wygląda wcale lepiej. Te dwa stowarzyszenia wyczerpują listę widocznych w internecie instytucji, których głównym celem jest promowanie badań teoretycznych w psychologii.

Amerykańskie Stowarzyszenie Psychologiczne wydaje 58 czasopism, z czego zdecydowana większość ma charakter empiryczny. Jedyne wydawane przez Elsevier czasopismo psychologiczne, zawierające stosunkowo dużo tekstów wyraźnie teoretycznych jakie udało mi się znaleźć, to skądinąd znakomity Journal of Mathematical Psychology. Kursy poświęcone psychologii matematycznej należą jednak do rzadkości, a od psychologów (wliczając w to osoby prowadzące kursy poświęcone metodologii badań i statystyce) nie wymaga się zwykle znacznie więcej ponad elementarną umiejętność obsługi jednego lub dwóch popularnych programów służących do analizy statystycznej (J. T. Townsend, Golden i Wallsten, 2005). Nie znam żadnych prestiżowych czasopism o globalnym zasięgu, poświęconych specyficznie problemom teoretycznym psychologii poznawczej, społecznej, poznawczo zorientowanej psychologii społecznej, klinicznej, motywacji, ani jakiejkolwiek innej. Zawartość wspomnianego JTPP ma jeszcze najwięcej wspólnego z nurtami współczesnej filozofii o raczej drugorzędnym znaczeniu, takimi jak egzystencjalizm, postmodernizm, hermeneutyka, fenomenologia, czy filozofia wschodu. Dominująca współcześnie filozofia analityczna reprezentowana jest skromnie.

W moim odczuciu dobrym źródłem analiz teoretycznych o rzeczywistym znaczeniu dla współczesnej praktyki badawczej są artykuły publikowane w takich renomowanych czasopismach, jak Psychological Review, Cognitive Psychology, czy wspomniany już Journal of Mathematical Psychology. Czasopisma te (może za wyjątkiem JMP) zawierają jednak głównie raporty z badań, których autorzy wydają się zajmować w znacznym stopniu pracą o charakterze empirycznym. Można mieć rozmaite opinie na temat wartości badań teoretycznych w psychologii, ale opinie te pozostaną spekulatywne, dopóki jakaś

większa grupa nie pozbawionych kontaktu z rozwojem badań empirycznych osób nie zacznie uprawiać teorii na serio. Stale rosnąca liczba mikroteorii i rozmaitych „terminów technicznych" nie jest oznaką takiego postępowania.

Wbrew pozorom rezultatem badań teoretycznych najczęściej nie jest sformułowanie nowej teorii. Wykrywanie ukrytych założeń, analiza spójności logicznej rozumowań, wyprowadzanie nieoczywistych wniosków, ustalanie związków logicznych między teoriami, poszukiwanie niezgodnych predykcji, interpretacja i eksplikacja pojęć, formułowanie nowych rozróżnień, poszukiwanie uogólnionych postaci twierdzeń, pojęć i teorii, ustalanie mocy wyjaśniającej hipotez, i tym podobne, to wszystko są aktywności, które mogłyby wypełnić cały czas jaki pozostawiają obowiązki dydaktyczne niebagatelnej liczbie psychologów, nawet gdyby dotyczyły tylko *wybranej* dziedziny szczegółowej w psychologii. Gdy idee podobne do falsyfikacjonizmu albo łączenia modeli w większe całości zalecane są jako najbardziej obiecujące, jeżeli nie jedyne, skuteczne narzędzie postępu, nie tylko nie spotyka się to z powszechną krytyką, ale wręcz zdaje się budzić ogólną aprobatę.

Cummins (2000) podjął interesującą w tym kontekście próbę określenia źródeł trudności związanych z oceną wyjaśnień psychologicznych. Poświęciwszy kilka początkowych akapitów umiarkowanie krytycznym uwagom na temat, zdaniem tego autora nadal popularnego wśród psychologów, modelu nomologiczno-dedukcyjnego, omówienie różnic między wyjaśnianiem a przewidywaniem i analizę specyfiki języka opisu wyników badań („procent wariancji wyjaśnionej") Cummins zauważa, że psychologowie zdają się nie cenić wyjaśnień wysoko. Twierdzi między innymi, że bardzo trudno jest opublikować artykuł zawierający samo wyjaśnienie, bez jakichkolwiek nowych wyników badań. Niemal koniecznym warunkiem publikacji ma być „potwierdzenie hipotez". W porównaniu z częścią empiryczną, część poświęcona na dyskusję jest według Cumminsa przeważnie luźna i mocno spekulatywna, zwykle też nie jest w ogóle czytana, a jej zawartość nie jest prawie nigdy cytowana w innych publikacjach. Diagnoza Cumminsa pokrywa się ze spostrzeżeniem innego filozofa nauki. Kukla (1989, 2001) uważa, że psychologowie zajmują się zwykle teorią w czasie gdy nie są akurat zajęci przeprowadzaniem badań i często nie odróżniają problemów o charakterze empirycznym od problemów teoretycznych. Nieuprzedzona lektura wydań Psychological Review z ostatnich dziesięciu lat skłania moim zdaniem do nieco mniej pesymistycznej oceny, niemniej czasopismo to wydaje się pod tym względem wyjątkowe.

Psychologowie są zdaniem Cumminsa przytłoczeni przez to, co ma być wyjaśniane, dysponując zarazem ubogim repertuarem samych wyjaśnień. Trudno powiedzieć, co w psychologii oznacza dobrze wyjaśnić zjawisko, ponieważ:

> Możemy być całkiem pewni, jak by [takie wyjaśnienie] nie wyglądało. Nie wyglądałoby jak *Principia Psychologica*. Zasady mechaniki Newtona zostały sformułowane jako system aksjomatyczny, będąc świadomą próbą naśladowania geometrii Euklidesowej, [a system aksjomatyczny stano-

wił] wyjątkowo wpływowy paradygmat w XVII wieku i od tej pory jest dominującym, paradygmatycznym przykładem wyjaśniającej teorii w nauce. Jest rzeczą dyskusyjną, na ile ten paradygmat jest rzeczywiście użyteczny w jakiejkolwiek dziedzinie nauki[3]. Z pewnością mechanika, nawet mechanika Newtona, nie jest w ten sposób współcześnie prezentowana. Jednakowoż, jeżeli celem jest ustalenie fundamentalnych zasad ruchu, podejście aksjomatyczne ma pewien sens. Istnieje, jak można przypuszczać, mała liczba fundamentalnych zasad określających ruch, a te zasady, w połączeniu z odpowiednimi definicjami, mogłyby pozwolić na wyprowadzenie równań, opisujących (być może wyidealizowane) zachowanie się każdego poszczególnego systemu mechanicznego: wahadła, sprężyny, układu słonecznego, i tak dalej. Tym, co czyni to podejście rozsądnym jest idea, że ruch jest wszędzie taki sam, niezależnie od tego, co się porusza i kiedy. To jest także ten rodzaj idei, który leży u podstaw rozpowszechnionego przekonania, że to fizyka jest najbardziej fundamentalną z nauk.

Odwrotnie, tym co leży u podstaw przekonania, że psychologia albo geologia nie są naukami fundamentalnymi, jest spostrzeżenie, że systemy psychologiczne i geologiczne są specyficzne (ang. *special*). Zasady psychologii, geologii i innych tak zwanych nauk szczegółowych nie dotyczą ogólnie rozumianej natury, ale jedynie systemów szczególnego rodzaju. Prawa psychologii i geologii są prawami *in situ*: to znaczy, prawami dotyczącymi szczególnego rodzaju systemu, ze względu na jego specyficzną konstytucję i organizację. Nauki szczegółowe nie dostarczają ogólnych praw natury, ale raczej praw obowiązujących dla specyficznych systemów, które są dla nich [tych nauk] właściwym przedmiotem. Prawa *in situ* określają efekty - regularne wzorce działania, charakterystyczne dla szczególnego rodzaju mechanizmów.

Kiedy już dostrzeżemy, że prawa nauk szczegółowych są specyfikacjami efektów, widzimy, że teorie w takich naukach nie mogą nawet w najmniejszym stopniu przypominać Principiów Newtona. Kto byłby zainteresowany aksjomatycznym wyprowadzeniem efektów obserwowanych w przypadku wątroby, albo silnika spalinowego?

(s. 121-122, Cummins, 2000)

Pisząc o efektach Cummins odnosi się tutaj do takich regularności jak efekt Stroopa, albo efekt zaoszczędzenia przy ponownym uczeniu się, czyli względnie systematycznie replikujących się regularności, obserwowanych na przykład w warunkach wykonywania pewnych zadań. Autor ten twierdzi, z czym trudno się nie zgodzić, że takie efekty nie są wyjaśnieniami, tylko czymś, co wyjaśnienia wymaga. Jednocześnie, co jest już znacznie

[3]Cumminsowi chodzi tutaj przypuszczalnie o nauki empiryczne, a nie formalne.

mniej oczywiste, to właśnie tego rodzaju zależności miałyby stanowić paradygmatyczny przykład praw w naukach takich jak psychologia.

Przytoczona wyżej argumentacja posiada łatwą do odnalezienia lukę. Trudno w zacytowanym artykule odnaleźć uzasadnienie tezy, że prawa nauk szczegółowych mogą być tylko i wyłącznie specyfikacjami efektów. Cummins po prostu stwierdza, że tak jest, bez jakiejkolwiek dalszej argumentacji. W szczególności, nie podaje żadnego dobrego uzasadnienia, dlaczego teoria „systemów szczególnego rodzaju" nie mogłaby być sformułowana na podobieństwo systemu aksjomatyczno-dedukcyjnego. Dlatego, że ewentualne prawa nie byłyby w wystarczającym stopniu ogólne? Jak można porównać ogólność teorii fizycznych i teorii psychologicznych, skoro mają zasadniczo inny przedmiot? Kto powiedział, że teoria aksjomatyczna musi być równie ogólna jak pewna teoria fizyczna? Jakie jest uzasadnienie stwierdzenia, że w przeciwieństwie do „systemu fizycznego", „system psychologiczny" ma pewne szczególne właściwości? Jak jest w ogóle możliwa empiryczna teoria czegoś, co szczególnych właściwości nie posiada? Luka w argumentacji bierze się zapewne z przekonania o nieodpartej oczywistości wniosków.

Być może Cummins ma rację kwestionując użyteczność aksjomatyczno-dedukcyjnej postaci teorii *jako wyjaśnienia*. Czym innym jest jednak wyjaśnienie zjawiska za pomocą rozumowania dedukcyjnego, a czym innym strategia badawcza, polegająca na próbach tworzenia teorii poprzez dedukcyjne wyprowadzanie wniosków z możliwie uzasadnionych lub narzucających się, ogólnych przesłanek. Cummins zdaje się sugerować, że po pierwsze, nawet w przypadku teorii sformułowanych w postaci aksjomatyczno-dedukcyjnej, zastosowanie tych teorii do wyjaśniania zjawisk nie polega na dedukowaniu eksplanandum z aksjomatów teorii, a po drugie, że z powodu szczególnego charakteru swojego przedmiotu, teorie w naukach takich jak geologia czy psychologia nie mogą być sformułowane w postaci aksjomatyczno-dedukcyjnej. Pierwsza sugestia zgadza się z tym, co na temat wyjaśniania piszą Van Frassen i Grobler. Dopóki nie będzie uzasadniona, druga sugestia pozostanie dyskusyjna.

4.4 Podsumowanie

Ocena wartości strategii integracyjnej albo strategii teorii lokalnych nie może być oparta na wnioskach dotyczących wartości strategii 20 pytań, ponieważ stosowanie strategii 20 pytań nie jest cechą charakterystyczną podejścia opartego na badaniu względnie wyizolowanych procesów, struktur czy funkcji. Ocena obu strategii nie może być również oparta na mocy predykcyjnej modeli zintegrowanych lub modeli lokalnych nie tylko dlatego, że moc predykcyjna modeli zintegrowanych jest trudna do określenia, ale przede wszystkim dlatego, że najważniejszym kryterium oceny wartości poznawczej hipotez w naukach empirycznych nie jest ich moc predykcyjna, tylko wyjaśniająca.

Należy więc zadać pytanie, w jaki sposób modele zintegrowane różnią się od modeli lokalnych pod względem jakości dostarczanych wyjaśnień i zakresu wyjaśnianych

zjawisk. Ściślej, pytanie, które należy zadać, brzmi zupełnie inaczej - problem dotyczy wyjaśniającej mocy teorii, a nie reprezentujących je modeli. O ile mi wiadomo nie zaproponowano do tej pory żadnych rozsądnych rozwiązań problemu ilościowej oceny mocy wyjaśniającej, ale nic nie stoi na przeszkodzie, aby spróbować ustalić, na jakie klasy pytań teorie określonego rodzaju z samej swojej natury mogą albo nie mogą dostarczyć odpowiedź. Zanim na dobre podejmę ten wątek, w następnym rozdziale opiszę przykłady całkiem udanych teorii, które przeczą tezie Cumminsa. Będą to teorie poznawcze, które można, oddając sprawiedliwość ich szczególnemu charakterowi, przedstawić w postaci aksjomatyczno-dedukcyjnej.

Rozdział 5

Trzy przykłady teorii racjonalnych

5.1 Sheparda teoria uniwersalnej generalizacji

W opublikowanym w roku 1987 artykule pod tytułem „Toward a Universal Law of Generalization for Psychological Science" Shepard zaproponował pewną teorię generalizacji, rozumianej jako funkcjonalna składowa uczenia się. Teoria Sheparda została niedawno przeformułowana i znacznie rozwinięta przez Griffithsa i Tenenbauma (2001), a także z innej perspektywy przez Chatera i Vitányiego (2003). Zanim w dalszej części tekstu odniosę się do rozwiniętej postaci teorii, chciałbym zwrócić uwagę na szczególny sposób, w jaki problem generalizacji został przez Sheparda ujęty.

Sądzę, że przede wszystkim analiza tego sposobu, a dopiero w dalszej kolejności samych rezultatów, dostarcza ważnych wskazówek na temat potencjalnych źródeł niezadowalającego tempa rozwoju psychologii poznawczej. Co więcej, ponieważ propozycja Sheparda jest konstruktywna, analiza tego przypadku pozwala jednocześnie ustalić coś nie tylko na temat tego, dlaczego tempo rozwoju nie jest tak imponujące, jak można by sobie tego życzyć, ale także na temat tego, co być może należałoby zrobić, aby stan rzeczy uległ poprawie. W pewnym sensie, klucz do częściowego rozwiązania zagadki efektywności procesu formułowania i oceny propozycji teoretycznych w psychologii poznawczej znajduje wyraz w następującym fragmencie:

> Czasem potrzebna jest nie większa ilość danych lub dane bardziej szczegółowe, tylko inne sformułowanie problemu. Newton odkrył uniwersalne prawa ruchu tylko dzięki odejściu od stanowiska Arystotelesa i Ptolemeusza, zgodnie z którym ziemia była traktowana jako konkretnie dany punkt odniesienia i zamiast tego wybrał abstrakcyjnie scharakteryzowaną przestrzeń, względem której wszystkie obiekty, włączając w to ziemię, poruszają się zgodnie z tymi samymi prawami.
>
> Analogicznie w psychologii prawo, które byłoby niezmienne niezależnie od wymiarów percepcyjnych, modalności, różnic indywidualnych i ga-

> tunków jest być może osiągalne tylko dzięki sformułowaniu tego prawa względem odpowiedniej, abstrakcyjnej przestrzeni psychologicznej.
>
> (s. 1318, Shepard, 1987)

Jak się wkrótce przekonamy, nawiązania do teorii Newtona nie są tu przypadkowe.

Tradycyjnie problem generalizacji przedstawiany jest zwykle z perspektywy wyników uzyskiwanych w badaniach nad procesami warunkowania. Między innymi w przypadku warunkowania klasycznego, u wielu gatunków i w wielu warunkach eksperymentalnych można zaobserwować, że wyuczony sposób reagowania na bodźce pojawia się nie tylko wobec bodźców zastosowanych na etapie uczenia, ale także wobec innych bodźców, pod różnymi względami do nich podobnych (Ghirlanda i Enquist, 2003). Na ogół obserwuje się wtedy, że intensywność reagowania, mierzona jako prawdopodobieństwo, częstotliwość, siła bądź szybkość reakcji, zmniejsza się w miarę malejącego podobieństwa fizycznego między bodźcami nowymi a tymi zastosowanymi na etapie warunkowania. Zależność między siłą reagowania a fizyczną różnicą między bodźcami nosi nazwę „gradientu generalizacji" (Guttman i Kalish, 1956).

Tak jak w przypadku każdej innej obserwowanej regularności, można zadać pytanie, do jakiego stopnia regularność ta zasługuje aby nazwać ją prawem. W tym kontekście termin „prawo" jest oczywiście wieloznaczny i nie ma jednego ustalonego zbioru warunków koniecznych i wystarczających, które powinny być spełnione, aby określenie dało się zastosować. W przypadku gradientu generalizacji, przynajmniej kilku badaczy (Lashley i Wade, 1946; Bush i Mosteller, 1951) argumentowało, że zależność ta nie może być prawem, dlatego że nie jest wystarczająco niezmienna ze względu na warunki eksperymentalne i badany organizm. Zależnie od gatunku organizmu i zastosowanej fizycznej miary podobieństwa gradient może nie być nawet monotoniczny.

Zdaniem Sheparda, szansa na ustalenie bardziej uniwersalnej regularności mogłaby wzrosnąć, gdyby dotyczyła ona zmiennych zdefiniowanych ze względu na abstrakcyjną przestrzeń psychologiczną, a nie zmiennych zdefiniowanych fizycznie. Odnosząc się do możliwych sposobów rozumienia naukowego jako ilościowego charakteru teorii psychologicznej, autor ten zwraca uwagę, że „fizyczny charakter pomiaru niekoniecznie gwarantuje niezmienność [obserwowanej zależności]" (s. 1317, tamże). Propozycja Sheparda ma między innymi znosić pewną trudność, związaną z jednym ze sposobów wyprowadzenia psychologicznej definicji podobieństwa:

> Zamiast tego, gdybyśmy poszukiwali psychologicznej [a nie fizycznej] miary różnicy jako zmiennej niezależnej, najbardziej podstawową taką miarą byłyby z pewnością same wyniki pomiaru generalizacji, ewidentnie sprawiając, że próby ustalenia funkcjonalnego prawa byłyby obarczone błędnym kołem.
>
> (s. 1318, tamże)

Tak jak prawa ruchu Newtona mają zawdzięczać swoją uniwersalność przez to, że są zdefiniowane na wystarczająco abstrakcyjnym poziomie, tak samo nie można wykluczyć, że generalizacja mogłaby stać się uniwersalnym prawem ze względu na bardziej abstrakcyjnie zdefiniowaną przestrzeń psychologiczną. Rozwiązanie zaproponowane przez Sheparda polega na ustaleniu funkcji psychofizycznej, przekształcającej wartości określone na wymiarach fizycznych w wartości na wymiarach zdaniem Sheparda psychologicznych. Dopiero w takiej psychologicznej przestrzeni należy poszukiwać uniwersalnego prawa. Shepard ustalił, że w wielu wypadkach można unikalnie zidentyfikować taką niezależną od fizycznych wymiarów funkcję, która pozwala na przekształcenie danych w wartości liczbowe, interpretowalne jako odległości w pewnej ciągłej przestrzeni metrycznej. Ponieważ odległości w przestrzeni metrycznej muszą spełniać określone warunki, tym silniejsze im mniej wymiarów ma ta przestrzeń, tak wyznaczone podobieństwo psychologiczne nie jest po prostu opisem danych w innym języku. Zarzut błędnego koła zostaje tym samym oddalony. Współcześnie rozwiązanie Sheparda znane jest jako skalowanie wielowymiarowe.

Zastosowana w badaniach ludzi i zwierząt, zarówno z bodźcami wzrokowymi jak i słuchowymi, metoda ta pozwoliła odkryć, że zależność generalizacji (zoperacjonalizowanej jako prawdopodobieństwo tej samej reakcji dla każdej pary bodźców) od podobieństwa psychologicznego przebiegała w każdym przypadku monotonicznie, w sposób przypominający funkcję wykładniczą. Uzyskanie zadowalającego dopasowania funkcji wykładniczej do wyników wymagało jedynie ustalenia odpowiedniej wartości parametru nachylenia. Jak wykazał Shepard, metoda skalowania wielowymiarowego nie wymusza takiej postaci zależności, chociaż wymusza monotoniczność. Nie ma potrzeby, abym w tym miejscu omawiał szczegółowo metodę skalowania wielowymiarowego. Na razie ważne jest tylko, że odkryta przez Sheparda, względnie prosta i uniwersalna zależność nie była tylko skutkiem ubocznym zastosowanej przez niego metody analizy.

Przypuszczalnie już po tym, jak udało mu się ustalić, że obserwowany gradient generalizacji przebiega zgodnie z wykładniczą funkcją podobieństwa psychologicznego, Shepard przedstawił dosyć specyficzne teoretyczne uzasadnienie dla tej obserwacji. Zauważył, że behawioralni teoretycy uczenia się zdają się błędnie przypuszczać, że prawa opisujące warunkowanie należy traktować jako pierwotne, a problem generalizacji, jako drugorzędny, można rozwiązać w dalszej kolejności. Wobec takiego stanowiska Shepard wysuwa argument:

> Ponieważ jest mało prawdopodobne, aby jakikolwiek obiekt lub sytuacja doświadczana przez organizm pojawiły się ponownie w dokładniej tej samej postaci i kontekście, sugeruję, że pierwszym ogólnym prawem psychologii powinno być prawo generalizacji. (...) Pełna charakterystyka zmiany wywołanej w organizmie, nawet przez pojedyncze zdarzenie w środowisku, musi pociągać za sobą specyfikację tego, w jaki sposób potencjał zachowaniowy uległ zmianie względem jakiejkolwiek następującej później sytu-

acji.

(s. 1317, tamże)

Założenie o zdolności do generalizowania jest konieczne do wyjaśnienia tego, jak w ogóle możliwe jest uczenie się, a więc w kontekście uczenia się nieuchronnie pojawia się pytanie o to, jak faktycznie zachodzi generalizacja u ludzi i zwierząt. To, że nie można zrozumieć generalizacji bez zrozumienia uczenia się i vice versa wynika więc ze znaczenia tych pojęć. Shepardowi udało się dedukcyjnie wyprowadzić wniosek o w przybliżeniu wykładniczym kształcie psychologicznego gradientu generalizacji, rozważając zagadnienie optymalnego wnioskowania na temat własności przyszłego bodźca na podstawie kontaktu z pojedynczym bodźcem wcześniejszym. „Prawo generalizacji" zostało więc wyprowadzone na podstawie rozumowania dotyczącego skrajnie nierealistycznej i uproszczonej sytuacji, podobnie jak prawa ruchu Newtona.

5.1.1 Rozwiązanie problemu generalizacji w ujęciu Sheparda

Wyobraźmy sobie, że w eksperymencie dotyczącym warunkowania klasycznego, w pierwszej z serii prób prezentowany jest bodziec dźwiękowy, na przykład ton o określonej wysokości. Zaraz potem prezentowany jest inny bodziec, o którym wiadomo, że zwiększa prawdopodobieństwo określonej reakcji bezwarunkowej. Gdyby bodźcem bezwarunkowym był pokarm, spodziewalibyśmy się między innymi wzrostu aktywności gruczołów wydzielających ślinę. Przebieg tej próby można opisać z perspektywy rozumującego racjonalnie organizmu. Jeżeli po jednokrotnej prezentacji tonu pojawił się pokarm, jakie jest prawdopodobieństwo, że po kolejnej prezentacji tonu również pojawi się pokarm? Albo jakie jest prawdopodobieństwo, że pokarm pojawi się również po prezentacji tonu o innej wysokości lub natężeniu?

Trudną do przecenienia zasługą Sheparda było zredukowanie problemu generalizacji do być może najprostszej możliwej postaci. Załóżmy, że organizm odbiera sygnały ze środowiska jako wartości na pewnej skali. Dla uproszczenia, niech ta hipotetyczna skala przyjmuje wartości od 1 do 10. Jeżeli pierwszy napotkany bodziec o wartości 6 miał pewną własność, na przykład okazał się smaczny, jakie jest prawdopodobieństwo, że kolejny bodziec o tej samej wartości będzie również smaczny? Wreszcie, jakie jest prawdopodobieństwo, że bodziec o innej wartości, na przykład 8, będzie smaczny?

Udzielenie odpowiedzi na te pytania nie jest oczywiście możliwe bez dodatkowych założeń na temat związku między hipotetyczną własnością a wartościami bodźca. Mówiąc w uproszczeniu, Shepard przyjął, że dopuszczalne są tylko takie własności, którym odpowiada interwał na jakiejś ciągłej psychologicznej skali, ale już nie dowolny zbiór punktów. Zamiast więc pytać o to, jakie jest prawdopodobieństwo, że następny bodziec y będzie miał tą samą własność co bodziec x (również będzie smaczny), można równoważnie zapytać, jakie jest prawdopodobieństwo, że y należy do tego samego, odpowiadającego własności bycia smacznym, nieznanego interwału. Zakładamy przy tym, że

organizm wie, że jeden z interwałów jest właściwy, to znaczy odpowiada dokładnie tym wartościom, dla których bodźcowi przysługuje dana własność, ale zupełnie nie wie który, a więc wszystkie możliwe interwały są a priori jednakowo prawdopodobne. Czytelnik zdążył się już zapewne domyślić, że zagadnienie to można przedstawić jako szczególny przypadek wnioskowania bayesowskiego. Na tym właśnie polega propozycja Griffithsa i Tenenbauma (2001).

5.2 Griffithsa i Tenenbauma uogólniona teoria generalizacji

Zamiast mówić o regionach czy interwałach, można rozważyć zbiór alternatywnych i wzajemnie wykluczających się hipotez H. Każda hipoteza $h \in H$ będzie odpowiadała unikalnie jednemu interwałowi $I(h)$, czyli $x \in I(h)$ będzie oznaczało, że zgodnie z hipotezą h własność przysługuje bodźcowi x. Dla uproszczenia załóżmy również na razie, że zarówno bodźce jak i granice interwałów mogą przyjmować wyłącznie wartości całkowite, a więc każda hipoteza odpowiada pewnemu zbiorowi następujących po sobie liczb całkowitych, a w przypadku najmniejszych interwałów, pojedynczej liczbie. Jeżeli jedyne co można powiedzieć, to że zgodnie z hipotezą własność albo przysługuje egzemplarzowi x, albo nie, każda hipoteza będzie odpowiadała następującemu rozkładowi prawdopodobieństwa:

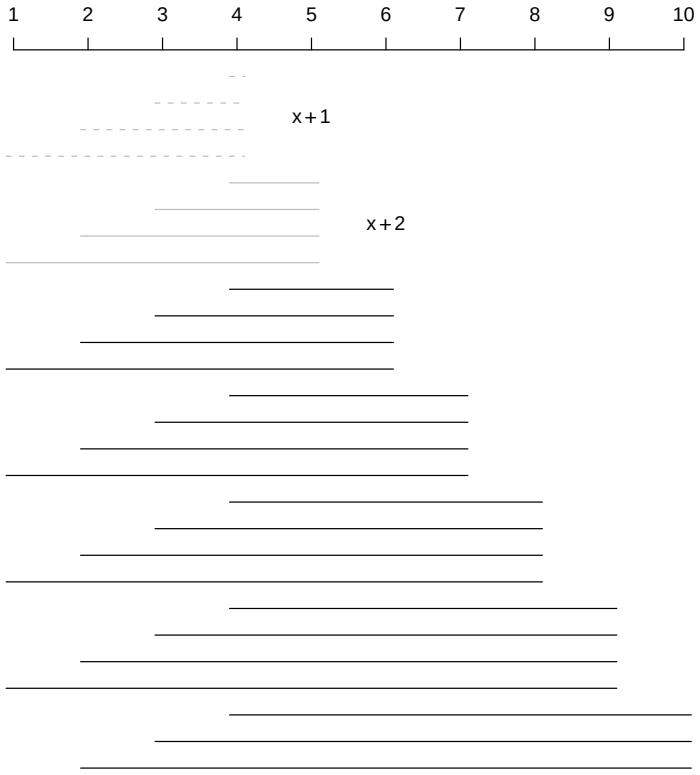
$$p(x|h) = 1 \text{ jeżeli } x \in I(h)$$
$$p(x|h) = 0 \text{ jeżeli } x \notin I(h)$$

Gdy dodatkowo określony jest rozkład aprioryczny na zbiorze hipotez, dla każdego elementu zbioru H możemy obliczyć aposterioryczne prawdopodobieństwo $p(h|x)$, które odpowiada uaktualnionej sile subiektywnego przekonania, że własność przysługuje dokładnie wszystkim tym egzemplarzom, które należą do $I(h)$. Zgodnie z założeniem o początkowej ignorancji organizmu, aprioryczne prawdopodobieństwo każdej hipotezy jest jednakowe i równe $p(h) = 1/|H|$. Wtedy aposterioryczne prawdopodobieństwo każdej hipotezy dane jest przez:

$$\begin{aligned} p(h|x) &= \frac{p(x|h)p(h)}{p(x)} \\ &= \frac{p(x|h)/|H|}{1/|H| \sum_{h' \in H} p(x|h')} \\ &= \frac{p(x|h)}{|h' : x \in I(h')|/|H|} \end{aligned}$$

Dla każdej hipotezy h niezgodnej z egzemplarzem $p(x|h) = 0$ implikuje $p(h|x) = 0$, w trakcie uczenia się hipotezy niezgodne z doświadczeniem są zatem odrzucane bezpowrotnie. Dla każdej hipotezy zgodnej $p(h|x)$ musi być jednakowe i równe $1/|h : x \in I(h)|$,

ponieważ priory są jednakowe. Przykładowy zbiór hipotez o niezerowym prawdopodobieństwie aposteriorycznym po napotkaniu jednego egzemplarza przedstawiłem graficznie na poniższym wykresie:



Rysunek 5.1: Zbiór hipotez o niezerowym prawdopodobieństwie aposteriorycznym zgodnych z wartościami $x + 1$ lub $x + 2$ po zaobserwowaniu egzemplarza $x = 4$

Linią przerywaną zaznaczyłem hipotezy odpadające po zaobserwowaniu następnego egzemplarza o wartości $x + 1$, a linią jasnoszarą hipotezy odpadające po zaobserwowaniu egzemplarza $x + 2$.

Dla każdej możliwej wartości następnego bodźca y określony jest teraz pewien zbiór hipotez o jednakowym niezerowym prawdopodobieństwie aposteriorycznym takich, że wartość y jest zgodna z tymi hipotezami. Im liczniejszy jest ten zbiór, tym więcej hipotez, które nie zostały dotąd odrzucone, wskazuje na to, że własność będzie przysługiwała również y-owi. Jak łatwo zauważyć, hipotez takich będzie tym mniej, im bardziej y jest oddalone od x.

Naturalną miarą siły przekonania, że własność przysługuje także kolejnemu egzemplarzowi y, jest teraz stosunek liczby interwałów zawierających jednocześnie y i x do licz-

by wszystkich interwałów zawierających x. Z wartością $x+1$ niezgodne są wszystkie nieodrzucone dotąd interwały zawierające x postaci $[n, x], 1 \leqslant n \leqslant x$, czyli niezgodnych jest x interwałów. Dla $x + 2$ dodatkowo niezgodne są interwały postaci $[n, x + 1], 1 \leqslant n \leqslant x$ (interwał $[x + 1, x + 1]$ został odrzucony po zaobserwowaniu x), których też jest x, a ogólnie, dla $x + n, x < n \leqslant 10$ niezgodnych jest nx interwałów zgodnych z x. Wynikający stąd gradient generalizacji jest liniowy, a nie wykładniczy.

Jak udało się ustalić Griffithsowi i Tenenbaumowi, gradient staje się w przybliżeniu wykładniczy[1] przy założeniu, że $p(x|h) = 1/|I(h)|$, czyli egzemplarze traktowane są jak próby pochodzące z nieznanego dyskretnego rozkładu jednostajnego. Posterior dla każdej hipotezy będzie wtedy odwrotnie proporcjonalny do wielkości jej interwału, wobec czego promowane będą hipotezy bardziej specyficzne. Żeby ustalić prawdopodobieństwo dla każdej możliwej przyszłej wartości nie można już teraz po prostu liczyć nieodrzuconych wcześniej hipotez zgodnych z daną wartością, posteriory hipotez zależą bowiem od wielkości ich interwałów. Rozkład predykcyjny, który można było wcześniej ustalić przez zliczanie hipotez nieodrzuconych, teraz można obliczyć sumując prawdopodobieństwa przyszłej wartości y po hipotezach, ważonych przez posteriory:

$$p(y|x) = \sum_{h \in H} p(y|h)p(h|x)$$

Tak określony rozkład predykcyjny wykorzystuje całą dotychczasową informację ze względu na rozważany zbiór hipotez. Nie jest to nic innego, jak uwzględniająca złożoność miara predykcji wynikających z modelu, tyle że rozkłady należące do modelu są tutaj indeksowane przez interwały. Wartość $p(y|x)$ zależy zarówno od tego, ile hipotez zgodnych z x jest również zgodnych z y, jak i od tego, jak bardzo te hipotezy są specyficzne. W przybliżeniu wykładniczy gradient generalizacji wynika stąd, że w miarę jak rośnie różnica między x i y, liniowo maleje proporcja hipotez zgodnych z x i y (co już ustaliliśmy), ale hipotezy bardziej specyficzne odpadają wcześniej, wobec czego $p(y|x)$ maleje coraz łagodniej. Shepardowi udało się wykazać, że w przybliżeniu wykładnicza postać gradientu jest zadziwiająco niewrażliwa na $p(h|x)$, o ile posteriory są niezależne od położenia interwału.

Jak dotąd udało się jedynie przedstawić pewne uzasadnienie teoretyczne w przybliżeniu wykładniczej postaci gradientu generalizacji dla wybitnie wyidealizowanego przypadku wnioskowania na temat pojedynczego przyszłego bodźca na podstawie kontaktu z pojedynczym bodźcem wcześniejszym, przy założeniu nieinformacyjnych priorów. Nie

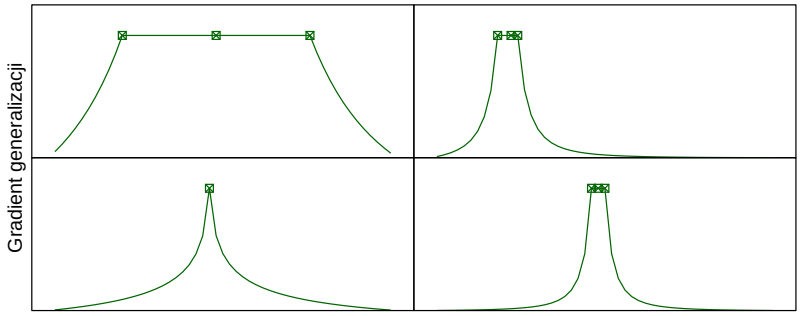
[1]Przybliżenie staje się jeszcze dokładniejsze, gdy hipotezy są ciągłe, a nieinformacyjny rozkład apriorycznyokreślony na zbiorze hipotez jest rozkładem Erlanga, takim samym jaki zastosował Shepard. Niestety, rozkład Erlanga został przyjęty arbitralnie i Griffiths nie potrafi podać żadnego teoretycznego uzasadnienia dla tego wyboru (Griffiths 2009, korespondencja osobista). Teoria generalizacji Griffithsa i Tenenbauma ma jednak szereg innych zalet, daleko ważniejszych niż dokładna zgodność przewidywanego gradientu z gradientem zwykle obserwowanym, dlatego problem ten nie będzie miał dla mnie większego znaczenia.

jest to jeszcze wynik szczególnie imponujący. Gdyby teoria sprowadzała się tylko do tego, nie byłoby o czym wspominać, konsekwencje reinterpretacji problemu w kategoriach wnioskowania bayesowskiego są jednak poważniejsze.

Reinterpretacja ta pozwala między innymi bez trudu uwzględnić dowolną ustaloną sekwencję napotykanych kolejno egzemplarzy. Przy założeniu, że egzemplarze są niezależnymi próbkami z tego samego rozkładu, posterior h ze względu na wektor egzemplarzy x dany jest przez:

$$\begin{aligned} p(h|x) =& \frac{p(x|h)p(h)}{p(x)} \\ =& \frac{\left[\prod_{i=1}^{n} p(x_i|h)\right] p(h)}{\sum_{h' \in H} (\prod_{i=1}^{n} p(x_i|h'))p(h')} \end{aligned}$$

Można teraz zobrazować kilka konsekwencji tej teorii:



Rysunek 5.2: Ciągłe gradienty generalizacji dla kilku przykładowych zbiorów egzemplarzy

Na zamieszczonym wyżej wykresie widać jak zmienia się gradient generalizacji w zależności od obserwowanych egzemplarzy. Zgodnie z oczekiwaniami, gradient opada tym

szybciej, im więcej jest bodźców skupionych w wąskim obszarze. Gdy bodźców jest dużo, a ich zmienność mała, niepewność co do prawdziwego interwału też powinna być mała. Im większa jest zmienność między bodźcami, tym łagodniejszy jest gradient. Na górnych wykresach widać też wyraźnie, że w zależności od tego, gdzie skupione są bodźce, gradient może być mniej lub bardziej asymetryczny.

Do tej pory hipotezy odpowiadały interwałom określonym na pewnej skali, jednak uogólniona wersja teorii nie wymusza takiego ograniczenia. Nic nie stoi na przeszkodzie, aby hipotezy reprezentowały interwały na dwóch lub większej liczbie wymiarów. Nie ma powodu, aby upierać się przy interwałach jako takich, czy nawet ograniczać się do przestrzeni metrycznych. Gdyby obserwowane egzemplarze były faktycznie liczbami całkowitymi, w zależności od kontekstu rozsądny zbiór hipotez mógłby wyglądać zupełnie inaczej, na przykład, w warunkach szkolnych mógłby zawierać hipotezy odpowiadające liczbom parzystym, nieparzystym, wielokrotnościom dziesiątek, i tak dalej. Zasadę rozmiaru, polegającą na premiowaniu hipotez bardziej specyficznych, można wtedy stosować w ten sam sposób, a rolę wielkości interwału będzie odgrywała wielkość zbioru. Jeżeli znajdą się po temu powody, można też zastosować priory informacyjne wyrażające początkowe przekonania osób badanych. Teoria Griffithsa i Tenenbauma daje się naturalnie, w sposób uzasadniony uogólniać na znacznie większy zakres przypadków niż teoria Sheparda.

Jak zauważyli autorzy, tradycyjnie traktowane jako wyjaśnienia konkurencyjne, oparta na pojęciu ciągłej przestrzeni metrycznej teoria Sheparda i oparty na założeniu dyskretności cech model subiektywnego podobieństwa Tversky'ego (1977) okazują się być szczególnymi przypadkami teorii generalizacji. Zgodnie z modelem Tversky'ego, subiektywne podobieństwo $S(x, y)$ między obiektami x i y ma być funkcją zbiorów cech wspólnych i dystynktywnych:

$$S(y, x) = \theta f(Y \cap X) - \alpha f(Y - X) - \beta f(X - Y) \tag{5.1}$$

gdzie X i Y to odpowiednio zbiory cech obiektów x i y, α, β i θ to wolne parametry, a f to pewna bliżej niesprecyzowana funkcja określona na zbiorach cech. Zgodnie z modelem Tversky'ego, subiektywne podobieństwo zależy od tego, ile cech wspólnych posiadają oba obiekty, a także od tego, ile mają cech specyficznych. Dzięki temu, że cechy dystynktywne każdego obiektu uwzględniane są oddzielnie z potencjalnie różnymi wagami, model można uzgodnić z często obserwowanym, niesymetrycznym charakterem subiektywnego podobieństwa. Model Tversky'ego przedstawiany jest czasem w takiej oto, alternatywnej postaci:

$$S(y, x) = 1/\left[1 + \frac{\alpha f(Y - X) + \beta f(X - Y)}{\theta f(Y \cap X)}\right] \tag{5.2}$$

Żeby zademonstrować, w jaki sposób ten model wynika z ogólnej teorii generalizacji, wystarczy zamienić interwały na poszczególne cechy. Wtedy $p(x|h) = 1$, gdy $Q(h) \in$

$Q(x)$ i $p(x|h) = 0$ gdy $Q(h) \notin Q(x)$, gdzie $Q(x)$ oznacza zbiór cech obiektu x, a $Q(h)$ pojedynczą cechę, odpowiadającą hipotezie h. Przy założeniu jednakowych prawdopodobieństw apriorycznych dla wszystkich hipotez/cech $p(y|x)$ będzie stosunkiem liczby cech wspólnych egzemplarzom x i y do liczby wszystkich cech egzemplarza x, a więc:

$$\begin{aligned} p(y|x) =& \frac{|Q(x) \cap Q(y)|}{|Q(x)|} \\ =& \frac{|Q(x) \cap Q(y)|}{(|Q(x) \cap Q(y)) \cup [Q(x) - Q(y)]|} \\ =& \frac{|Q(x) \cap Q(y)|}{|Q(x) \cap Q(y)| + |[Q(x) - Q(y)]|} \\ =& 1/\left[1 + \frac{|Q(x) - Q(y)|}{|Q(x) \cap Q(y)|}\right] \end{aligned}$$

co staje się równe modelowi Tversky'ego (5.2) i monotonicznie związane z (5.1), gdy tylko rozluźnić założenia co do priorów i przyjąć, że funkcja f jest addytywna, $\alpha = 0$ i $\beta = 1$ (Tenenbaum i Griffiths, 2001).

Subiektywne podobieństwo znajduje teraz wyjaśnienie jako generalizacja w obie strony, to jest $p(y|x)$ i $p(x|y)$. Potencjalna asymetryczność relacji podobieństwa, przyjmowana w modelu Tversky'ego jako aksjomat, wynika naturalnie z ogólnej teorii generalizacji. Im więcej cech dystynktywnych względem y ma x, tym mniejsza będzie wartość $p(y|x)$ i odpowiednio dla cech dystynktywnych y względem x i $p(x|y)$. Nic w modelu Tversky'ego nie pozwala stwierdzić, od czego miałoby zależeć występowanie lub stopień asymetryczności subiektywnego podobieństwa. Model jedynie dopuszcza taką możliwość, dzięki zastosowaniu wolnych parametrów i bliżej nieokreślonej funkcji na zbiorach cech. W porównaniu z ogólną teorią generalizacji, propozycja Tversky'ego jest tylko zwięzłym, formalnym zapisem znanej obserwacji, że subiektywne podobieństwo jest bliżej nieokreśloną, potencjalnie asymetryczną funkcją cech wspólnych i dystynktywnych. Sam model Tversky'ego nie pozwala zrozumieć, dlaczego tak się dzieje, ponieważ nie wnosi nic nowego do naszego rozumienia natury oceny podobieństwa i generalizacji.

Wszystkie jakościowe własności gradientu generalizacji wynikają z ogólnej teorii bez użycia wolnych parametrów i bez odwoływania się do mechanizmu obliczeniowego. Ustalenie, że teorie uznawane wcześniej za konkurencyjne są w istocie szczególnymi przypadkami teorii ogólniejszej jest przykładem znacznego postępu, nie wymagającego przeprowadzenia ani jednego eksperymentu. W połączeniu z intrygującym wyjaśnieniem nieoczywistego związku między generalizacją a trudniejszym do zdefiniowania podobieństwem (Goodman, 1972), dopóki nie zostaną przedstawione argumenty świadczące o konieczności rozróżnienia tych pojęć, wiedza na temat pewnej klasy zjawisk ulega znacznemu uproszczeniu. Uogólniona teoria generalizacji stanowi teraz najlepsze wyjaśnienie

obszernej klasy zjawisk, do której należą między innymi zjawiska wyjaśniane przez teorie Sheparda i Tversky'ego.

Nie wiadomo, czy z teorii Griffithsa i Tenenbauma wynika jakieś „uniwersalne prawo generalizacji". Wręcz przeciwnie, wydaje się, że generalizacja powinna być w znacznym stopniu zależna od kontekstu, na który składa się między innymi model jakim dysponuje organizm i historia interakcji tego organizmu ze środowiskiem. Jeżeli można tu mówić o uniwersalnych prawach, to są to prawa wnioskowania bayesowskiego, a nie prawa opisujące regularności w obserwowalnym zachowaniu. Wygląda na to, że Shepard przecenił uniwersalność wykładniczego prawa generalizacji, a nie docenił innego waloru swojego podejścia, polegającego na próbie dedukcyjnego wyprowadzenia teorii w oparciu o pewne ogólne założenia, dotyczące samej możliwości uczenia się w warunkach niepewności. Raczej te założenia, a nie wykładnicze prawo generalizacji, wypada moim zdaniem uznać za uniwersalne.

Teoria Sheparda i jej uogólniona wersja interesuje mnie w tej pracy jako przykład ilustrujący pewne wnioski metateoretyczne. Omówię jeszcze jedną teorię opartą na takim samym podejściu, to jest sformułowaną również przez Griffithsa i Tenenbauma teorię elementarnej oceny związku przyczynowo-skutkowego, nazywaną dalej w skrócie teorią oceny kauzalnej. Teoria oceny kauzalnej posłuży mi jako punkt wyjścia do analizy zagadnień nieco trudniej dostrzegalnych w przypadku teorii generalizacji.

5.3 Griffithsa i Tenenbauma teoria oceny kauzalnej

W najprostszym przypadku ocena związku przyczynowo-skutkowego dotyczy pojedynczego czynnika i pojedynczego efektu, traktowanych jako zmienne binarne. Problem polega wtedy na wnioskowaniu na temat tego związku na podstawie danych informujących o występowaniu efektu w zależności od występowania czynnika. Stosunkowo popularny paradygmat eksperymentalny, służący do badania oceny kauzalnej u ludzi, polega na prezentowaniu takich danych albo w postaci sekwencyjnej, kiedy to kolejno prezentowane są przypadki wystąpienia efektu lub jego braku gdy czynnik był lub nie był zastosowany, albo w postaci listy, kiedy cała sekwencja prezentowana jest jednocześnie, albo wreszcie w postaci zbiorczej, za pomocą tabeli wielodzielczej.

Stosuje się zwykle dwa rodzaje zależności przyczynowo-skutkowych między czynnikiem i efektem, to jest „generatywną" i „prewencyjną". Czynnik generatywny zwiększa prawdopodobieństwo wystąpienia efektu, a czynnik prewencyjny zmniejsza to prawdopodobieństwo. Dla uproszczenia wywodu ograniczę się do postaci zbiorczej i czynników generatywnych.

Przed zaprezentowaniem danych osoby badane informowane są, że każda tabela zawiera wyniki odrębnego, hipotetycznego eksperymentu. Zadaniem osób badanych jest udzielenie w oparciu o te wyniki odpowiedzi na pytanie w rodzaju „na ile Twoim zdaniem między czynnikiem a efektem występuje związek przyczynowo-skutkowy" (Lober

i Shanks, 2000; Buehner, Cheng i Clifford, 2003). Odpowiedzi udzielane są zwykle na skali, na której wartość minimalna reprezentuje przekonanie, że związek w ogóle nie zachodzi, a wartość maksymalna reprezentuje przekonanie, że związek zachodzi w największym możliwym stopniu. Najczęściej przyjmuje się, że liczba prób w grupie kontrolnej hipotetycznego eksperymentu jest taka sama jak liczba prób w grupie eksperymentalnej, a łączna liczba wszystkich prób jest taka sama dla wszystkich tabel.

Do najważniejszych zbiorów danych, służących współcześnie do testowania racjonalnych teorii kauzalnych, należą wyniki eksperymentów Buehnera i Cheng (2003) i Lobera i Shanksa (2000). Przed pojawieniem się teorii Griffithsa i Tenenbauma do wyników tych i wielu innych eksperymentów stosunkowo najlepiej pasowały dwie teorie racjonalne, to jest historycznie pierwsza teoria siły kauzalnej (Allan, 1980) i późniejsza teoria mocy kauzalnej (Cheng, 1997).

Zgodnie z racjonalną teorią siły kauzalnej, ocena związku przyczynowo-skutkowego powinna być funkcją wielkości danej przez:

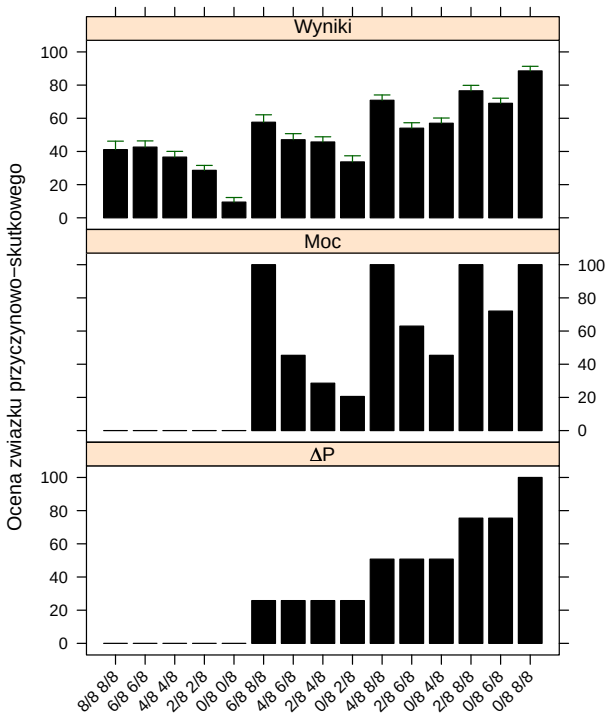
$$\begin{aligned}\Delta P &= \frac{N(e^+, c^+)}{N(e^+, c^+) + N(e^-, c^+)} - \frac{N(e^+, c^-)}{N(e^+, c^-) + N(e^-, c^-)} \\ &= p(e^+|c^+) - p(e^+|c^-)\end{aligned}$$

gdzie $N(e^+, c^+)$ oznacza liczbę wystąpień efektu w warunkach oddziaływania czynnika (znak plus oznacza wystąpienie, a znak minus niewystąpienie efektu lub czynnika), a $p(e^+|c^+) = N(e^+, c^+)/[N(e^+, c^+) + N(e^-, c^+)]$ to oszacowanie warunkowego prawdopodobieństwa wystąpienia efektu gdy czynnik oddziaływuje i odpowiednio dla wszystkich pozostałych kombinacji e^+, e^-, c^+ i c^-. Wartość ΔP informuje o tym, o ile prawdopodobieństwo wystąpienia efektu zwiększa się wraz z pojawieniem się czynnika. Równanie mocy kauzalnej od ΔP odróżnia wprowadzenie dodatkowego wyrażenia w mianowniku:

$$\text{moc kauzalna} = \frac{\Delta P}{1 - p(e^+|c^-) = p(e^-|c^-)}$$

Interpretacja równania mocy jest nieco bardziej skomplikowana. Jak trafnie zauważyła Cheng (1997), ΔP jest miarą współzmienności, a nie przyczynowości. Jeżeli w grupie eksperymentalnej efekt wystąpił u 20 jednostek na 30, a w grupie kontrolnej u 10 na 30 ($\{k : 10/30, e : 20/30\}$), siła przyczynowa wyniesie $\Delta P = 2/3 - 1/3 = 1/3$ i dla wyników $\{k : 20/30, e : 30/30\}$ również $\Delta P = 3/3 - 2/3 = 1/3$. Wydaje się jednak, że siła oddziaływania czynnika była w tych eksperymentach inna. Najlepszą miarą prawdopodobieństwa spontanicznego (niezależnego od zastosowanego czynnika) wystąpienia efektu są wyniki w grupie kontrolnej. Opierając się na tym założeniu, można stwierdzić, że w pierwszym badaniu w grupie eksperymentalnej efekt zamanifestował się w połowie tych przypadków, u których najprawdopodobniej nie wystąpiłby spontanicznie ($20 - 10/(30 - 10)$), a w drugim badaniu efekt zamanifestował się tak silnie,

jak tylko mógł $(30 - 20/(30 - 20))$. Propozycja Cheng polega właśnie na uwzględnieniu informacji o tym, w jakim stopniu wpływ czynnika miał szansę się zamanifestować $(p(e^-|c^-))$. Dzięki uprzejmości Marca Buehnera mogę przedstawić tu wyniki jednego z eksperymentów Buehnera i Cheng z 1997 roku (eksperyment $1B$), omawianego również przez Griffithsa i Tenenbauma[2]:



Rysunek 5.3: Zgodność predykcji modelu siły (ΔP) i mocy kauzalnej z wynikami eksperymentu 1B Buehnera, Cheng i Clifforda

Na wykresie widoczne są predykcje obu modeli i średnie wyniki dla wszystkich osób badanych, wraz z odchyleniami standardowymi (R^2 dla modelu ΔP wyniosło $0{,}79$, a dla modelu mocy $0{,}76$). Byłoby wielką niespodzianką, gdyby zastosowana skala oceny była liniowo związana z predykcjami modelu ΔP i mocy kauzalnej, dlatego podobnie jak Griffiths i Tenenbaum zastosowałem wobec predykcji każdego modelu jednoparametrową funkcję potęgową[3], z wartością parametru γ maksymalizującą korelację między

[2]Nie potrafię zobrazować tych wyników lepiej, niż uczynili to Griffiths i Tenenbaum, dlatego postarałem się stworzyć wykres możliwie podobny do tego, jaki znajduje się w cytowanym artykule.

[3]$sign(x)abs(x)^\gamma$, gdzie x to predykcje modelu

predykcjami a wynikami średnimi. Wyniki przedstawione są w kolejności rosnącej wartości ΔP, a w każdej grupie warunków z jednakowym ΔP układ odpowiada malejącemu prawdopodobieństwu bazowemu, czyli malejącej liczbie efektów zaobserwowanych w grupie kontrolnej.

Przy ustalonej wartości ΔP, model mocy kauzalnej przewiduje spadek wartości oceny wraz z malejącym prawdopodobieństwem bazowym. Pomijając na razie warunki, dla których $\Delta P = 0$, można zaobserwować pewne podobieństwo między wynikami a wzorcem przewidywanym przez model mocy. Oba modele kompletnie nie pasują do danych z warunków przedstawionych na wykresie po stronie lewej.

Jako racjonalne, modele siły i mocy muszą być oparte na błędnych założeniach, skoro wyniki $\{k : 8/8, e : 8/8\}$ mówią coś zupełnie innego na temat ewentualnego związku przyczynowo-skutkowego niż wyniki $\{k : 0/8, e : 0/8\}$, co można wykazać opierając się na tym samym rozumowaniu jak to, które przytoczyłem jako uzasadnienie modelu Cheng. Skoro efekt wystąpił u wszystkich jednostek badawczych zarówno w warunku eksperymentalnym jak i kontrolnym, o ewentualnym generatywnym wpływie czynnika nie da się w zasadzie nic powiedzieć, dlatego że ze względu na efekt sufitowy wpływ nie miał szansy się zamanifestować. Gdy jednak efekt nie ujawnia się w ogóle w obu grupach, wypada stwierdzić, że przypuszczalnie czynnik nie działa, albo działa słabo. Żaden z wymienionych modeli nie korzysta z tej intuicji.

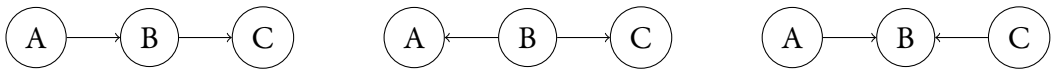
Kluczowa dla uzasadnienia propozycji Griffithsa i Tenenbauma jest obserwacja, że ocena związku przyczynowo-skutkowego polega na rozstrzygnięciu co najmniej dwóch różnych kwestii, to jest *istnienia* związku i jego *siły*. Z perspektywy wnioskowania statystycznego kwestia istnienia związku odpowiada selekcji modelu, a kwestia siły odpowiada szacowaniu wolnych parametrów.

W ujęciu bayesowskim problem wnioskowania na temat związków przyczynowo-skutkowych znajduje eleganckie i sprawdzające się w praktyce rozwiązanie w postaci wnioskowania na modelach graficznych (Pearl, 1998, 2000), nazywanych też sieciami bayesowskimi. Modele graficzne to modele probabilistyczne reprezentowane w postaci grafu. Związek między grafami a modelami probabilistycznymi został poddany formalnej analizie po raz pierwszy przez Pearla (1998) i zdaniem niektórych autorów, dokonania te należą do najważniejszych odkryć we współczesnym wnioskowaniu statystycznym (Gelman i in., 1995) i sztucznej inteligencji (Korb i Nicholson, 2003).

Gdy celem wnioskowania bayesowskiego jest ustalenie czegoś na temat wartości parametrów modelu, najczęściej nie chodzi o łączny rozkład prawdopodobieństwa określony na całym wektorze parametrów, tylko o rozkłady brzegowe dla poszczególnych parametrów. Uśrednianie, a w przypadku parametrów ciągłych całkowanie, pojawia się także w sytuacji wnioskowania na temat modelu. Jeden z problemów technicznych, które do niedawna uniemożliwiały praktyczne zastosowanie wnioskowania bayesowskiego do bardziej złożonych zagadnień, wynika właśnie z wymagań obliczeniowych całkowania w wielu wymiarach. Problem ten stał się znacznie mniej dotkliwy między innymi dzięki zastosowaniu grafów jako reprezentacji umożliwiającej efektywną dekompozycję

łącznego rozkładu prawdopodobieństwa.

W zastosowaniu do modeli probabilistycznych graf reprezentuje bezpośrednie zależności między stałymi, zmiennymi lub parametrami za pomocą krawędzi łączących węzły odpowiadające tym stałym, zmiennym lub parametrom. Graf jest więc reprezentacją strukturalną czy też jakościową modelu. Bodaj najpopularniejszą postacią grafów używaną do wnioskowania statystycznego są tak zwane grafy skierowane acykliczne (ang. *Directed Acyclic Graph*, w skrócie DAG). DAG składa się ze skończonej liczby węzłów i łączących je krawędzi skierowanych (kierunkowych). Acykliczność oznacza, że niedopuszczalne jest, aby można było wędrując wzdłuż krawędzi zgodnie z ich kierunkiem trafić dwa razy na ten sam węzeł. Jeżeli od węzła A prowadzi krawędź do węzła B, węzeł B jest następnikiem węzła A, a węzeł A jego poprzednikiem. Węzły, do których można dotrzeć po krawędziach zgodnie z ich kierunkiem rozpoczynając od węzła A to potomkowie węzła A. Na razie ograniczę się do przypadków, w których węzły reprezentują zmienne losowe. Istnieją trzy typy najprostszych połączeń między trzema węzłami, to jest szeregowe, rozbieżne i zbieżne:



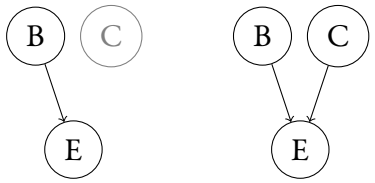
Rysunek 5.4: Trzy podstawowe typy połączeń między węzłami grafu skierowanego acyklicznego

Aby zrozumieć, w jaki sposób struktura grafu pozwala na wnioskowanie na temat (nie)zależności między zmiennymi, dogodnie jest interpretować strzałki jako kanały przepływu informacji, albo jako bezpośrednie probabilistyczne związki przyczynowo-skutkowe. W przypadku połączenia szeregowego, jeżeli znana jest wartość węzła B, węzły A i C są niezależne, ponieważ jakiekolwiek wartości przyjęłaby zmienna A, znajomość tej wartości nie dostarczy żadnej dodatkowej informacji na temat zmiennej C. Na przykład, jeżeli dowiadujemy się o kimś, że jest przeziębiony (A), rośnie prawdopodobieństwo, że ma katar (B), a jeżeli ma katar, to rośnie prawdopodobieństwo, że będzie kichał (C). Jeżeli jednak wiemy, że ma katar, informacja na temat źródła kataru (przeziębienie) nie powinna znacząco wpłynąć na ocenę prawdopodobieństwo kichania. Informacja przepływa przez połączenie szeregowe wtedy i tylko wtedy, gdy wartość zmiennej środkowej nie jest znana.

W przypadku połączenia rozbieżnego, jeżeli wiemy coś na temat wartości A, możemy także powiedzieć coś na temat wartości B, a więc również C. Na przykład, jeżeli wiemy, że ktoś ma katar (A), możemy przypuścić, że jest przeziębiony (B), a więc może mieć podwyższoną temperaturę (C). Jeżeli jednak wiemy, że ktoś jest przeziębiony, informacja na temat kataru nie powinna wpłynąć na ocenę prawdopodobieństwa gorączki. Informacja przepływa przez połączenie rozbieżne wtedy i tylko wtedy, gdy wartość zmiennej środkowej nie jest znana.

Najmniej oczywiste, a przez to najciekawsze własności ma połączenie zbieżne. Połączenie takie może reprezentować między innymi związek między skutkiem a dwoma możliwymi przyczynami. Na przykład, znużenie (B) może być wywołane wysiłkiem fizycznym (A) albo lekturą niezbyt pasjonującej monografii matematycznej (C). Przypuśćmy, że dysponując takim uproszczonym modelem chcemy wywnioskować coś na temat źródła znużenia napotkanego właśnie kolegi. Zapytany, co ostatnio czytał, kolega twierdzi, że niedawno skończył ostatni rozdział „Wstępu do matematyki współczesnej". Skoro tak, to pewnie właśnie dlatego jest znużony, a więc subiektywne prawdopodobieństwo, że uprawiał niedawno sport, jest teraz mniejsze niż było, zanim poznaliśmy odpowiedź. Można powiedzieć, że lektura i wysiłek fizyczny są alternatywnymi (przyczynowymi) wyjaśnieniami tego samego zdarzenia. Jeżeli wzrasta wiarygodność jednego z wyjaśnień, spada wiarygodność wyjaśnienia alternatywnego. Gdybyśmy jednak nie wiedzieli nic na temat samopoczucia kolegi, stąd, że niedawno uprawiał sport, nie wynikałoby jeszcze nic na temat tego, co też mógł w międzyczasie czytać. Informacja przepływa przez połączenie zbieżne wtedy i tylko wtedy, gdy wartość zmiennej środkowej jest znana.

Problem wnioskowania na temat probabilistycznego związku przyczynowo-skutkowego można przedstawić jako wnioskowanie w oparciu o dwa, wzajemnie wykluczające się modele graficzne, takie jak te poniżej:



Graf 1 Graf 2

Rysunek 5.5: Grafy reprezentujące alternatywne modele dla problemu wnioskowania o zależności przyczynowo-skutkowej

Graf pierwszy reprezentuje model, w którym efekt E zależy od bliżej nieokreślonych przyczyn B („tło przyczynowe"), ale nie zależy od czynnika eksperymentalnego C. Graf drugi reprezentuje model, w którym czynnik i pozostałe przyczyny wypływają niezależnie na występowanie efektu. Wnioskowanie na temat związku przyczynowo-skutkowego między C i E polegałoby więc na ustaleniu, który model jest przypuszczalnie prawdziwy, a o ile prawdziwy jest model drugi, jaka jest siła związku między C i E.

Żeby zastosować wnioskowanie bayesowskie trzeba jeszcze zinterpretować strukturę obu modeli w kategoriach rozkładów zmiennych losowych i zależności między tymi rozkładami. Dla danych w postaci tabel wielodzielczych i zmiennych binarnych naturalnym wyborem jest model złożony ze zmiennych o rozkładzie dwumianowym.

Griffiths i Tenenbaum przyjęli całkiem rozsądne założenie, że B i C mogą wywołać efekt niezależnie. Ponieważ prawdopodobieństwo sumy dwóch niezależnych zdarzeń

jest równe sumie ich prawdopodobieństw minus prawdopodobieństwo łącznego zajścia obu zdarzeń[4], zależność między parametrami w modelu reprezentującym występowanie związku przyczynowo-skutkowego (graf drugi) powinna być następująca:

$$\begin{aligned}\theta_k &= \theta_B \\ \theta_e &= \theta_B + \theta_C - \theta_B\theta_C\end{aligned}$$

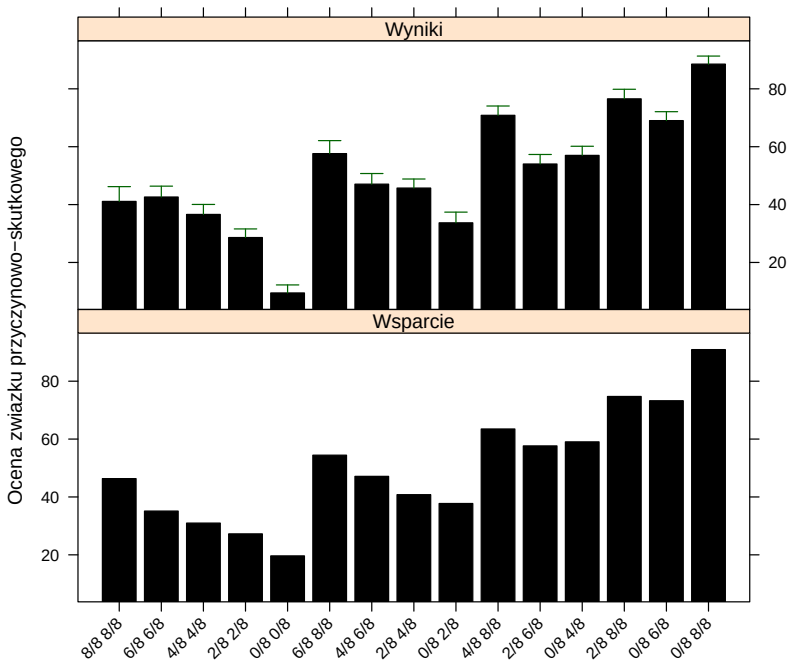
gdzie θ_i to prawdopodobieństwo wystąpienia efektu w grupie kontrolnej ($i = k$) i eksperymentalnej ($i = e$), a θ_B i θ_c to prawdopodobieństwa wywołania efektu odpowiednio przez tło przyczynowe i czynnik eksperymentalny. Hipotetyczne wyniki eksperymentu traktujemy jako realizacje zmiennej o rozkładzie dwumianowym, wobec czego prawdopodobieństwo danych ze względu na model wynosi:

$$\binom{N_k}{N_{ek}}\theta_k^{N_{ek}}(1-\theta_k)^{N_k-N_{ek}}\binom{N_e}{N_{ee}}\theta_e^{N_{ee}}(1-\theta_e)^{N_e-N_{ee}}$$

gdzie N_i to liczba wszystkich obserwacji, a N_{ei} to liczba wystąpień efektu w grupie kontrolnej ($i = k$) i eksperymentalnej ($i = e$). W modelu reprezentowanym przez graf pierwszy prawdopodobieństwo wystąpienia efektu zależy od jednego wolnego parametru, określającego prawdopodobieństwo zadziałania tła przyczynowego.

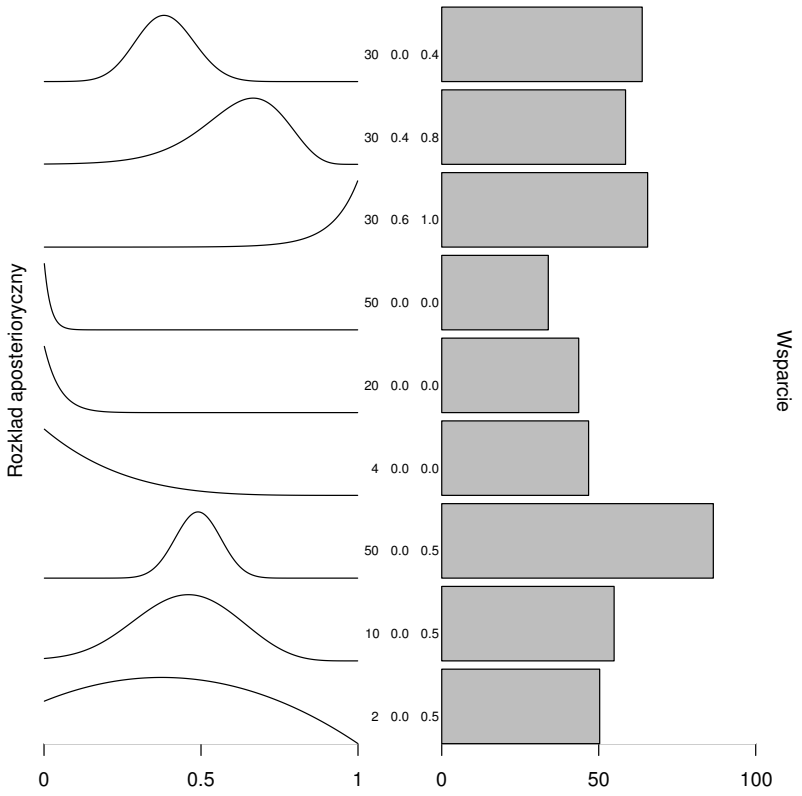
Przyjmując dla obu modeli i wszystkich parametrów priory nieinformacyjne (rozkład jednostajny dla parametrów, priory równe $1/2$ dla modeli), na podstawie wyników hipotetycznego eksperymentu można teraz obliczyć wartość czynnika Bayesa (K). Wartość $K < 1$ oznacza, że wyniki wspierają bardziej model pierwszy (brak związku między C i E), a $K > 1$ oznacza silniejsze wsparcie dla modelu drugiego. Na poniższym wykresie zestawiłem dane uzyskane przez Buehnera i Cheng i logarytm wartości czynnika Bayesa dla obu modeli, przeskalowany za pomocą jednoparametrowej funkcji potęgowej:

[4]Bez tego odejmowania łączne zajście obu zdarzeń niezależnych byłoby liczone podwójnie. $p(X \cup Y) = p(X) + p(Y)$ tylko wtedy, gdy $p(X \cap Y) = 0$, a więc zdarzenia X i Y wzajemnie się wykluczają, czyli nie są niezależne.



Rysunek 5.6: Zgodność predykcji modelu wsparcia z wynikami eksperymentu 1B Buehnera, Cheng i Clifforda

Osoby badane zachowują się zadziwiająco zgodnie z predykcjami modelu wsparcia, o czym można się przekonać również na podstawie wartości statystyki R^2, równej $0{,}94$, w tym przypadku dość wiarygodnej, ponieważ model nie jest dopasowywany. To oznacza, że formułowane przez ludzi w warunkach laboratoryjnych, na podstawie stosunkowo sztucznych bodźców, intuicyjne oceny związku przyczynowo skutkowego są bardzo podobne do ocen racjonalnego agenta, dokonującego wnioskowania bayesowskiego na temat struktury modelu, na podstawie danych zawartych w tabeli wielodzielczej. Następny wykres ilustruje kilka kluczowych własności tego modelu:



Rysunek 5.7: Związek między brzegowym rozkładem aposteriorycznym dla prawdopodobieństwa wystąpienia efektu na skutek oddziaływania czynnika (θ_C), a relatywnym wsparciem dla modelu reprezentującego występowanie zależności przyczynowo-skutkowej

Środkowa kolumna zawiera kolejno od lewej liczbę obserwacji w każdej z grup, proporcję zaobserwowanych efektów w grupie kontrolnej i odpowiednią proporcję w grupie eksperymentalnej. Trzy dolne wykresy obrazują wpływ całkowitej liczby obserwacji dla ustalonych proporcji efektów w obu grupach (0 w grupie kontrolnej i $0{,}5$ w grupie eksperymentalnej), wskazujących na istnienie związku przyczynowo-skutkowego. Na trzech środkowych wykresach widać wpływ wielkości próby, gdy efekt w ogóle nie jest obserwowany. Sześć dolnych wykresów pozwala zrozumieć, w jaki sposób wsparcie zależy od dwóch rodzajów informacji - punktowego oszacowania związku θ_C i aposteriorycznej niepewności co do wartości tego oszacowania. Trzy górne wykresy ilustrują wpływ prawdopodobieństwa bazowego (wystąpienia efektu w grupie kontrolnej) przy tej samej całkowitej liczbie prób i ustalonej wartości ΔP (tutaj $0{,}4$).

Wyprowadzenie predykcji z modelu Griffithsa i Tenenbauma nie polega na szacowaniu wolnych parametrów, ponieważ zaproponowany przez tych autorów model tako-

wych nie posiada. Wsparcie obliczane jest wyłącznie na podstawie liczebności w tabeli wielodzielczej, a nie na podstawie wyników eksperymentu w postaci ocen uzyskanych od osób badanych. Badacz może sformułować predykcje oparając się tylko na prezentowanych osobom badanym bodźcach, a mimo to predykcje te pasują znakomicie do wyników uzyskanych przez Buehnera i Cheng i wielu innych autorów.

Model można łatwo zastosować do oceny kauzalnej na podstawie bodźców w postaci częstości występowania efektu w jednostce czasu, wystarczy zamienić rozkład dwumianowy na rozkład Poissona, co też autorzy zrobili, znowu uzyskując wysoką zgodność predykcji z ocenami rzeczywistymi. Griffiths i Tenenbaum sprawdzili też, jak dobrze model radzi sobie z wynikami badań Lobera i Shanksa (2000), a także dwunastu innych eksperymentów, w których zastosowano różne instrukcje i formaty prezentacji, nie tylko tabele wielodzielcze, czasami manipulowano też całkowitą liczbą obserwacji. Wzorce uzyskane w tych badaniach nie są wcale proste. Model wsparcia przewiduje między innymi niemonotoniczne zmiany oceny w funkcji prawdopodobieństwa bazowego $p(e^+|c^-)$ przy ustalonej wartości ΔP. Model siły nie przewiduje wtedy żadnych zmian, a model mocy przewiduje liniowy spadek oceny. Ta niemonotoniczność również znalazła potwierdzenie empiryczne.

Nawet w przypadku stosunkowo elastycznych modeli, dopasowanie bywa niewiele lepsze lub nawet gorsze niż dopasowanie pozbawionego jakiejkolwiek elastyczności modelu wsparcia do nieoczywistych wyników imponującej liczby zróżnicowanych badań. Być może jest to świadectwem mojej ignorancji, ale jak dotąd nie spotkałem się z niczym porównywalnym, ani w literaturze z zakresu poznawczej psychologii eksperymentalnej, ani w literaturze eksperymentalnej z zakresu psychologii w ogóle, co nie byłoby przykładem zastosowania teorii racjonalnej.

Przykładów mniej lub bardziej udanego zastosowania teorii racjonalnych jest więcej. Przegląd tego typu teorii i badań znajduje się u Griffithsa, Kempa i Tenenbauma (w druku), Andersona (1991b) i Chatera i Oaksforda (2000), pojawił się też osobny numer Trends in Cognitive Sciences, poświęcony probabilistycznym modelom poznania (Chater, Tenenbaum i Yuille, 2006b). Cechą wspólną znakomitej większości tych propozycji jest wykorzystanie na poziomie teorii wnioskowania bayesowskiego lub bayesowskiej teorii decyzji. Nie oznacza to oczywiście, że wszystkie takie próby są równie udane. Anderson (1990) zaproponował własną wersję racjonalnego modelu bayesowskiego dla zadań oceny kauzalnej, zawierającego początkowo, w co być może trudno uwierzyć, 7 wolnych parametrów, a w wersji uproszczonej 4 (Anderson i Sheu, 1995). W modelu tym każda kombinacja wartości czynnika i tła przyczynowego jest reprezentowana przez wolny parametr. Te parametry lub określone na nich priory są szacowane na podstawie danych, a mimo to model dopasowuje się słabo, nie radzi sobie między innymi ze wspomnianym już niemonotonicznym efektem prawdopodobieństwa bazowego przy ustalonej wartości ΔP (Griffiths i Tenenbaum, 2005).

5.4 Podsumowanie

Z perspektywy racjonalnej procesy poznawcze traktowane są jako rozwiązania określonych problemów. Nic nie gwarantuje, że teoria racjonalna będzie dostarczała choćby akceptowalnie trafnych predykcji tylko dlatego, że jest racjonalna. Nic nie gwarantuje również, że sformułowanie problemu i jego rozwiązania będzie od razu trafne, tylko dlatego, że rozwiązanie sprawia wrażenie racjonalnego, a problem wydaje się być właśnie tym, który faktycznie rozwiązują osoby badane. Racjonalne podejście do tworzenia teorii, wyjaśniania i przewidywania zachowania ma jednak pewne cechy szczególne, wyraźnie odróżniające to podejście od tego, które jest tradycyjnie najczęściej stosowane w psychologii poznawczej.

Dzięki temu, że teoria racjonalna abstrahuje od mechanizmu obliczeniowego, możliwe staje się znaczące ograniczenie elastyczności teorii. Własności mechanizmu, takie jak czas trwania i kolejność poszczególnych procesów, struktura reprezentacji, albo dynamika zmiany sił asocjacji i aktywacji, z samej swojej natury wymagają wprowadzenia do modelu wolnych parametrów. Mechanizm obliczeniowy nie jest bezpośrednio obserwowalny i to, w jaki sposób dokładnie działa, nie może być rozstrzygnięte teoretycznie. Zanim nie przeprowadzi się odpowiedniego eksperymentu, nie wiadomo ile czasu może zajmować każdy z hipotetycznych procesów składowych, jakie jest prawdopodobieństwo błędu związane z tymi procesami, jakie są początkowe siły asocjacji między elementami zapisanymi w pamięci, i tak dalej.

Po dopasowaniu modelu wartości tych parametrów będą zawsze do pewnego stopnia zależały od czynników bezpośrednio nie interesujących badacza, takich jak dokładny czas prezentacji bodźców, warunki oświetleniowe w laboratorium, pora dnia, wiek osób badanych, i wielu innych. W konkretnym przypadku tylko część tych czynników może badacza interesować, pozostałe zaś muszą być traktowane jako źródło niekontrolowanej zmienności, dlatego wartości tylko niektórych wolnych parametrów będą dostarczały cennej informacji. Model mechanizmu obliczeniowego zawsze musi być stosunkowo elastyczny, tym bardziej, im bardziej złożony jest modelowany proces. Teoria racjonalna jest pozbawiona tej kłopotliwej elastyczności, ceną jest jednak niezdolność do udzielenia odpowiedzi na wiele pytań, przede wszystkim na wszystkie te pytania, które dotyczą właśnie działania mechanizmu obliczeniowego.

Przytoczone w tym rozdziale przykłady teorii mają też inne, zasadnicze zalety, które nie dają się sprowadzić do dokładności predykcji. Wydaje się, że dzięki ich specyfice teorie te można naturalnie uogólniać na inne zadania, chociaż oczywiście a priori nic nie gwarantuje, że uogólnienia będą trafne. Co jeszcze ważniejsze, jeżeli teoria racjonalna prowadzi do nieoczywistych predykcji i te predykcje okażą się zgodne z obserwacjami, coś wynikającego z charakteru tych teorii sprawia, że taka zgodność jest przekonującym świadectwem trafności interpretacji nieobserwowalnego procesu *jako rozwiązania zadania* postulowanego przez daną teorię racjonalną. Inaczej mówiąc, jeżeli na przykład osoby badane, poproszone o dokonanie oceny związku przyczynowo-skutkowego za-

chowują się w dużym stopniu zgodnie z nieoczywistymi predykcjami teorii racjonalnego wnioskowania o zależnościach przyczynowo-skutkowych, wypada po prostu stwierdzić, że dokonują oceny związku przyczynowo-skutkowego w znaczeniu zakładanym przez tą teorię. Taki wniosek nie jest bynajmniej trywialny. Zwykle nie wiadomo, jakie właściwie zadanie rozwiązują osoby badane - wiadomo tylko, jakie zadanie zostało im narzucone przez eksperymentatora. Utożsamienie a priori zadania narzuconego i faktycznie wykonywanego byłoby równoznaczne z przypisywaniem eksperymentatorowi nadludzkich mocy.

Wymienione cechy teorii racjonalnych wskazują na to, że takie teorie mogą odgrywać w psychologii poznawczej szczególną rolę. Analizę tego zagadnienia rozpocznę od przyjrzenia się bliżej poglądom najbardziej wpływowych zwolenników podejścia racjonalnego.

Rozdział 6

Analiza racjonalna

W późnych latach osiemdziesiątych Anderson (1988/1991c) zaproponował pewną strategię badawczą, która miała między innymi dostarczyć częściowego rozwiązania problemu identyfikacji architektury. Na skutek zastosowania tej strategii powstało kilka nowych, racjonalnych teorii pamięci, kategoryzacji i innych procesów poznawczych. Teorie te doprowadziły między innymi do odkrycia nowych regularności, które zostały później wykorzystane w celu skorygowania niektórych założeń ACT-R'a, przypuszczalnie najpopularniejszego i poddanego najbardziej wszechstronnej weryfikacji empirycznej spośród wszystkich rozwijanych współcześnie zintegrowanych modeli umysłu. Aby zrozumieć powody, dla których Anderson uznał zastosowanie analizy racjonalnej za konieczne, dobrze jest najpierw, przynajmniej w zarysie, zapoznać się z samym ACT-R'em. Opis tego modelu oparłem głównie na wersji 4.0 (Anderson i Lebiere, 1998). Aktualna wersja ma numer 6.0 i zawiera szereg ważnych dodatków, jest przez to jednak o wiele bardziej złożona. Ze względu na wnioski do których zmierzam, wystarczy jeżeli ograniczę się do wersji prostszej.

6.1 Zarys struktury i mechanizmów działania modelu zintegrowanego ACT-R

Jak już wspomniałem, ACT-R jest aktualnie być może najpopularniejszym zintegrowanym modelem umysłu. Względnie wyczerpujący opis tego modelu zająłby tutaj zdecydowanie zbyt wiele miejsca. Nawet gdyby udało mi się go zwięźle i wyczerpująco opisać, Czytelnik nieposiadający doświadczenia w jego użyciu miałby i tak nikłe szanse na zrozumienie, jak model działa. Z podręcznika dostępnego na stronie domowej ACT-R'a można się dowiedzieć, że nabycie elementarnej wprawy w stosowaniu modelu zajmuje studentowi przeciętnie około 9 tygodni, a nie da się dobrze zrozumieć jak ACT-R działa inaczej, niż poprzez samodzielne tworzenie i testowanie konkretnych modeli.

Typowy model zintegrowany (nazywany częściej architekturą poznawczą) składa się

z następujących elementów (Langley, Laird i Rogers, 2009):

- pamięci krótko i długoterminowej, zawierającej informacje o przekonaniach, celach i wiedzy agenta,
- reprezentacji elementów zapisanych w tych magazynach pamięciowych i ich organizacji w większe struktury,
- procesów operujących na tych strukturach, włączając w to mechanizmy, które wykorzystują je do wykonywania zadań i mechanizmy uczenia się, które zmieniają zawartość pamięci.

W przypadku ACT-R'a wyróżnia się systemy pamięci deklaratywnej i proceduralnej. Pamięć deklaratywna (świadoma) zawiera „porcje wiedzy" (ang. *chunks*) dotyczącej faktów, takich jak wydarzenia, własności znanych obiektów, na przykład liczb, i tym podobne. Każda taka porcja wiedzy ma określony typ i pewną liczbę pól określonego typu. Na przykład, fakt dotyczący dodawania (typ porcji) może określać, że 2 dodać 2 (pola typu „liczba dodawana") równa się 4 (pole typu „wynik"). Pamięć proceduralna (nieświadoma) składa się z tak zwanych reguł produkcji, czyli reguł postaci jeżeli-to. Ważną cechą reguł produkcji jest ich kierunkowość. Po lewej stronie reguły (strona „jeżeli") określone są warunki stosowalności, a po stronie prawej (strona „to") działania, będące skutkiem zastosowania reguły. Warunki stosowalności dotyczą zawartości rozmaitych buforów, takich jak bufor celów, bufor pamięci deklaratywnej i bufory sensoryczne. Bufory są interfejsami do poszczególnych podsystemów (nazywanych „modułami"), do których należą między innymi podsystem pamięci deklaratywnej, percepcji wzrokowej, słuchowej, czy generowania reakcji motorycznych. W dowolnej chwili każdy bufor może zawierać co najwyżej jedną porcję informacji, na przykład bufor celów może zawierać tylko jeden aktywny cel[1]. Rezultatem zastosowania reguły może być modyfikacja zawartości buforów, wysłanie żądania do buforów, lub żądanie wykonania określonych działań, takich jak reakcje motoryczne.

Pamięć deklaratywna i proceduralna są najważniejszymi elementami ACT-R'a. W każdym kolejnym kroku symulacji, na podstawie zawartości buforów ustalany jest zbiór reguł w danej chwili spełnionych (dopasowanie wzorca po lewej stronie reguły do zawartości buforów). Spośród reguł spełnionych wykonywana jest jedna, wyłoniona ze względu na aktualny poziom aktywacji. Nowe produkcje mogą powstawać z porcji wiedzy („kompilacja produkcji"), a nowe porcje wiedzy mogą powstawać w konsekwencji realizacji celów, albo na skutek procesów percepcyjnych. Mechanizmem kontroli jest tutaj mechanizm odpowiedzialny za selekcję reguł.

[1]Nie znaczy to jednak, że ACT-R realizuje zawsze tylko jeden cel. Aktualna *hierarchia* celów reprezentowana jest za pomocą stosu.

Wyróżnia się poziom symboliczny (dyskretne elementy składowe reguł i porcji) i subsymboliczny. Na poziomie subsymbolicznym działają procesy odpowiedzialne za zmianę parametrów ciągłych, reprezentujących na przykład siłę asocjacji porcji z zawartościami pól, albo poziom aktywacji bazowej porcji lub reguł. Ten subsymboliczny proces, z trudnych do ustalenia powodów określany przez Andersona jako „neuropodobny", działa w dużym stopniu równolegle. Zdaniem samego autora, działanie całego system jest w większym stopniu konsekwencję procesów subsymbolicznych niż symbolicznych (s. 9, Anderson i Lebiere, 1998). Model faktycznie składa się ze stosunkowo prostych elementów, ale nazwanie go „prostą teorią złożonych procesów poznawczych" wymaga dosyć szczególnego rozumienia prostoty. Jak piszą Anderson i Lebiere w pierwszym rozdziale podręcznika (Anderson i Lebiere, 1998):

> Relatywnie drobne poprawki teoretyczne w systemie ACT-R doprowadziły łącznie do uzyskania przez porcje i produkcje głębokiego sensu psychologicznej rzeczywistości. Osiągnięty został moment, w którym uważamy, że zidentyfikowaliśmy atomowe komponenty poznania, tak jak to zapowiada tytuł tej książki. Uważamy teraz, że możemy uznać produkcje za definicyjne jednostki w jakich przebiega myśl, a porcje za definicyjne jednostki, w jakich wiedza jest wydobywana z pamięci deklaratywnej.
>
> (...) Używając terminu „atom" odwołujemy się do metafory nauk fizycznych i uważamy, że jest to stosowna metafora. Porcje i produkcje w ACT-R 4.0 są na tak niskim poziomie, jak to tylko można osiągnąć w procesie symbolicznej dekompozycji myślenia.
>
> (...) Tak jak identyfikacja atomu była w znacznym stopniu warunkiem umożliwiającym postęp naukowy na poziomie subatomowym i superatomowym, tak jesteśmy przekonani, że aktualna wersja ACT-R umożliwia podobny postęp w badaniach dotyczących poznania.
>
> (s. 13, tamże)

ACT-R ma pozwalać na wyjaśnienie wielu zjawisk bez pomocy „czarodziejskich sztuczek", czyli ma być:

- Ugruntowany eksperymentalnie, co oznacza, że model jest w pewnym sensie samowystarczalny. W przeciwieństwie do wielu wysokopoziomowych teorii poznawczych, mechanizm odpowiedzialny za percypowanie bodźców i generowanie reakcji jest wyrażony w modelu jawnie, za pomocą systemów percepcyjnych i motorycznych, powiązanych z działaniem produkcji.
- Źródłem detalicznych i precyzyjnych predykcji[2], co oznacza predykcje dotyczące

[2]Dokłanie, „detailed and precise accounting of data" (s. 15). Angielski termin „account" można tłumaczyć jako przewidywanie, wyjaśnianie, bądź jedno i drugie. Z kontekstu wynika, że chodzi raczej o przewidywanie.

każdego aspektu badanego procesu, jaki jest tylko rejestrowany przez oprogramowanie eksperymentalne.

- Oparty na uzasadnionej parametryzacji[3] - wartości wolnych parametrów nie zmieniają się arbitralnie między eksperymentami w celu zapewnienia dopasowania do danych.

- Wiarygodny neuronalnie, przez co należy rozumieć między innymi to, że ACT-R zawiera specyfikację związków (mapowanie) między elementami modelu a obszarami mózgu.

(s. 14-17, tamże)

Anderson i Lebiere wymieniają jeszcze zdolność uczenia się z doświadczenia i możliwość modelowania złożonych procesów poznawczych, takich jak uczenie się rozwiązywania zadań z fizyki.

Bez przesady można powiedzieć, że lista zróżnicowanych wyników badań, które wydają się wspierać ten model, jest imponująca. Na stronie domowej ACT-R'a dostępne są bezpłatnie liczne publikacje, dotyczące wyników badań uwagi selektywnej, pamięci krótko i długoterminowej, percepcji, kategoryzacji, przełączania się między zadaniami, rozwiązywania mniej, lub bardziej złożonych problemów, „kontroli wykonawczej", zadań jednoczesnych, psychologicznego okresu refrakcji, niektórych aspektów rozumienia języka naturalnego i produkcji mowy, w tym rozumienia metafor, zachowania się podczas prowadzenia samochodu lub samolotu, nawigacji po stronach internetowych, i tak dalej. Autorzy modelu mogą się również pochwalić niebagatelnymi sukcesami aplikacyjnymi. Do wyjątkowo udanych należą oparte na tym modelu systemy tutoringu, czyli zautomatyzowane systemy nauczania, dostarczające uczniom instrukcje i informacje zwrotne w trakcie wykonywania zadań (Psotka i Mutter, 1988). Pomimo tego, że wiele zróżnicowanych wyników zdaje się na pierwszy rzut oka świadczyć o trafności ACT-R'a, trudno powiedzieć, na ile trafne są *poszczególne* założenia tego modelu.

Zdolność modelu do symulowania przebiegu działania procesów poznawczych na poziomie milisekund, którą chwali się Anderson, sprowadza się w istocie do tego, że wykonanie każdej reguły zajmuje zwykle dokładnie 50 milisekund[4]. Zakłada się, że przeszukiwanie czegoś, co odpowiada pamięci krótkoterminowej (bufor celów), przebiega równolegle, a wydobywanie elementów z pamięci deklaratywnej przebiega szeregowo. Niemal każde zadanie, którego przebieg jest modelowany za pomocą ACT-R'a znacząco obciąża ten bufor, nie wiadomo jednak, które właściwie wyniki zmuszają do zgody na

[3](ang. *principled parameters*)

[4]Dopuszcza się, aby czas wykonania reguły produkcji uległ zmianie na skutek procesu uczenia się, ale możliwość ta rzadko jest wykorzystywana.

równoległe przeszukiwania pamięci krótkoterminowej. Nie wiadomo też, w jakich warunkach to konkretne założenie dałoby się testować, używając do tego całego modelu.

Odgrywający, jak przyznał to Anderson, niezwykle ważną rolę mechanizm uczenia się jest tak skonstruowany, że wynika z niego potęgowa zależność poziomu wykonania od czasu ćwiczenia. Podobnie jak wielu innych autorów, Anderson wziął sobie do serca rezultaty podjętych przez Newella i Rosenblooma (1981) prób dopasowywania krzywych do wyników w postaci czasów reakcji. Na podstawie tych prób autorzy stwierdzili, że uczenie się zdecydowanej większości zadań powoduje zmiany poziomu wykonania, do których lepiej pasuje krzywa potęgowa, a nie na przykład zaproponowana przez Ebbinghausa krzywa wykładnicza. Uniwersalność tego rezultatu zrobiła na Newellu i Rosenbloomie tak duże wrażenie, że efekt nazwali „potęgowym prawem uczenia się". Wkrótce prawie wszystkie teorie uczenia się „wyjaśniały to prawo". Logan (2002) twierdzi wręcz, że gdyby miało się okazać, że zależność nie jest potęgowa, większość współczesnych teorii uczenia się należałoby odrzucić. Jeżeli Logan ma rację, wsparcie empiryczne dla tych teorii byłoby bardzo słabe niezależnie od tego, jaki jest prawdziwy kształ krzywej zapominania, pozostaje więc żywić nadzieję, że opinia ta jest przesadzona. Z opublikowanych w roku 2000 wyników analiz Heathcote'a, Browna i Mewhorta (2000) wynika jednak, że „prawo" powinno być raczej w przybliżeniu wykładnicze, a nie potęgowe. Pomijając już inne mankamenty metodologiczne, Newell i Rosenbloom dopasowywali krzywe do danych uśrednionych po osobach badanych. W ogólnym przypadku więcej niż jedna funkcja liniowa po uśrednieniu jest nadal liniowa, ale uśrednienie więcej niż jednej funkcji potęgowej (lub wykładniczej) przestaje być funkcją potęgową (lub wykładniczą). Na podstawie wyników znacznie bardziej rozsądnej analizy, bo między innymi uwzględniającej dane na poziomie indywidualnym, Heathcote i in. stwierdzili, że funkcja wykładnicza pasuje lepiej niż potęgowa do *wszystkich* z 40 zbiorów danych, pochodzących od 475 osób, biorących udział w 24 różnych eksperymentach. Pomimo ogromnej liczby badań, których rezultaty okazały się zgodne z „detalicznymi i precyzyjnymi" predykcjami ACT-R'a, problematyczność tego założenia nie dała o sobie znać. Dostarczający „wiarygodności neuronalnej" dla tej „zunifikowanej teorii umysłu" poziom subsymboliczny działa w taki sposób, aby wynikało stąd między innymi „prawo potęgowe". Ani potęgowe, ani wykładnicze prawo uczenia się nie mają jednak jak dotąd żadnego przekonującego uzasadnienia teoretycznego.

Nie jest łatwo odpowiedzieć na pytanie, które konkretnie założenia ACT-R'a należy uznać za względnie uzasadnione, a które można spokojnie wymienić na inne, bez większej szkody dla i tak nie dającej się bliżej ustalić trafności predykcji modelu. Jest oczywiste, że problem identyfikowalności architektury nie zniknie tylko dzięki temu, że model obliczeniowy będzie zintegrowany, czyli bardziej wyczerpujący. Pod wieloma względami model jest znakomity i niewątpliwe sprawdził się wielokrotnie jako narzędzie aplikacyjne, ale nie można powiedzieć, że reprezentuje prostą teorię złożonych procesów poznawczych, albo że jego stosowanie znacząco ułatwia rozstrzygnięcie podstawowych teoretycznych kwestii. Z powodów, które przytoczę w dalszej części pracy, nie jest też

wcale oczywiste, co właściwie ACT-R wyjaśnia. Program analizy racjonalnej został zaproponowany w moim odczuciu po to, żeby poradzić sobie z tego rodzaju problemami, chociaż deklarowane przez autora tego programu powody tylko do pewnego stopnia się z nimi pokrywają.

6.2 Krótka historia rozwoju programu analizy racjonalnej

Analiza racjonalna miała między innymi dostarczyć częściowego rozwiązania problemu identyfikacji architektury. Zdaniem Andersona (1988/1991c), zwolennicy określonych architektur powołują się zwykle na pewne charakterystyczne regularności (ang. *signature phenomena*), takie jak potęgowe prawo uczenia się, które mają znajdywać naturalne wyjaśnienie właśnie dzięki przyjęciu tej a nie innej architektury. Jak trafnie zauważa autor, regularności te nie wystarczają jednak do rozstrzygnięcia między architekturami, ponieważ liczba zgodnych z nimi, możliwych architektur, opartych na wykluczających się założeniach, dotyczących mechanizmu i struktury umysłu jest zbyt duża. Anderson powołuje się przy tym na problem identyfikacji systemów szeregowych i równoległych, odrębności systemów pamięci krótko i długoterminowej i problem identyfikacji struktury reprezentacji. Analiza racjonalna ma być realizacją, cytowanego przez Andersona, genialnego w swej prostocie spostrzeżenia Marra:

> Zrozumienie algorytmu staje się często dużo łatwiejsze, gdy zrozumie się rozwiązywany problem, niż gdy bada się mechanizm (wraz z implementacją sprzętową), za pomocą którego ten problem jest rozwiązywany.
>
> (s. 27, Marr, 1982)

Anderson w ten sposób ujmuje zalety tego, jak sam to przyznaje, niedocenianego wcześniej przez siebie podejścia:

> 1. Jak podkreślił Marr, zrozumienie natury problemu dostarcza wyraźnych wskazówek na temat możliwych mechanizmów. To jest szczególnie ważne w przypadku procesów poznawczych wyższego rzędu, gdzie trzeba się zmierzyć z oszałamiającym bogactwem możliwych mechanizmów i astronomiczną przestrzenią ich możliwych kombinacji, której przeszukiwanie jest konieczne do zidentyfikowania prawdziwej architektury.
> 2. Znowu jak to podkreślił Marr, to [podejście] umożliwia głębsze zrozumienie mechanizmów. Możemy zrozumieć dlaczego działają tak a nie inaczej, zamiast postrzegać je jako arbitralne (ang. *random*) konfiguracje elementów obliczeniowych.

> 3. Psychologia poznawcza staje w obliczu fundamentalnych nierozstrzygalności, takich jak szeregowość-równoległość przetwarzania, status odrębnego magazynu pamięci krótkoterminowej, albo struktura wewnętrznej reprezentacji. Skupienie się na problemie pozwala wznieść się na poziom abstrakcji, który znajduje się ponad tym, na którym musimy sobie poradzić z tymi nierozstrzygalnościami.
>
> (s. 2, Anderson, 1988/1991c)

Program analizy racjonalnej ma być oparty na tak zwanej „zasadzie racjonalności", zgodnie z którą „system poznawczy optymalizuje adaptację organizmu do środowiska" (s. 3, tamże). Zasadę racjonalności należy zdaniem Andersona oceniać nie na podstawie uzasadnienia dostarczonego przez badania ewolucyjne, ale na podstawie sukcesu predykcyjnego opartych na niej teorii. Nieracjonalność zachowania jest według Andersona znacznie bardziej wątpliwa, niż mogłyby na to wskazywać liczne wyniki dotychczasowych badań. Wyniki te bywają błędnie interpretowane, ponieważ nie uwzględnia się ani kosztów związanych z „obliczaniem zachowania", ani środowiska, do którego organizm w procesie ewolucji faktycznie się zaadaptował, ani celów o rzeczywistym znaczeniu dla organizmu. Uwzględnienie kosztów związanych z procesami obliczeniowymi miało przypuszczalnie pozwalać również na uzgodnienie programu analizy racjonalnej z tezą Simona o ograniczonej racjonalności organizmów żywych (Simon, 1972, 1975), o których będzie mowa później.

Konstruowanie teorii racjonalnej ma przebiegać w 6 krokach:

1. Dokładne określenie celów systemu poznawczego.
2. Stworzenie formalnego modelu środowiska, do którego system się zaadaptował.
3. Przyjęcie minimalnych założeń dotyczących kosztów obliczeniowych.
4. Ustalenie optymalnego rozwiązania ze względu na wyniki przeprowadzenia kroków 1-3.
5. Ustalenie zgodności predykcji ze znanymi wynikami badań.
6. Jeżeli predykcje są niezgodne, powrót do kroków wcześniejszych.

Anderson twierdzi dalej, że w ten sposób uzyskuje się teorię abstrahującą od mechanizmu, opartą w znacznym stopniu na założeniach dotyczących tego, co znajduje się na zewnątrz umysłu, czyli środowiska. Do wczesnych przykładów zastosowania analizy racjonalnej należą teoria pamięci (Anderson, 1990), kategoryzacji (Anderson, 1991a), i inne. W racjonalnej teorii pamięci zakłada się, że celem jest wydobycie informacji w danej chwili potrzebnych, formalny model środowiska oparty jest na modelu systemów

magazynowania i wydobywania informacji takich jak biblioteki, a założenia dotyczące ograniczeń mechanizmu sprowadzają się do nieznanego kosztu wydobycia informacji. Anderson potraktował przeszukiwanie pamięci jako szczególny przypadek racjonalnego, bayesowskiego podejmowania decyzji. Oszacowane przez organizm prawdopodobieństwo, że zapisany w pamięci element będzie potrzebny w danym momencie, ma zależeć od historii użycia i aktualnego kontekstu. Na podstawie tych wszystkich założeń udało mu się wyjaśnić między innymi potęgowy charakter krzywej uczenia się i niektóre efekty rozkładu powtórek.

We wszystkich omawianych przez Andersona przypadkach zastosowania analizy racjonalnej predykcje okazywały się co najmniej tak dobre, jak predykcje alternatywnych modeli obliczeniowych, pomimo że teorie racjonalne pozwalały na wyprowadzenie tych predykcji przy użyciu stosunkowo małej liczby wolnych parametrów, na przykład teoria kategoryzacji wymagała tylko jednego wolnego parametru. Jeszcze w roku 1988 Anderson tak rozumiał rolę analizy racjonalnej:

> Jedną z możliwych reakcji na relatywnie dobre wyniki zastosowania analizy racjonalnej byłyby wzmożone wysiłki stworzenia lepszej architektury poznawczej. (...) Te rezultaty można by więc potraktować raczej jako świadectwa na rzecz odrzucenia trzech rozważanych architektur, a nie jako przykłady zalet analizy racjonalnej. Wydaje mi się jednak, że nie chodzi po prostu o poszukiwanie lepszej architektury. Po pierwsze, nie byłoby to próbą poradzenia sobie z fundamentalnymi problemami identyfikowalności, które nawiedzają nas w trakcie poszukiwania takich mechanizmów mentalnych. Co jeszcze ważniejsze, zignorowalibyśmy w ten sposób zasadniczy wgląd w nieprzypadkowość działania architektur odpowiadających ludzkiemu zachowaniu. [Architektury] Działają w ten sposób, ponieważ to właśnie jest optymalne ze względu na świat, który je otacza.
>
> Inną możliwą reakcją byłoby potraktowanie tych wyników jako zarzutu wobec mechanistycznych teorii umysłu i wezwanie do powrotu do behawioryzmu. (...) Optymalne zachowanie często będzie jednak złożone obliczeniowo, a wyjaśnienia mechanistyczne pozwalają wyrazić tą złożoność, czego nie można powiedzieć o prostych asocjacjach między bodźcami i reakcjami.
>
> Chociaż jestem pewien, że obie wymienione reakcje są błędne, nie mam już takiej pewności co do własnych pozytywnych propozycji, ale oto one: Moje pierwsze przypuszczenie jest takie, że architektury poznawcze powinny być traktowane jako systemy zapisu, służące do wyrażania funkcji behawioralnych, wyłaniających się jako rozwiązania problemów optymalizacji w ramach analizy racjonalnej. Prawdziwa teoria zawiera się w założeniach dotyczących problemu optymalizacji, to jest założeniach dotyczących celów, świata i ograniczeń obliczeniowych. Te założenia nie są związane z ta-

> kimi samymi problemami identyfikowalności co modele mechanistyczne i prowadzą do znacznie głębszych wyjaśnień danych zjawisk. Niemniej, coś takiego jak architektura, posiadająca moc obliczeniową równoważną mocy maszyny Turinga, jest konieczne, jeżeli mamy być w stanie wyrazić rozwiązania tych problemów.
>
> (...) posługując się określonym sformułowaniem problemu, dla każdej architektury można znaleźć pewną konfigurację mechanizmów, umożliwiającą obliczenie optymalnego zachowania. Wybór między architekturami nie ma być zatem oparty na trafności empirycznych predykcji. Powinien się raczej opierać na łatwości, z jaką w danej architekturze można ustalić, jakie jest optymalne zachowanie. Łatwość użycia jest klasycznym kryterium wyboru systemu zapisu. Trafność empiryczna jest zarezerwowana dla teorii.
>
> Moje drugie przypuszczenie (będące wariacją na temat pierwszego) jest takie, że architektury mogą odgrywać pewną rolę na etapie formułowania problemu optymalizacji. Proszę zwrócić uwagę, że krok 3 analizy racjonalnej polega na przyjęciu założeń dotyczących kosztów obliczeniowych. (...) Z tego punktu widzenia, wiele szczegółów, które przypisujemy architekturze, może się sprowadzać do kwestii sposobu zapisu, ale mogą się również pojawiać kluczowe, treściwe założenia. (...) W ramach analizy racjonalnej architektura wyznaczałaby po prostu granice optymalizacji, albo, używając metafory wspinaczkowej Simona, wyznaczałaby kontury powierzchni przeszukiwanej w celu odnalezienia lokalnego optimum.
>
> (s. 21-22, Anderson, 1988/1991c)

Podsumowując, w 1988 roku Anderson twierdził, że w porównaniu do modeli zintegrowanych, a ogólnie modeli mechanistycznych, rezultatem zastosowania analizy racjonalnej bywają teorie procesów poznawczych o większej wartości predykcyjnej, wyższy poziom abstrakcji pozwala uniknąć wielu problemów związanych z identyfikowalnością architektury czy mechanizmu, dostarcza również „głębszego wyjaśnienia" obserwowalnych regularności. Racjonalne wyjaśnienia, jako „bardziej prawdziwe", mogą wręcz czasem zastąpić wyjaśnienia architektoniczne czy mechanistyczne.

Stwierdzenie, że architektury są tylko systemami zapisu wydaje się co najmniej zagadkowe. Założenia architektoniczne mają przecież konsekwencje w postaci (niekoniecznie dokładnych) predykcji dotyczących zachowania. Anderson zdawał się jeszcze wtedy nie rozumieć, że hipotezy obliczeniowe czy mechanistyczne dostarczają odpowiedzi na inne pytania, niż hipotezy wyrażone na poziomie abstrakcji analizy racjonalnej. Ten powód w zupełności wystarcza, aby nie odrzucać hipotez mechanistycznych jako takich. Warto też w tym miejscu zauważyć, że łatwość użycia (prostota) jest zawsze ważnym kryterium oceny wartości poznawczej hipotez empirycznych, a trafność predykcji co najwyżej bywa takim kryterium, szczególnie wobec systemów z natury słabo przewidywalnych.

Nie jest prawdą, że do wyrażenia optymalnego rozwiązania jakiegoś problemu konieczna jest architektura o mocy równoważnej mocy maszyny Turinga. Można czasem dokładnie określić, na czym polega optymalne rozwiązanie pewnych problemów, nawet jeżeli to rozwiązanie nie będzie efektywnie obliczalne. Efektywna obliczalność nie jest kryterium podziału rozwiązań na autentyczne i pozorne, tylko, mówiąc w uproszczeniu, na rozwiązania możliwe i niemożliwe do zrealizowania w skończonym czasie za pomocą mechanicznej procedury (algorytmu). Wiele problemów znacznie prostszych niż te, z którymi nieustannie radzą sobie ludzie, ma to do siebie, że określenie ich optymalnego rozwiązania nie jest zadaniem trywialnym nawet przy założeniu nieograniczonych możliwości obliczeniowych (Huber, 2004).

W komentarzu do stanowiska Andersona Simon (1988/1991) przedstawił kilka ważnych argumentów, które miały pogrążyć program analizy racjonalnej. Jak pisze sam autor:

> W psychologii chodzi zawsze o ograniczoną racjonalność, a badanie ograniczonej racjonalności nie jest badaniem optymalizacji ze względu na środowiska zadaniowe. To jest badanie tego, jak ludzie nabywają strategie radzenia sobie z tymi środowiskami, jak te strategie wyłaniają się z definicji przestrzeni problemowej i jak wbudowane, fizjologiczne ograniczenia kształtują i ograniczają proces nabywania [reprezentacji] przestrzeni problemowych i strategii. Na każdym kroku jest miejsce dla alternatywnych procesów, z których każdy może w stopniu wystarczającym (a nie optymalnym) spełnić wymagania środowiska zadaniowego. Środowisko nie wystarcza do przewidywania, która z tych alternatyw będzie odpowiedzialna za wygenerowanie adaptacyjnego zachowania.
>
> (s. 35, tamże)

Simon omawia problemy związane ze stosowaniem zasady optymalności w naukach ekonomicznych, teorii ewolucji i psychologii. Sens przytaczanych przez niego przykładów można sprowadzić do kilku zastrzeżeń, które przedstawię tutaj w skrócie, bez cytowania odpowiednich fragmentów, jako że wszystkie są w oczywisty sposób prawdziwe:

1. Cel organizmu nie jest najczęściej znany.

2. Proces ewolucji gwarantuje co najwyżej, że rozwiązania będą albo lokalnie lepsze niż rozwiązania konkurencyjne, albo wystarczająco dobre, ale nie gwarantuje optymalności.

3. Rozwiązując to samo zadanie ludzie mogą stosować różne strategie, a to, jakiej strategii użyją, zależy między innymi od ograniczeń systemu poznawczego i wcześniejszej historii uczenia się. Wybór strategii zwykle nie jest całkowicie zdeterminowany przez charakter zadania.

4. Wiedza, jaką dysponuje organizm, jest zawsze częściowa i nie można jej określić analizując tylko środowisko.

5. Żeby zrozumieć, w jaki sposób rozwiązywane jest zadanie, trzeba prawie zawsze założyć coś na temat mechanizmu selekcji uwagowej, mechanizmu przeszukiwania przestrzeni problemu i innych właściwości, decydujących o możliwościach radzenia sobie z wymaganiami zadania.

Simon podkreśla też, że poszukiwanie racjonalnych teorii może łatwo prowadzić do formułowania pozornych, arbitralnych wyjaśnień:

> Skoro zachowanie zostało już zaobserwowane, zwykle łatwo jest znaleźć takie założenia pomocnicze, że zaobserwowane zachowanie będzie optymalne. Takie założenia jest szczególnie łatwo znaleźć, jeżeli nie wymagamy dla nich żadnego bezpośredniego wsparcia empirycznego. Staje się to jeszcze łatwiejsze, jeżeli możemy wprowadzać dodatkowe założenia pomocnicze dla każdego nowego zjawiska, które chcemy wyjaśnić. (...) O ile zbiór wyjaśnianych zjawisk nie jest obszerny i zróżnicowany, liczba stopni swobody wynikająca z przyjmowanych ad hoc założeń może bez trudu przerosnąć liczbę punktów danych, domagających się wyjaśnienia.
>
> (s. 28-29, tamże)

Racjonalne teorie Andersona mają zdaniem Simona zawdzięczać swój sukces predykcyjny przede wszystkim dosyć wątpliwym założeniom pomocniczym. Simon konkluduje, że wbrew temu co twierdzi Anderson, wiedza o środowisku nie wystarcza do przewidywania zachowania.

Atak ze strony Simona był przypuszczalnie jednym z powodów, dla których z czasem stanowisko Andersona uległo zmianie. W publikacji z roku 1991 (Anderson, 1991b) Anderson nadal twierdzi, że analiza racjonalna ma dostarczać częściowego rozwiązania problemu identyfikacji architektury i redukować poziom niepewności związany z predykcjami wyprowadzanymi z hipotez mechanistycznych. Propozycja, aby zastąpić modele mechanistyczne teoriami racjonalnymi zostaje jednak złagodzona na rzecz stanowiska, zgodnie z którym teorie racjonalne są sformułowane na innym, wyższym niż mechanistyczny, to jest komputacyjnym (w terminologii Marra) poziomie analizy. Pomysł, aby traktować procesy poznawcze jako optymalne jest według Andersona hipotezą empiryczną i powinien być oceniany ze względu na to, jak dobrze „organizuje dane" (s. 472, tamże). Program analizy racjonalnej wydaje się w tym ujęciu raczej przedsięwzięciem polegającym na zdaniu się na szczęśliwy zbieg okoliczności, niż metodą systematycznie dostarczającą wartościowych wyników:

> (...) rozwijanie teorii racjonalnej pewnego aspektu ludzkiego poznania to przedsięwzięcie związane zarazem z większym ryzykiem i większymi możliwymi zyskami, niż normalnie stosowane w psychologii poznawczej podejście, polegające na tworzeniu teorii mechanistycznych. Zwiększone ryzyko wynika z tego, że dany aspekt ludzkiego poznania nie musi się wcale okazać optymalny w żadnym interesującym znaczeniu. Obecnie nie ma natomiast powodów aby wątpić, że ludzkie poznanie jest zrealizowane przez pewien zbiór mechanizmów. Podejście racjonalne związane jest z większym możliwym zyskiem, ponieważ byłoby znaczącym odkryciem, gdyby udało się ustalić, że pewne aspekty ludzkiego poznania są optymalne w jakimś interesującym znaczeniu. Byłoby to pouczające, podczas gdy odkrycie, że umysł jest zrealizowany mechanistycznie, nie jest już żadną nowością.
>
> (s. 472, tamże)

Anderson oparł argumentację na rzecz tezy, że procesy poznawcze u ludzi nie muszą przebiegać optymalnie, na przesłankach ewolucyjnych, podobnych do tych które przytoczył Simon (ewolucja jako proces lokalnej optymalizacji, silnie zależny od trudnego do zidentyfikowania kontekstu). W osłabionej wersji założenia programu analizy racjonalnej można streścić w następujących punktach:

- Procesy poznawcze czasami okazują się być w przybliżeniu optymalnie przystosowane do określonych warunków naturalnych.
- Nie wiemy, kiedy procesy poznawcze są optymalne w interesującym znaczeniu, wobec czego ustalenie, kiedy tak jest, byłoby ważnym odkryciem.
- Analiza racjonalna umożliwia, chociaż nie gwarantuje, wykrycie tej optymalności, dzięki uwzględnieniu modelu środowiska, do którego w procesie ewolucji dany organizm się przypuszczalnie zaadaptował i domniemanych celów o rzeczywistym znaczeniu dla organizmu.
- Dzięki temu, że analiza racjonalna dostarcza teorii na wyższym niż mechanistyczny poziomie abstrakcji:
  - dowiadujemy się czegoś na temat regularności obecnych w zachowaniu, które byłoby trudniej przewidzieć opierając się na przesłankach mechanistycznych,
  - uzyskujemy dodatkowe ograniczenia dla teorii mechanistycznej, zwiększając szanse na identyfikację architektury.

Inni zwolennicy stosowania analizy racjonalnej (Chater i Oaksford, 1999, 2000; Chater, Tenenbaum i Yuille, 2006a; Griffiths i in., w druku) często za jej główny cel uznają, nieosiągalne za pomocą teorii mechanistycznych, wyjaśnienie zadziwiającego sukcesu adaptacyjnego ludzi.

Niektóre tezy Simona są ewidentnie fałszywe. Nie jest prawdą, że „w psychologii zawsze chodzi o *ograniczoną* racjonalność". Udzielenie odpowiedzi na wiele ważnych pytań wymaga analizy procesów poznawczych również ze względu na ich, ograniczoną lub nie, ale jednak *racjonalność*. Chcemy nie tylko wiedzieć, jak ludzie rozwiązują określone zadania, ale także na ile, w jaki sposób i dlaczego te rozwiązania odbiegają od teoretycznie osiągalnego ideału. Najczęściej tego ideału nie znamy, co już wystarcza, aby analizę racjonalną uznać za użyteczną. Zgodnie z narzucającą się interpretacją Anderson wcale nie twierdzi, że do przewidywania i wyjaśniania zachowania wystarcza czasami wiedza o środowisku, tylko wiedza o *zadaniu*, które pewne procesy rozwiązują, a to zadanie jest zawsze określone w kategoriach *interakcji* organizmu ze środowiskiem.

Trudno z kolei nie zgodzić się z tym, że ani cele, ani wymiary środowiska, które ukształtowały mechanizm działania procesów poznawczych, najczęściej nie są znane, a teoria i badania ewolucyjne mogą co najwyżej dostarczyć na ten temat obiecujących wskazówek. Nie ma też żadnych zniewalających powodów, aby a priori rozstrzygać, że jakieś rzeczywiste procesy działają optymalnie w interesującym znaczeniu. To jednak nie wystarcza do odrzucenia innego podejścia do analizy racjonalnej.

Opisane w poprzednim rozdziale przykłady teorii racjonalnych nie są oparte na założeniu, że pewne własności środowiska zdecydowały w procesie ewolucji o charakterze rozwiązań. W przeciwieństwie do propozycji Andersona, na podstawowym poziomie, teorie te nie są wcale oparte na żadnym modelu rzeczywistego środowiska, do którego organizm się zaadaptował, tylko na modelu skrajnie wyidealizowanym i nierealistycznym, którego największą zaletą jest to, że pozwala uchwycić niedostrzegalne gołym okiem, istotne wymiary problemów, z którymi ludzie muszą sobie nieustannie radzić.

6.3 Podsumowanie

Wydaje się, że osłabiony program analizy racjonalnej nie musi być sprzeczny z tezą o ograniczonej racjonalności. Brak wyczerpującej wiedzy na temat zadań albo wymagania czasowe interakcji można potraktować jako nieoczywiste, a przez to interesujące właściwości tych zadań. Nie da się wtedy wykluczyć, że stosowane przez ludzi strategie okażą się w przybliżeniu optymalne. Taką racjonalność w warunkach ograniczonych możliwości obliczeniowych próbowali wykazywać między innymi Gigerenzer i Goldstein (1996); Chase, Hertwig i Gigerenzer (1998), badając na przykład szybkie i uproszczone (ang. *fast and frugal*) heurystyki.

Tym, co łączy poglądy Simona, późniejszą wersję analizy racjonalnej Andersona, czy wreszcie propozycje Gigerenzera i Goldsteina jest traktowanie teorii racjonalnych przede

wszystkim jako narzędzia o wartości heurystycznej, od czasu do czasu przydającego się do wyprowadzenia nieoczywistych i stosunkowo dokładnych predykcji. Głównym celem analizy racjonalnej byłaby wtedy identyfikacja architektury. Teorie racjonalne wydają się być przez tych autorów porzucane w momencie, w którym okazują się względnie trafne. Zostaje po nich hipoteza o (zaskakującej) optymalności niektórych procesów, ewentualne nieznane wcześniej obserwowalne regularności i udoskonalona teoria mechanistyczna. Na przykład, w modelu ACT-R mechanizm kategoryzacji i równania opisujące mechanizm zmian sił asocjacji zostały zrewidowane w taki sposób, aby uzgodnić predykcje modelu z wynikami zastosowania analizy racjonalnej. Żeby wyjaśnić, dlaczego moim zdaniem program przypominający analizę racjonalną może mieć dla psychologii poznawczej znaczenie daleko większe niż heurystyczne, w następnym rozdziale pozwolę sobie na kilka uwag dotyczących funkcjonalizmu.

Rozdział 7

Dwa funkcjonalizmy

Przypuszczalnie większość psychologów poznawczych za oczywistą uznaje wartość podejścia interdyscyplinarnego. Trudno znaleźć przykłady praktyk stosowanych przez autorów określających się raczej jako kognitywiści, które byłyby odrzucane jako zbędne przez większą grupę autorów uważających się raczej za psychologów poznawczych. Jako że kognitywistyka jako dziedzina charakteryzowana jest zwykle przez sposób (interdyscyplinarny) badania umysłu, granice między psychologią poznawczą a kognitywistyką wydają się czasem dosyć rozmyte. Wyraźne różnice ujawniają się między innymi w postawie wobec kwestii (meta)teoretycznych podstaw własnej dziedziny.

Za autora terminu „psychologia poznawcza" uznaje się zwykle Ulrica Neissera, który w opublikowanej w 1967 roku książce pod tym samym tytułem tak scharakteryzował przedmiot tej dyscypliny:

> Termin „poznanie" odnosi się do wszelkich *procesów, na skutek których* wejście sensoryczne jest przekształcane, redukowane, opracowywane, magazynowane, odzyskiwane i używane. (...) Ale chociaż psychologia poznawcza zajmuje się wszystkimi formami aktywności człowieka, a nie tylko jakąś ich częścią, robi to z określonej perspektywy. Inne punkty widzenia są równie prawomocne i konieczne. Psychologia Dynamiczna, dla której punktem wyjścia są raczej motywy niż wejście sensoryczne jest dobrym przykładem. Zamiast pytać, w jaki sposób działania i wrażenia człowieka wynikają z tego, co zobaczył, zapamiętał, lub o czym był przekonany, psycholog dynamiczny pyta, jak zależą one od jego celów, potrzeb i instynktów.

Jest rzeczą godną uwagi, że dla Neissera procesy poznawcze nie były jeszcze wtedy wyraźnie związane z celami działania. Jeżeli za autora terminu „psychologia poznawcza" można uznać Neissera, to popularyzację terminu „przetwarzanie informacji" należy prawdopodobnie zawdzięczać Broadbendtowi, który z kolei tak opisuje początkowy okres wyłaniania się tej nowej dyscypliny:

> W tym czasie ci z nas, którzy zajmowali się problemem interakcji człowieka z maszyną mieli problem. Zorientowaliśmy się, że potrzebujemy szczególnego rodzaju psychologii aby nadać sens naszym wynikom. Czasopisma akademickie były jednak zdominowane przez inny sposób myślenia. Jak wykazał Paul Fitts, można było uzyskać uwagę odbiorców rozpoczynając artykuł w ogólnie przyjętym języku bodźca i reakcji (S-R), by potem powoli przejść do bardziej użytecznego języka przetwarzania informacji. (...) jednakże z czasem stawało się to coraz trudniejsze, ponieważ oznaczało wykonywanie całej roboty w każdym artykule; potrzebowaliśmy jakiegoś ogólnego stanowiska, na którym można by się było oprzeć.
>
> (s. 134, Broadbent, 1984)

Od wymienionych publikacji minęło wiele lat, jednak nadal nie bardzo wiadomo, na czym właściwie polega specyfika podejścia poznawczego. Bez trudu można odróżnić zwolenników tego podejścia od zwolenników niektórych wersji behawioryzmu, wciąż jednak na pytanie o najważniejszy rys charakterystyczny pada mniej więcej odpowiedź w rodzaju „badanie umysłu", „procesów poznawczych", „procesów przetwarzania informacji", albo „procesów i stanów poznawczych scharakteryzowane przez ich funkcję" (Smith, 2002), przy czym samo pojęcie funkcji pozostaje wieloznaczne (Krzyżewski, 1989; Smith, 2002).

W najogólniejszym znaczeniu funkcja oznacza rolę, jaką proces lub stan pełni w pewnym większym systemie. To, że procesy lub stany złożonego systemu muszą być scharakteryzowane częściowo przez ich związek z innymi procesami lub stanami jest oczywiste i nie mówi zbyt wiele o specyfice teorii poznawczych. Z powodów, które wkrótce staną się jasne, stanowisko najlepiej odpowiadające moim zdaniem praktyce badawczej w psychologii poznawczej będę określał jako „mechanistyczne", a stanowisko alternatywne będę nazywał „racjonalnym", chociaż być może na początek bardziej adekwatne byłoby jakieś słabsze określenie, na przykład „teleologiczne".

Od razu pragnę zaznaczyć, że moja znajomość bogatej rodziny poglądów filozoficznych, nazywanych takimi lub innymi funkcjonalizmami jest skromna. Odniosę się tutaj jedynie do stanowiska określanego czasem jako „funkcjonalizm obliczeniowy", albo „teoria identyczności stanów funkcjonalnych" (Piccinini, 2004). Głównym celem będzie nie tyle wyczerpująca analiza założeń, na których stanowisko to jest oparte, ile raczej zwrócenie uwagi Czytelnika na sposób myślenia, który znajduje w nim wyraz.

7.1 Funkcjonalizm obliczeniowy

Niektórzy filozofowie zajmujący się problematyką umysłu, tacy jak Putnam (1960, 1967/1980), Fodor (1965), Block (1980, 1996/2006) i Pylyshyn (1989) podjęli próby dokładniejszej

analizy kwestii, jakiego rodzaju teoriami są, przynajmniej niektóre, teorie umysłu. Teorie te mają być „funkcjonalistyczne":

> ... roboczą hipotezą znacznej części kognitywistyki jest [założenie], że umysł jest dosłownie rodzajem komputera. Można by zapytać, ze względu na jaką właściwość maszyna Turinga osiąga uniwersalność albo programowalność, która uzasadnia traktowanie jej jako modelu inteligencji?
>
> (s. 5-6, Pylyshyn, 1989)

> Funkcjonalizm jest jednym z głównych osiągnięć teoretycznych dwudziestowiecznej filozofii analitycznej i dostarcza podstaw pojęciowych dla wielu badań w kognitywistyce.
>
> (Block, 1996/2006)

Nie jest dla mnie do końca jasne, w jakim stopniu stanowisko, które wkrótce opiszę nieco dokładniej, można uznać zdaniem tych autorów za poprawną rekonstrukcję podstaw teoretycznych w psychologii poznawczej jako takiej zamiast w kognitywistyce. Przynajmniej w kilku miejscach Putnam, Block, Pylyshyn i Fodor zdają się twierdzić, że chodzi im o kognitywistykę jako dziedzinę zawierającą w sobie psychologię poznawczą, czasem też wyraźnie twierdzą, że mają na myśli psychologię poznawczą, ale nie jest łatwo znaleźć fragmenty, w których powoływaliby się na jakieś konkretne teorie psychologiczne, demonstrując przy tym dokładnie, na czym polega ich funkcjonalistyczny (w zaproponowanym rozumieniu) charakter. O ile ten stan rzeczy jest jeszcze zrozumiały w przypadku funkcjonalizmu rozumianego jako rekonstrukcja podstaw psychologii potocznej, o tyle budzi już poważniejsze wątpliwości w przypadku funkcjonalizmu, który miałby być rekonstrukcją podstaw psychologii naukowej. Najczęściej w tym kontekście przytaczanym, a więc być może zdaniem autorów reprezentatywnym przykładem teorii psychologicznej jest bliżej nieokreślona psychologiczna teoria stanu czucia bólu.

Funkcjonalizm obliczeniowy opiera się między innymi na trzech, nietrywialnych i rzadko jawnie rozróżnianych założeniach. Rozdzielenie tych założeń zawdzięczam znakomitej pracy Piccininiego (2004). Zgodnie z pierwszym założeniem, stany i procesy[1] mentalne powinny być scharakteryzowane przez ich funkcję. Zgodnie z drugim założeniem, ta funkcjonalna charakterystyka polega na określeniu związków między abstrakcyjnie rozumianymi procesami, stanami, wejściami i wyjściami systemu bądź organizmu. Zgodnie z trzecim założeniem, te związki mają mieć charakter przyczynowo-skutkowy.

[1]Co ciekawe, w cytowanych tutaj, klasycznych pracach dotyczących funkcjonalizmu, Putnam i Fodor nieodmiennie snują rozważania na temat stanów mentalnych, procesom poświęcając już znacznie mniej miejsca. Jeżeli funkcjonalizm obliczeniowy ma być rekonstrukcją podstaw teoretycznych psychologii czy kognitywistyki, musi oczywiście dotyczyć zarówno stanów jak i procesów, dlatego będę dalej pisał o jednych i drugich.

Charakterystyka procesów i stanów mentalnych polegałaby więc na określeniu ich funkcji, rozumianej jako sieć zależności przyczynowo-skutkowych, łączących te stany i procesy z innymi stanami lub procesami, a także wejściami i wyjściami systemu. Tak mniej więcej wygląda funkcjonalizm w ujęciu (wczesnego) Putnama, Fodora i Pylyshyna. Teoria oparta na wymienionych metateoretycznych założeniach byłaby teorią procesów i stanów mentalnych jako procesów i stanów obliczeniowych.

Stanowisko funkcjonalizmu obliczeniowego zostało sformułowane po raz pierwszy przez McCulloha i Pittsa (1943), jednak najczęściej autorstwo przypisuje się Putnamowi (1960), który w latach 60-tych o pomysłach McCulloha i Pittsa prawdopodobnie nie wiedział (Piccinini, 2004). Propozycja Putnama opiera się na analogii między umysłem i abstrakcyjnymi automatami, zdolnymi do „obliczania" dowolnej efektywnie obliczalnej funkcji, czyli tak zwanymi maszynami Turinga. Właśnie ta elastyczność maszyn Turinga sprawia, że wydają się one mieć coś wspólnego z umysłem i zachowaniem, ponieważ umysł i zachowanie na pierwszy rzut oka sprawiają wrażenie niezwykle elastycznych.

Putnam szybko przestał korzystać z pojęcia maszyny Turinga i zaczął używać mniej zobowiązującego terminu „automat probabilistyczny" (Putnam, 1967/1980). O ile dobrze rozumiem Putnama, automat probabilistyczny to dowolny system dający się opisać w kategoriach stanów, wartości na wejściach i wyjściach i warunkowego rozkładu prawdopodobieństwa, określonego na przyszłych stanach i wyjściach ze względu na aktualny stan i wejście[2].

O dowolnym, scharakteryzowanym fizycznie lub biologicznie systemie można powiedzieć, że jest realizacją określonego automatu probabilistycznego wtedy i tylko wtedy, gdy da się zidentyfikować fizyczne bądź biologiczne stany tego systemu takie, że dynamika następstw tych stanów jest taka jak tego automatu probabilistycznego. Stany fizyczne lub biologiczne mogą być przy tym scharakteryzowane na dowolnym poziomie abstrakcji. Taka identyfikacja pozwala mówić o stanach obliczeniowych danego systemu ze względu na opis w kategoriach pewnego abstrakcyjnego automatu. Hipoteza stanu funkcjonalnego głosi, że stany mentalne mają być utożsamione ze stanami obliczeniowymi, zgodnie z najlepszą dostępną teorią psychologiczną (Putnam, 1967/1980). Teorie psychologiczne mają wyjaśniać zachowanie w kategoriach tak rozumianych stanów obliczeniowych i charakteryzować procesy psychiczne jako procesy obliczeniowe.

Być może najważniejszą zaletą funkcjonalizmu obliczeniowego jest możliwość wyminięcia wielu problemów filozoficznych, związanych z relacją ciało-umysł, co uzyskuje się dzięki abstrahowaniu od substancji. Rzeczywiście istniejący system jest albo nie jest realizacją określonego automatu probabilistycznego niezależnie od tego, z jakiej substancji, na przykład mentalnej czy fizycznej, ten system jest „zrobiony". Putnam uważał również, że dopiero taka obliczeniowa teoria psychologiczna, w przeciwieństwie do teorii opartych na utożsamieniu stanów mentalnych ze stanami mózgu, może dostarczyć

[2]Putnam ujmuje to nieco swobodniej, opisując automat probabilistyczny jako coś przypominającego maszynę Turinga, tyle że działającego niedeterministycznie (Putnam, 1967/1980).

praw i definicji stanów i procesów mentalnych obowiązujących niezależnie od gatunku organizmu.

Od czasu pojawienia się tej koncepcji wielu filozofów kwestionowało jej założenia a funkcjonalizm przyjął liczne nowe postaci. W filozofii współczesnej istnieje cała rodzina mniej lub bardziej pokrewnych funkcjonalizmów, na przykład funkcjonalizm analityczny Lewisa (Lewis, 1980), homuncularny Dennetta (Dennett, 1978), i inne. Te nowsze wersje zostały zaproponowane między innymi jako próby odpowiedzi na kontrprzykłady dla tez Putnama, takie jak „chiński mózg" Blocka (Block, 1980), „chiński pokój" Searla (Searle, 1980), „odwrócone spektrum" (Block i Fodor, 1972), czy argument z bliźniaczej ziemi, którego autorem był sam Putnam (Putnam, 1973), chociaż sformułował go w związku z innym, podjętym przez siebie problemem. Nie widzę powodu, aby podejmować w tym miejscu próby analizy zawiłego rozwoju tej idei, tym bardziej że brakuje mi do tego kompetencji. Muszę jednak zwrócić uwagę Czytelnika na kilka fragmentów pochodzących z „The Nature of Mental States" (Putnam, 1967/1980):

> Hipotezę, że „czucie bólu jest funkcjonalnym stanem organizmu" można wyrazić dokładniej w następujący sposób:
>
> 1. Wszystkie organizmy zdolne do czucia bólu są Automatami Probabilistycznymi.
> 2. Każdy organizm zdolny do czucia bólu posiada przynajmniej jeden Opis *pewnego rodzaju* (to jest, bycie zdolnym do czucia bólu jest tym samym, co posiadanie *odpowiedniej* Organizacji Funkcjonalnej).
> 3. ...
>
> (s. 227, tamże)

Podkreślenia we wszystkich cytatach z Putnama pochodzą ode mnie. I dalej:

> ... stan funkcjonalny o jaki nam chodzi [stan czucia bólu] jest stanem polegającym na odbieraniu na wejściach sensorycznych wartości, które *odgrywają pewną rolę* w Organizacji Funkcjonalnej. Ta rola jest określona przynajmniej częściowo przez to, że organy zmysłowe odpowiedzialne za rozważane wejścia są organami, *których funkcją jest wykrywanie uszkodzenia ciała, albo niebezpiecznie skrajnych temperatur, ciśnienia, etc., i przez to, że same te „wejścia" ... reprezentują warunki, którym organizm przypisuje niską użyteczność.*
>
> (s. 229, tamże)

> (...) W rzeczy samej, badanie tej hipotezy oznacza podjęcie prób stworzenia „mechanicznych" modeli organizmów - a czy w pewnym sensie nie o to właśnie chodzi w psychologii?

(s. 227, tamże)

Proszę zwrócić uwagę, jak ciężką pracę wykonuje w tych fragmentach wyrażenie „odpowiednia Organizacja Funkcjonalna".

Putnam sam zauważa, że zawarte w pierwszym z cytowanych fragmentów twierdzenie (1) jest „puste, ponieważ wszystko można potraktować jako automat probabilistyczny ze względu na jakiś opis" (s. 227, tamże). To oznacza jednak, że bez dodatkowego dookreślenia kryteriów odpowiedniości organizacji funkcjonalnej nie dowiadujemy się nic na temat teorii psychologicznych poza tym, że mogą, choć nie muszą, abstrahować od substancji mentalnej i że wyjaśniają zjawiska należące do dziedziny psychologii w kategoriach częściowo nieobserwowalnych mechanizmów przyczynowo-skutkowych. Kiedy jednak kryteria odpowiedniości relacji zostają przez Putnama w zarysie dookreślone, tak jak ma to miejsce w ostatnim fragmencie, są to kryteria funkcjonalne w znaczeniu *celów*, jakim służą wejścia sensoryczne i teleologicznie, a nie przyczynowo-skutkowo określonej roli, jaką stany i procesy mentalne odgrywają w procesie realizacji tych celów. W tej sytuacji nie można się najwyraźniej obejść bez pojęcia „przypisywania przez organizm użyteczności". To samo pęknięcie pojawia się w przytoczonym fragmencie pochodzącym od Neissera, kiedy pisze „Termin „poznanie" odnosi się do wszelkich *procesów, na skutek których* wejście sensoryczne jest przekształcane, redukowane, opracowywane, magazynowane, odzyskiwane i używane." Przekształcanie, redukowanie, zapisywanie, przechowywanie i odtwarzanie wejścia sensorycznego, percypowanie, zwracanie lub odwracanie uwagi, myślenie, wyobrażanie sobie, zgadywanie, przypuszczanie, przewidywanie, podejmowanie decyzji to wszystko są funkcje rozumiane jako rozwiązania pewnych zadań, a *mechanizmy, które te rozwiązania realizują*, to coś zupełnie innego.

Opis w kategoriach zadania i jego rozwiązania jest czymś innym, niż opis w kategoriach mechanicznego, to jest przyczynowo-skutkowego następstwa stanów, wejść i wyjść. Można to uzasadnić powołując się na ten sam argument, którego w innym celu użył Putnam (1967/1980). System scharakteryzowany racjonalnie, albo używając słabszego określenia, teleologicznie, może być zrealizowany przez nieskończenie wiele różnych automatów probabilistycznych. Putnam zauważył, że z tezy o wielorakiej realizowalności wynikają kłopoty dla zwolenników tezy o identyczności stanów mózgowych i stanów mentalnych. Zdaniem tego autora jest znacznie mniej prawdopodobne, aby udało się kiedykolwiek ustalić biologiczne właściwości mózgu takie, że wszystkie przypadki wystąpienia jakiegoś stanu mentalnego będą ściśle skorelowane z pewnym biologicznie wyróżnionym stanem mózgowym, niż że uda się odkryć skorelowane ze stanami mentalnymi własności, wyrażone w kategoriach bardziej abstrakcyjnych stanów obliczeniowych. Jeżeli jednak stany obliczeniowe utożsamiane są ze stanami mentalnymi na mocy zaobserwowanych korelacji między tymi pierwszymi a tymi drugimi, muszą istnieć jakieś niezależne od teorii obliczeniowej kryteria identyfikacji stanów mentalnych!

Wystarczy w argumencie przeciwko tezie o identyczności stanów mentalnych i stanów mózgowych zamienić własności mózgowe na własności obliczeniowe, a własno-

ści obliczeniowe na własności komputacyjne lub racjonalne, jednocześnie usuwając kłopotliwy wymóg stwierdzenia „ścisłej korelacji". Argument przyjąłby wtedy postać - jest znacznie mniej prawdopodobne, że uda się kiedykolwiek znaleźć własności obliczeniowe takie, że wszystkie wystąpienia jakiegoś stanu mentalnego będą ściśle skorelowane z tymi własnościami, niż że uda się znaleźć takie własności komputacyjne, że wystąpienie tego stanu mentalnego będzie pociągało za sobą wystąpienie tych własności na mocy definicji tego stanu mentalnego jako rozwiązania określonego zadania. Uzyskujemy w ten sposób argument przeciwko stanowisku funkcjonalnej identyczności stanów i można w końcu dostrzec, dlaczego konieczne było odwoływanie się do „odpowiedniej" organizacji funkcjonalnej. Putnam wybrał niewłaściwy poziom abstrakcji. Jak można zauważyć przyglądając się teoriom racjonalnym, nawet optymalne, a więc w jakimś sensie unikalne rozwiązanie, można scharakteryzować abstrahując od mechanizmu obliczeniowego.

W późnych latach 80-tych Putnam ostatecznie odrzucił funkcjonalizm obliczeniowy. Podobnie jak Fodor, doszedł do wniosku, że model obliczeniowy jest zawieszony w próżni i przyjął stanowisko, zgodnie z którym stany mentalne jako stany mentalne muszą być częściowo zdeterminowane zewnętrznie, ponieważ stwierdził, że inaczej nie sposób poradzić sobie z zagadnieniem reprezentacji i znaczenia (Putnam, 1988, 1999; Fodor i Lepore, 1992). Sformułował też (Putnam, 1988) przeciwko funkcjonalizmowi obliczeniowemu kilka argumentów o bardziej technicznym charakterze, między innymi argument z trywialności opisu obliczeniowego (każdy opis obliczeniowy można zastosować do dowolnego systemu fizycznego), z wielorakiej obliczeniowej realizowalności dowolnego stanu mentalnego i inne, argumenty te wydają się jednak z różnych względów problematyczne (Buechner, 2007). W warstwie konstruktywnej nowszych propozycji Putnama i Fodora środowisko i interakcja z nim są jednak moim zdaniem uwzględnione w stopniu niewystarczającym, w każdym razie niewystarczającym jak na potrzeby psychologii poznawczej. Nadal brakuje jawnego i dostatecznie konsekwentnego uwzględnienia celowości, a zwłaszcza wyuczalności ze względu na cel i prób formułowania teorii środowiska.

Rozróżnienie na warstwę mechanistyczną i racjonalną teorii psychologicznych zostało bardziej docenione przez innych autorów. Niewykluczone, że to Marr (1982) pierwszy zauważył, że aby zrozumieć „system przetwarzający informacje" trzeba przeprowadzić analizę na trzech różnych poziomach, które nazwał komputacyjnym, algorytmicznym i implementacyjnym. Na poziomie komputacyjnym analiza dotyczy tego, co system robi, to znaczy jakie problemy rozwiązuje i tego, dlaczego (ze względu na te problemy) robi to co robi. Analiza w kategoriach mechanizmu obliczeniowego odpowiada poziomowi algorytmicznemu, wreszcie poziom implementacyjny odpowiada fizycznej lub biologicznej realizacji. Terminologia jest tu niestety dość kłopotliwa. Zwolennicy probabilistycznych modeli poznania poziom opisu w kategoriach zadania i jego rozwiązania nazywają obliczeniowym (ang. *computational*), a poziom algorytmiczny w terminologii Marra, który funkcjonaliści obliczeniowi nazywają obliczeniowym, określają jako mechanistyczny. Ponieważ nie jest wcale oczywiste (Piccinini, 2004), że charakterystyka

funkcjonalna to charakterystyka obliczeniowa, mechanistyczny jest również każdy opis implementacyjny, a potrzebny jest mi w tej pracy termin na określenie teorii wyrażonych na tym samym poziomie analizy co teorie racjonalne, ale nie będących teoriami racjonalnymi, zdecydowałem się na dalekie od doskonałości rozwiązanie, aby mówić o poziomach komputacyjnym (czasem, zależnie od kontekstu, nazywanym też poziomem teleologicznym albo racjonalnym), obliczeniowym albo mechanistycznym (czyli algorytmicznym w terminologii Marra) i implementacyjnym.

Jeszcze więcej na temat trzech wymienionych poziomów można się dowiedzieć od Pylyshyna (1989):

> ... [W ramach Klasycznego Stanowiska, zgodnie z którym umysł jest rodzajem komputera] zakłada się, że zarówno komputery jak i umysły mają przynajmniej trzy odrębne poziomy organizacji:
>
> 1. *Poziom semantyczny*, albo *poziom wiedzy*. Na tym poziomie wyjaśniamy, dlaczego ludzie lub odpowiednio zaprogramowane komputery robią pewne rzeczy, mówiąc co wiedzą i jakie są ich cele i demonstrując, że są one [te cele, wiedza i działania] powiązane w sensowny, lub nawet racjonalny sposób.
> 2. *Poziom symbolu*. Zakłada się, że semantyczna zawartość wiedzy i celów jest zakodowana w wyrażeniach symbolicznych. Takie ustrukturyzowane wyrażenia mają części, z których każda koduje jakąś semantyczną zawartość. Kody i ich struktura, jak również reguły według których są manipulowane, to osobny poziom organizacji systemu.
> 3. *Poziom fizjologiczny* (lub *biologiczny*). Aby cały ten system mógł działać, musi być zrealizowany w fizycznej postaci. Struktura i zasady zgodnie z którymi działa obiekt fizyczny odpowiadają poziomowi fizycznemu lub biologicznemu.
>
> (s. 8, tamże)

Pylyshyn podaje przykład, jak taka wielopoziomowa analiza może wyglądać:

> Przypuśćmy, że masz kalkulator z przyciskiem pierwiastka kwadratowego. Jeżeli chcesz wyjaśnić, dlaczego udziela dziwnych odpowiedzi, lub przestaje działać gdy baterie są na wyczerpaniu, albo kiedy odetniesz jedno z jego połączeń, albo gdy temperatura jest niska, musisz odwołać się do fizycznych właściwości kalkulatora, do poziomu fizycznego. Jeżeli chcesz wyjaśnić, dlaczego pojawiają się pewne błędy wynikające z zaokrąglania w niższych cyfrach odpowiedzi, albo dlaczego obliczenie rozwiązań pewnych problemów zajmuje więcej czasu niż innych, musisz odwołać się do tego,

> jak liczby są zakodowane symbolicznie i do tego, jaka konkretnie sekwencja przekształceń tych symbolicznych wyrażeń zachodzi (to znaczy, do używanego algorytmu). To jest wyjaśnienie na poziomie symbolu. Ale z drugiej strony, gdy chcesz wykazać, że algorytm zawsze udzieli poprawną odpowiedź, musisz się odnieść do faktów i twierdzeń teorii liczb; to jest, do semantyki symboli.
>
> (s. 8-9, tamże)

Muszę powiedzieć, że nawet teraz, kiedy piszę te słowa, oparcie się przekonującej sile tego przykładu nie przychodzi mi łatwo. Nie mogę w tym miejscu pominąć jeszcze jednego cytatu:

> Jeżeli opis na poziomie wiedzy jest poprawny, to musimy wyjaśnić jak to jest możliwe, aby system fizyczny taki jak człowiek zachowywał się w sposób odpowiadający zasadom na poziomie wiedzy, jednocześnie będąc ograniczonym przez prawa fizyczne. Zawartość wiedzy jest powiązana ze stanem systemu przez relację *semantyczną*, która jest całkiem inna od tych pojawiających się w prawach naturalnych (na przykład, przedmiot tej relacji nie musi istnieć [chodzi o terminy posiadające sens, których desygnat nie istnieje, na przykład „jednorożec"]). Obecnie istnieje tylko jedno obiecujące wyjaśnienie tego, jak zasady na poziomie wiedzy mogą być zrealizowane fizycznie i jest to wyjaśnienie oparte na ideach pochodzących od Boole'a, Hilberta, Turinga, Fregego i innych logików. Mówi ono, że wiedza jest *zakodowana* przez system kodów symbolicznych, które z kolei są zrealizowane fizycznie i że to właśnie fizyczne właściwości kodów powodują dane zachowanie.
>
> (s. 13, tamże)

Inaczej niż w cytowanych pismach Putnama, rozróżnienie na mechanizm obliczeniowy i zadanie, jakie ten mechanizm realizuje jest tu wyraźne. Co więcej, Pylyshyn pisze, że wyjaśnienie działania umysłu musi ostatecznie zawierać wszystkie trzy warstwy, z czym trudno się nie zgodzić. Propozycja Pylyshyna jest jednak problematyczna jako teoria teorii umysłu, w szczególności jako metateoretyczna charakterystyka psychologicznej teorii umysłu i zachowania. Zacznę od spraw mniej zasadniczych.

Trudno w teoriach psychologicznych odnaleźć ślady idei sformułowanych przez Boole'a, Hilberta, Turinga czy Fregego. Nie jest tak dlatego, że psychologowie raczej nie przytaczają tych autorów z nazwiska, chociaż w istocie niejawnie korzystają z ich ważnych dokonań, tylko dlatego, że bezpośrednio z tych dokonań nie korzystają. Nazwiska Boole'a, Hilberta i Fregego łączą podejmowane przez nich próby sformułowania logicznych podstaw matematyki. Turing pasuje do tego towarzystwa, ponieważ posługując się

pojęciem abstrakcyjnej maszyny udało mu się wykazać, że cele programu formalizacji podstaw matematyki nie mogą być zrealizowane. Nieco łatwiej byłoby dostrzec związki z psychologią, gdyby ograniczyć się jedynie do Boole'a, Fregego (być może Pylyshyn miał na myśli „prawa myślenia" jako coś, co ma być oparte na prawach logicznych lub matematycznych) i Turinga (przypuszczalnie chodzi przede wszystkim o test Turinga). Nie jest wcale łatwo przejść od Turinga do komputerów takich jak te, które stanowią niewyczerpane źródło inspiracji dla wielu teorii w psychologii poznawczej, ponieważ komputery te powstały prawdopodobnie nie dzięki dokonaniom Turinga, ale w znacznej mierze niezależnie od nich, jak przekonująco argumentuje Sloman (2002). Jeszcze trudniej jest przejść od komputerów do zawierających wszystkie trzy wymienione przez Pylyshyna warstwy teorii psychologicznych, o czym wkrótce.

Pewne wątpliwości budzi przywiązanie Pylyshyna, które zresztą dzieli z Fodorem, do reprezentacji symbolicznych i w ogóle do pewnego rodzaju architektur. Niejeden zwolennik koneksjonizmu oceniłby te pomysły jako zbyt oderwane od elementarnej wiedzy o mózgu, z czego Pylyshyn zdaje sobie doskonale sprawę i w kilku miejscach odnosi się do samego Rumelharta, nie omieszkując przy tym wspomnieć (w nawiasie) o wolnych parametrach. Twierdzi jednak, że hipoteza umysłu jako symbolicznego systemu obliczeniowego jest najlepszą jaką na razie dysponujemy, a przecież w tej sytuacji „racjonalną strategią jest kontynuowanie w oparciu o klasyczne założenie, aż nie pojawi się jakaś lepsza alternatywa. Przynajmniej taka właśnie strategia jest stosowana w każdej dojrzałej nauce" (s. 14, tamże).

Pod wieloma względami umysł różni się jednak od kalkulatora. W przypadku kalkulatora jego podstawowa funkcja jest doskonale znana, a w przypadku umysłu nie. W przypadku kalkulatora przestrzeń hipotez dotyczących (obserwowalnego) mechanizmu obliczeniowego jest bardzo przyjazna, a w przypadku umysłu nie.

Problemy związane z trzema warstwami teorii psychologicznej nie mają jednakowej wagi na każdym etapie procesu badawczego. Dopóki funkcja nie jest zbyt dobrze poznana, niewiele można powiedzieć na temat mechanizmów obliczeniowych, ani ich związku z poziomem implementacji. Kiedy funkcja nie jest znana można się dowolnie długo przyglądać (w tej sytuacji zidentyfikowanemu arbitralnie) mechanizmowi obliczeniowemu lub jego implementacji i nic ciekawego z tego nie wyniknie. Analiza przebiega przede wszystkim w kierunku od funkcji do mechanizmu, chociaż czasami *przyjęte wcześniej* założenia na temat funkcji mogą być zrewidowane na podstawie analizy mechanizmu obliczeniowego lub implementacji. Nawet kiedy coś więcej wiadomo na temat funkcji, jej status nadal jest uprzywilejowany, dlatego że ostatecznie mechanizm obliczeniowy i jego związki z działaniem mózgu interesują psychologów tylko o tyle, o ile mają związek z zachowaniem, a zachowanie jest tym czym jest, to znaczy właśnie zachowaniem, ze względu na cel jaki realizuje.

Układ nerwowy jest jak wiadomo systemem niezwykle złożonym i rozgrywają się w nim liczne fascynujące procesy, ale tylko znikoma część tych procesów może mieć jakiś uchwytny związek z mechanizmami obliczeniowymi, będącymi realizacjami stanów

i procesów mentalnych. Nie może być inaczej, opis w kategoriach mechanizmu obliczeniowego jest bowiem abstrakcyjnym opisem działania mózgu jako systemu biologicznego albo fizycznego. Mechanizmy obliczeniowe z kolei różnią się pod względem tego, na ile decydują o przebiegu obserwowalnego zachowania. Psychologowie poznawczy zdają sobie doskonale sprawę, że nie wszystkie mechanizmy obliczeniowe są jednakowo ważne, inaczej nie używaliby określenia „fenomen laboratoryjny" w pejoratywnym znaczeniu. Nie można być nigdy do końca pewnym, czy jakiś fenomen laboratoryjny nie okaże się kiedyś całkiem interesujący, ale nie wynika stąd, że można skutecznie uprawiać psychologię poznawczą ani jakąkolwiek inną nie stosując uzasadnionych kryteriów podziału zagadnień na ważne i mniej ważne. Takie kryteria opierają się głównie na przypuszczeniach dotyczących funkcji jako rozwiązania zadań związanych ostatecznie z celowym zachowaniem.

Trafność analogii komputerowej staje się jeszcze bardziej wątpliwa, gdy tylko przyjrzymy się bliżej środowisku w jakim działa komputer. Komputer jako obiekt fizyczny działa w tym samym środowisku co człowiek, ale komputer jako mechanizm obliczeniowy działa w środowisku złożonym jedynie z wejść i wyjść. To, że w ogóle można zrozumieć działanie takiego mechanizmu redukując środowisko do wejść i wyjść oznacza, że musi istnieć jakaś ważna różnica między komputerem i umysłem, albo działaniem komputera i zachowaniem się człowieka. Z perspektywy psychologicznej zachowanie można zrozumieć tylko jako proces celowej interakcji ze środowiskiem. Procesy poznawcze są tym czym są, to znaczy procesami poznawczymi, ze względu na pojawiające się w ramach tej interakcji zadania, które są przez te procesy (częściowo, w mniejszym lub większym stopniu) rozwiązywane.

Dopóki teorie poznawcze będą silnie inspirowane metaforą komputerową, te brakujące elementy nie będą mogły się dostatecznie wyraźnie ujawnić. Istnieje bogaty słownik pojęć psychologicznych dotyczących procesów i stanów mentalnych i korzystając z tych pojęć sformułowano wiele użytecznych teorii, ale nie istnieje nawet w przybliżeniu tak bogaty słownik dotyczący środowiska *jako przedmiotu zainteresowania psychologii*. Do tego „przeoczenia" wrócę jeszcze w dalszej części pracy.

Kalkulator wydaje się liczyć w zbliżonym znaczeniu do tego, jakiego używamy gdy mówimy, że człowiek liczy, z powodów częściowo związanych z przeoczeniem kwestii środowiska. Niemniej ani kalkulator, ani komputer nie liczy w znaczeniu psychologicznym, tylko co najwyżej *służy do liczenia*. Typowe komputery osobiste nie robią nic w tym samym znaczeniu co ludzie, ponieważ te komputery *nie działają celowo*. Nie jestem pewien, czy można powiedzieć, że mózg działa celowo, ale tą kwestię zostawię badaczom o większym zacięciu filozoficznym.

Dostrzegalne choć odległe podobieństwa między komputerem a umysłem biorą się stąd, że komputery zostały zaprojektowane po to, żeby służyły ludziom do realizowania niektórych *ich* celów. Elastyczność komputerów wynikająca z ich programowalności jest tylko elastycznością wynikającą z programowalności. Elastyczność to zdecydowanie za mało, żeby uznać, że komputery jako takie są zdolne do zachowania celowego. Kiedy

już raz zacznie się konsekwentnie myśleć o zachowaniu jako o działaniu celowym, rozgrywającym się w interakcji organizmu ze środowiskiem, o procesach poznawczych jako o rozwiązaniach zadań wyznaczone przez wymagania tej interakcji, a o mechanizmach działania tych procesów jako o mechanizmach działania tych procesów, cały obraz zacznie ulegać radykalnej zmianie.

7.2 Funkcjonalizm z przełomu wieków

Zarzut niedostatecznego uwzględnienia szeroko rozumianej celowości w psychologii został wyraźnie wyartykułowany już pod koniec dziewiętnastego wieku przez przedstawicieli zupełnie innego funkcjonalizmu. Pod wodzą Williama Jamesa autorzy tacy jak Angell, Moore, Mead i Dewey przypuścili atak na założenia i metody strukturalizmu[3]. Zarzucali akcentowanie struktury świadomości i procesów mentalnych zamiast ich funkcji, rozumianej przede wszystkim jako rola, jaką te procesy odgrywają w codziennej adaptacji organizmu do naturalnych warunków. Zwracali uwagę na wady metody introspekcyjnej i inne problemy natury metodologicznej, dostrzegli także ewidentną niewydolność aplikacyjną strukturalizmu, wynikającą z ignorowania teleologicznie rozumianej funkcji procesów psychicznych i roli środowiska.

Funkcjonalizm z przełomu dziewiętnastego i dwudziestego wieku ostatecznie zniknął z mapy psychologii, a miejsce obowiązującej doktryny na długie lata zajął behawioryzm. Lektura pism tych autorów po przeszło stu latach brzmi jednak zadziwiająco aktualnie. W tekście, który można uznać za manifest funkcjonalistów, Angell (1907) przedstawił, sprowadzające się jego zdaniem w istocie do jednego spójnego stanowiska, możliwe rozumienia funkcjonalizmu i przemawiające za nim argumenty. W przeciwieństwie do strukturalizmu celem psychologii funkcjonalnej miały być:

> ... starania, aby rozróżnić i scharakteryzować typowe *operacje* świadomości w naturalnych warunkach, w przeciwieństwie do prób analizy i opisu jej elementarnych i złożonych części składowych. Strukturalna psychologia percepcji stara się na przykład ustalić liczbę i charakter różnych nieanalizowalnych jakości zmysłowych, takich jak rodzaje koloru, tonu, smaku, i tak dalej. Funkcjonalna psychologia percepcji jako właściwy obszar swoich zainteresowań uznałaby z kolei charakter różnych aktywności percepcyjnych ze względu na to, jak różnią się one w swoim *modus operandi*, zarówno między sobą jak i od innych procesów mentalnych, takich jak ocenianie, wyobrażanie sobie, dążenie do czegoś i innych.
>
> (s. 62-63, tamże)

[3]Bogaty zbiór tekstów funkcjonalistów z przełomu wieków znajduje się pod adresem *http* : *//psychclassics.yorku.ca*.

„Operacje" świadomości i „modus operandi" aktywności percepcyjnych można jak się wydaje rozumieć teleologicznie lub mechanistycznie, ale już w innych fragmentach da się zauważyć znacznie silniejszy akcent na cel, jakiemu te procesy mają służyć. Przytoczony wyżej zarzut staje się trudny do odróżnienia od jednego z tych, który sformułowałem w tej pracy, gdy tylko zamienić „strukturę świadomości i procesów mentalnych" na „przyczynowo-skutkowe mechanizmy obliczeniowe i strukturę reprezentacji" i abstrahować od kwestii świadomości. Jak wskazuje kolejny fragment, podobieństwa są nawet głębsze:

> Należy dodać, że gdy wprowadza się rozróżnienie na psychiczną strukturę i psychiczną funkcję, często zapomina się o dziwacznym miejscu zajmowanym przez strukturę jako kategorię [służącą do opisu] umysłu. Cała adekwatność terminu struktura [jako nadającego się do opisu zjawisk] w życiu mentalnym zależy od tego, że każdy chwilowy stan świadomości może być potraktowany jako pewna złożoność poddająca się analizie, a terminy na które takie złożoności są rozkładane przez nasze analizy są analogiami, w dodatku oczywiście bardzo wątłymi i wadliwymi analogiami [nawiązującymi do] struktury anatomicznej i morfologii.
>
> (s. 65, tamże)

Bardzo podobny problem pojawia się moim zdaniem współcześnie, ponieważ teorie obliczeniowe opierają się mocno na - może nieco mniej wątłych - analogiach między umysłem i komputerem, a trafność tych analogii jest z wymienionych wcześniej powodów niemal równie dyskusyjna co trafność analogii biologicznych czy chemicznych. Angell dostrzega kolejny wymiar szczególnego statusu teorii funkcjonalnej względem teorii strukturalnej, kiedy pisze:

> Co więcej, gdy udaje nam się za pomocą takich lub innych środków zabezpieczyć to, co nazywamy tym samym wrażeniem lub tą samą ideą, nie tylko nie mamy gwarancji, że kolejne wystąpienie [tego wrażenia lub idei] jest naprawdę repliką pierwszego, ale mamy wręcz sporo powodów by sądzić, że z punktu widzenia treści oryginał nigdy nie jest i nigdy nie będzie dokładnie odtworzony.
>
> Funkcje z drugiej strony pozostają niezmienne w świecie mentalnym i fizycznym. Możemy nigdy nie mieć dwa razy tej samej idei z punktu widzenia struktury wrażeniowej, ale wydaje się, że nic nie przeszkadza nam posiadać tak często, jak tylko sobie tego życzymy, treści świadomych, które *znaczą* to samo. Odgrywają praktycznie jedną i tą samą rolę, niezależnie od tego, jak rozbieżna [względem wcześniejszej] jest ich chwilowa postać.
>
> (s. 65-66, tamże)

Pozwolę sobie dalej podążać tropem analogii obliczeniowe-strukturalne. Chociaż nie przypisywałbym hipotezom dotyczącym funkcji tak dużej odporności na rewizję, całkowicie zgadzam się ze spostrzeżeniem, że zrozumienie funkcji powinno zwykle owocować rezultatami teoretycznymi o większej trwałości i znaczeniu, niż oparte na niejawnych lub słabo wyartykułowanych intuicjach na temat tych funkcji teorie obliczeniowe. Odpowiednio zinterpretowany, problem zmienności i nieuchwytności struktury świadomości ma wiele wspólnego ze statystycznym problemami oceny ilościowej złożonych modeli mechanistycznych, identyfikowalności architektury i wielorakiej obliczeniowej realizowalności komputacyjnie scharakteryzowanych stanów i procesów mentalnych. Dodałbym jeszcze, wynikającą z samego charakteru teorii funkcjonalnej, naturalną zdolność do uogólnienia i podatność na teoretyczne środki oceny, analizy i rewizji, które są dostępne w znacznie mniejszym stopniu, gdy teleologicznie rozumiana funkcja zajmuje w teorii miejsce drugoplanowe. W pewnym sensie, zbliżone stanowisko można odnaleźć również u Angella:

> Na przykład, pamięć i wyobraźnia często są traktowane w sposób do tego stopnia podkreślający ich odrębność, że niemal wyklucza się występujące między nimi funkcjonalne podobieństwa. Z perspektywy funkcjonalnej nie są one oczywiście niczym innym jak wariantami pojedynczej i podstawowej [funkcji] kontroli. W szczególności, najprostsza postać strukturalizmu [ang. *austere structuralism*] jest nieuchronnie skazana na wyolbrzymianie różnic, na skutek czego życie psychiczne zaczyna się coraz bardziej rozpadać; gdy już zostanie [to życie psychiczne] z powrotem złożone w całość, sprawia wrażenie, jakby utraciło część swego wigoru i żywotności.
>
> (s. 73-74, tamże)

Angell pisze tutaj wyraźnie o dwóch rodzajach „unifikacji", to jest unifikacji dzięki stworzeniu ogólniejszej, wyrażonej na wyższym poziomie abstrakcji teorii i unifikacji przez połączenie w całość lokalnych teorii strukturalnych, sugerując przy tym istnienie bliżej nieokreślonych wad tej ostatniej.

Zarzuty sformułowane przez Angella mają dwa, odrębne wymiary. Jeden z nich to wymiar świadomościowy, a drugi to funkcjonalno-strukturalny. Gdy tylko usunąć elementy związane ze świadomością, podobieństwa między strukturalizmem a współczesną psychologią poznawczą stają się ewidentne. Nie przypadkiem Anderson obwieścił odkrycie atomowych składników myślenia.

Przywiązanie do świadomości jako centralnego przedmiotu psychologii było przypuszczalnie jedną z ważniejszych przyczyn upadku funkcjonalizmu[4]. Można by spekulować, dlaczego program funkcjonalizmu psychologicznego nie został zrealizowany w

[4]Wykluczając świadomość i wolę z obszaru zainteresowań psychologii naukowej, Dewey stanowił pod tym względem wyjątek. Uważał, że psychologię powinno interesować przede wszystkim zachowanie, tak jak się ono rozgrywa w interakcji ze środowiskiem naturalnym, ale już nie świadomość czy wola.

stopniu pozwalającym na jego przetrwanie. Być może przeszkodą na drodze do stworzenia wystarczająco przekonujących i trwałych teorii funkcjonalnych był brak odpowiedniego aparatu pojęciowego. Funkcjonalistów z przełomu wieków łączyła fascynacja pojęciami i metodami biologii, szczególnie teorią ewolucji Darwina, ale także aparatem pojęciowym fizjologii. Podejmowane przez nich próby teoretyzowania na temat funkcji procesów psychicznych były przez to czasami mocno oparte na analogiach biologicznych. Lektura tekstów z tego okresu pozwala zrozumieć, jak bardzo te skojarzenia okazywały się czasem niejasne i nieproduktywne.

Niezależnie od tego, z jakich powodów program funkcjonalizmu psychologicznego został porzucony (nadejście behawioryzmu), a z jakich powodów po prostu się nie powiódł, pozostaje faktem, że główny postulat, aby uczynić z funkcji centralny przedmiotem zainteresowania, nie został nigdy zrealizowany. We współczesnej psychologii poznawczej teleologicznie rozumiana funkcja pozostaje w tle oficjalnych rozważań, żeby jednak teoria dotycząca procesów poznawczych w ogóle brzmiała sensownie, nie sposób oderwać się od myślenia w kategoriach teleologicznych.

Pojęcia takie jak uwaga, pamięć, uczenie się albo percepcja są w psychologii poznawczej najczęściej definiowane teleologicznie, bo najprawdopodobniej inaczej się nie da. Te funkcje są tym czym są ze względu na problemy, które częściowo rozwiązują. Uwaga selektywna ma być tym, co jest konieczne do poradzenia sobie ze złożonością nieustannie zmieniającej się stymulacji, pamięć ma być tym, co pozwala na uwzględnienie wiedzy o historii interakcji, uczenie się ma być tym, co umożliwia wzrost skuteczności działań, i tak dalej. Żeby określić mechanizm obliczeniowy trzeba najpierw poznać nie tyle konkretną relację między wejściami i wyjściami, ile raczej zadanie, które ten mechanizm ma rozwiązywać. Rozwiązanie tego zadania można potem rozbić na teleologicznie scharakteryzowane elementy składowe, na przykład przeszukiwanie pamięci krótkoterminowej (zadanie) można rozbić na procesy porównywania, czyli hipotetyczne operacje, które służą do ustalenia, czy elementy są takie same czy różne, i procesy kontroli przebiegu porównywania, czyli między innymi rozwiązania problemu kolejności porównywania (szeregowość-równoległość, kolejność etapów) i przetargu czas-poprawność.

Na tym etapie formułowania teorii analizowany proces lub system można przedstawić w postaci diagramu przepływu, takiego jak te, które w 1967 roku omawiał Neisser. Za pomocą takich nadal bardzo popularnych i użytecznych diagramów charakteryzuje się proces lub system jako coś złożonego z *komponentów*, które spełniają w ramach procesu lub systemu określone zadanie, które do czegoś służą (Deutsch, 1960; Cummins, 1975, 2000). Wyróżnione komponenty można dalej rozkładać na elementy składowe, coraz bardziej uszczegóławiając strukturę funkcjonalną procesu lub systemu. Deutsch (1960) nazywa takie teorie teoriami etapu pierwszego. Dopiero teorie etapu drugiego określają abstrakcyjnie ujęte procesy fizjologiczne, odpowiedzialne za realizację tych funkcji.

Mechanizm obliczeniowy nie jest niczym innym jak stosunkowo abstrakcyjnym opisem realizacji fizjologicznej. Na poziomie mechanizmu obliczeniowego jako takiego nie chodzi już o zadania i ich rozwiązania. Na przykład, opis mechanizmu i struktury pa-

mięci deklaratywnej w kategoriach sieci powiązanych elementów o zmieniającej się aktywacji nie jest opisem w kategoriach teleologicznych, tylko w kategoriach mechanistycznych. Wyjaśnienie zjawisk pamięciowych w terminach mechanistycznych nie jest jednak możliwe, dopóki mechanizm nie będzie zinterpretowany teleologicznie. Dopiero wtedy można powiedzieć, że osoba badana udzieliła takiej a nie innej odpowiedzi, ponieważ tak a nie inaczej zmieniła się struktura asocjacji i taka a nie inna była relatywna aktywacja powiązanych ze sobą elementów. Za każdym razem trzeba założyć, że ta osoba wykonywała pewne, na przykład pamięciowe, zadanie, trzeba założyć że *coś robiła*.

Intuicje na temat zadania i jego możliwych rozwiązań są podstawowym kryterium wstępnego zawężenia i dookreślenia przestrzeni alternatywnych mechanizmów i struktur obliczeniowych. Teoria komputacyjna ogranicza tą przestrzeń nie tylko przez to, że może dostarczać nieznanych wcześniej, trafnych predykcji, ale również, a nawet przede wszystkim przez to, że określa psychologiczny sens mechanizmu.

Teoria identyfikacji systemów przetwarzających skończoną liczbę elementów Townsenda dopuszcza tylko takie systemy, które *a priori* wydają się być rozsądnymi, względnie skutecznymi rozwiązaniami zadania przetwarzania skończonej liczby elementów w warunkach presji czasowej. Dokładnie z takich przesłanek korzystał Sternberg uzasadniając swoje hipotezy (przeszukiwanie samowygaszające jako przypuszczalnie optymalne). W teorii Townsenda możliwość wielokrotnego porównywania tych samych dwóch elementów nie jest w ogóle rozważana, tak samo jak możliwość przerwania przeszukiwania zanim (być może błędnie) element poszukiwany zostanie odnaleziony. To konieczne wstępne zawężenie przestrzeni alternatywnych mechanizmów nie wynika ze znajomości konkretnej relacji między „wartościami na wejściu i wyjściu", ponieważ taka relacja nie jest zadaniem, tylko funkcją w znaczeniu matematycznym. Rozumienie procesu poznawczego jako względnie skutecznego rozwiązania określonego zadania odgrywa cały czas podstawową, regulacyjną rolę na etapie identyfikacji mechanizmu. Własności teleologiczne, a nie mechanistyczne, czynią go tym, czym jest dla psychologa, czyli na przykład przeszukiwaniem pamięci.

Uogólnianie teorii poznawczej *zawsze* polega na uogólnianiu ze względu na zadanie. Można powiedzieć, że przeszukiwanie pamięci krótkoterminowej jest szczególnym przypadkiem przetwarzania skończonej liczby elementów, albo decyzja leksykalna jest szczególnym przypadkiem wyboru z dwóch alternatyw, albo ocena kauzalna jest szczególnym przypadkiem wnioskowania bayesowkiego na modelach graficznych. W uogólnionym mechanizmie obliczeniowym jako mechanizmie obliczeniowym nie ma nic, co zachowywałoby jego psychologiczny charakter. Mechanizmy obliczeniowe interesują psychologów poznawczych tylko o tyle, o ile są rzeczywistymi lub potencjalnymi realizacjami rozwiązań pewnych zadań.

Jak trafnie zauważył Piccinini (2004), relacje funkcjonalne nie muszą być wcale obliczeniowe czy mechanistyczne. Czym innym jest stwierdzenie, że stany lub procesy są określone przez ich funkcję, a czym innym, że są określone przez zależności przyczynowo-skutkowe z innymi stanami, procesami, wejściami i wyjściami. Na podstawie szczegóło-

wych analiz teksów źródłowych Piccinini moim zdaniem przekonująco demonstruje, że Fodor i Putnam utożsamili te dwa rodzaje relacji bezwiednie.

Wydaje się więc, że Anderson miał rację gdy pisał, że prawdziwa teoria poznawcza znajduje się na wyższym niż obliczeniowy poziomie opisu. Taka teoria jest „prawdziwa" w tym znaczeniu, że dostarcza (niekoniecznie od razu poprawnych) definicji procesów lub stanów poznawczych jako procesów lub stanów poznawczych. Dopiero na tym poziomie można powiedzieć, co wie, czego nie wie, co robi i co stara się zrobić system, którego mechanizm działania jest opisany przez zintegrowany albo lokalny model obliczeniowy. Teoria na poziomie mechanistycznym jest tylko o tyle trafna, o ile zgadza się z poprawną teorią komputacyjną, co nie znaczy oczywiście, że teoria komputacyjna jest poprawna tylko dlatego, że jest komputacyjna.

Należy podkreślić, że teoria komputacyjna nie musi być racjonalna. Scharakteryzowanie procesów lub stanów jako elementów rozwiązania pewnych zadań nie wymaga, żeby te rozwiązania były optymalne, tylko żeby były, jak to określił Simon, „wystarczająco dobre", inaczej nie byłyby żadnymi rozwiązaniami. Odpowiedź na pytanie o to, czy ludzie faktycznie są racjonalni, jest oczywista. Ludzie nie muszą zawsze działać racjonalnie w żadnym interesującym znaczeniu i w wielu wypadkach z pewnością tak nie działają. Odpowiedź na pytanie, czy psychologiczna teoria komputacyjna powinna być racjonalna, oczywista już nie jest. Davidson i Dennett są autorami dwóch odmiennych, klasycznych stanowisk dotyczących regulacyjnej roli kryteriów racjonalności w teoriach psychologicznych.

7.3 Regulacyjna rola kryterium racjonalności według Davidsona i Dennetta

Podobnie jak Putnam i Fodor, Davidson podjął próbę udzielenia odpowiedzi na pytanie, co sprawia, że zjawiska mentalne są mentalnymi, a także co sprawia, że dane zjawisko mentalne jest tym właśnie zjawiskiem mentalnym, na przykład zapominanie jest zapominaniem. Bycie zjawiskiem mentalnym to niekoniecznie to samo, co bycie zjawiskiem poznawczym, zjawiska poznawcze są jednak zjawiskami mentalnymi, dlatego wszystko co pisze Davidson stosuje się wprost do psychologii poznawczej. Jak pisze Davidson (1984/1992):

> Mentalnymi możemy nazywać te słowa, które wyrażają postawy propozycjonalne, takie jak wierzenie, noszenie się z zamiarem, pragnienie, żywienie nadziei, poznawanie, postrzeganie, przypominanie sobie i tak dalej. (...) są to słowa psychologiczne w tych użyciach, w których tworzą konteksty wyraźnie nieekstensjonalne. (...) cechą wyróżniającą tego, co mentalne, nie jest jego prywatny charakter, subiektywność czy niematerialność, ale to, iż wykazuje własność, którą Brentano nazwał intencjonalnością.

(s. 168-169, tamże)

Według Brentana (1999) tym, co odróżnia fizyczne od mentalnego, jest właśnie intencjonalność. Każde zjawisko czy akt psychiczny ma to do siebie, że jest na coś skierowany (co niekoniecznie znaczy, że celowy), posiada jakąś zawartość, na przykład lęk jest lękiem przed czymś albo o coś, strach jest lękiem przed czymś bliżej nieokreślonym, zapamiętywanie, przechowywanie i odtwarzanie jest zapamiętywaniem, przechowywaniem i odtwarzaniem czegoś, i tak dalej.

Cechą charakterystyczną zdań przypisujących zjawiska czy akty inten*cjonalne* jest to, że takie zdania tworzą zwykle kontekst inten*sjonalny*. Inten*cjonalność* to cecha zjawisk mentalnych, inten*sjonalność* to cecha wyrażeń językowych. Gdy kontekst jest intensjonalny, a nie ekstensjonalny, zamiana wyrażeń na synonimy, czyli wyrażenia o tej samej ekstensji, nie musi zachowywać prawdziwości zdania. Na przykład, w zdaniu „adam chciał zjeść jabłko" zmiana wyrażenia „jabłko" na wyrażenie „owoc, który rósł w znajdującym się nieopodal ogródku" może sprawić, że z prawdziwego zdanie zmieni się w fałszywe, zależnie od tego, co Adam wie na temat swojego otoczenia i co ma akurat na myśli. Będzie tak nawet jeżeli faktycznie jedyne owoce rosnące w znajdującym się nieopodal ogródku to jabłka. Chęć zrobienia czegoś jest tak zwaną postawą propozycjonalną, to znaczy relacyjnym zdarzeniem czy stanem mentalnym, łączącym osobę z treścią jakiegoś sądu. Nieekstensjonalny charakter zdań opisujących postawy propozycjonalne sprawia, że trudno określić warunki ich prawdziwości.

Być może jednym z powodów, dla których wczesny Putnam i inni zwolennicy funkcjonalizmu obliczeniowego tak często posługiwali się przykładem czucia bólu, jest domniemany ekstensjonalny charakter zdań przypisujących ten stan organizmowi. Teoria mechanistyczna jest ekstensjonalna, a mechanistyczność wydaje się wyróżnikiem teorii naukowych, tak w każdym razie postrzegał ją Putnam w zacytowanym wcześniej fragmencie. Putnamowi nie udało się jednak uniknąć odwołania sie do kryteriów racjonalności stanu czucia bólu, a więc do odpowiedniego związku między czuciem bólu a tym, co organizm postrzega i robi. Nie udało mu się tym samym wysłowić warunków prawdziwości tych zdań, nie tworząc przy tym kontekstu nieekstensjonalnego. Davidson (1984/1992) twierdzi, że indywiduacja[5] postaw propozycjonalnych nie jest możliwa bez założenia racjonalności na poziomie całego systemu:

> Podobnie jak nie możemy w zrozumiały sposób przypisywać długości żadnym przedmiotom, o ile nie istnieje wszechstronna teoria dotycząca tego rodzaju przedmiotów, nie możemy też przypisywać działającemu postawy propozycjonalnej inaczej niż w ramach aparatu pojęciowego pewnej samodzielnej teorii jego przeświadczeń, pragnień, zamierzeń i decyzji.
>
> Nie można przypisywać nikomu pojedynczych przeświadczeń na podstawie zachowania werbalnego, wyborów czy innych fragmentarycznych

[5]Indywiduacja to (re)identyfikacja w różnych warunkach.

> oznak, niezależnie od tego, jak byłyby one wyraźne lub oczywiste, ponieważ poszczególne przeświadczenia rozumiemy tylko wtedy, gdy zgadzają się z innymi przeświadczeniami, preferencjami i zamierzeniami, nadziejami, lękami, oczekiwaniami itp. (...) treść postawy propozycjonalnej zależy od miejsca, jakie zajmuje ona w całym schemacie.
>
> Przypisywanie wysokiego stopnia konsekwencji nie może uchodzić jedynie za przejaw życzliwości; jest ono nieuniknione, jeżeli mamy sensownie oskarżać ich [to znaczy ludzi] o błąd i pewien stopień irracjonalności. Zupełne pomieszanie, podobnie jak totalny błąd, jest nie do pomyślenia; nie dlatego, że wzdraga się przed tym wyobraźnia, ale dlatego, że zbyt duże pomieszanie nie pozostawia niczego, co można by ze sobą pomieszać, a całkowity błąd powoduje erozję podłoża prawdziwego przeświadczenia, na którego tle jedynie można błąd zinterpretować. (...) W tej mierze, w jakiej nie udaje nam się odkryć spójnego i wiarygodnego schematu w postawach i działaniach innych ludzi, tracimy po prostu szansę traktowania ich jako osób.
>
> (s. 186-187, tamże)

Wymóg przybliżonej racjonalności na poziomie całego systemu ma jednak przekreślać możliwość „ścisłych praw psychofizykalnych":

> Ścisłe psychofizykalne prawa nie istnieją z powodu zasadniczo odmiennego uwikłania mentalnego i fizykalnego schematu pojęciowego. Rzeczywistość fizyczna charakteryzuje się tym, że zmianę fizyczną można wyjaśnić za pomocą praw, które łączą ją z innymi zmianami i warunkami opisywanymi fizykalnie. Cechą tego, co mentalne, jest fakt, że przypisywanie jednostce zjawisk mentalnych musi harmonizować z tłem jej motywów, przeświadczeń i zamierzeń. Między tymi dziedzinami nie ma ścisłych związków, jeśli każda z nich pozostaje wierna właściwemu sobie źródłu świadectw. Nomologiczna nieredukowalność tego, co mentalne, nie bierze się jedynie ze spoistej natury świata myśli, preferencji i zamiarów, ponieważ taka współzależność występuje powszechnie w teorii fizykalnej i daje się pogodzić z istnieniem jedynego właściwego sposobu interpretowania ludzkich postaw, bez relatywizacji do metody przekładu. Nieredukowalność ta nie polega też po prostu na istnieniu wielu równie dobrych metod, to bowiem pozostaje w zgodzie z arbitralnością wyboru metody, według której dokonuje się przyporządkowania cech mentalnych. Rzecz polega raczej na tym, że kiedy posługujemy się pojęciami przeświadczenia, pragnienia, itp. musimy być gotowi, w miarę gromadzenia świadectw, do modyfikowania naszej teorii w świetle wymogów całościowej spójności; konstytutywny ideał racjonalności częściowo kontroluje każdą fazę ewolucji powstającej stopniowo teorii.

> (...) Trzeba, jak sądzę, wyprowadzić stąd wniosek, że nomologiczne pęknięcie między tym, co mentalne, a tym, co fizyczne, jest nieuniknione dopóty, dopóki traktujemy człowieka jako istotę racjonalną.
>
> (s. 188-189, tamże)

Zjawiska mentalne byłyby więc mentalnymi przez to, że są intencjonalne, a określone zjawiska mentalne byłyby tymi właśnie, a nie innymi zjawiskami mentalnymi, na przykład domniemane pragnienie zjedzenia jabłka byłoby faktycznie pragnieniem zjedzenia jabłka, ze względu na ich pozycję w całym systemie racjonalnie powiązanych postaw propozycjonalnych. Podobnego zdania są Fodor i Pylyshyn, mówiąc o „systematycznej" czy może raczej „systemowej" naturze myśli (ang. *sistematicity of thought*) (Fodor i Pylyshyn, 1988).

Pozornie poglądy Davidsona mogą się wydawać zbliżone do teorii zawartości stanów i aktów psychicznych Dennetta (1989). Według Denneta, określenie, przewidywanie i wyjaśnianie zachowania ze względów praktycznych wymaga często przyjęcia tak zwanego „stanowiska intencjonalnego":

> Oto jak to działa: najpierw decydujesz się potraktować obiekt, którego zachowanie ma być przewidywane, jako racjonalnego agenta; potem próbujesz odgadnąć, jakie przekonania agent powinien posiadać, ze względu na jego miejsce w świecie i cel. Potem ustalasz, ze względu na te same przesłanki, jakie powinien mieć pragnienia i ostatecznie przewidujesz, że ten racjonalny agent będzie działał w taki sposób, aby przybliżyć się do osiągnięcia tych celów w świetle swoich przekonań. Odrobina praktycznego wnioskowania z wybranego zbioru przekonań i pragnień w większości przypadków doprowadzi do wniosku, co agent powinien zrobić; to jest właśnie tym, co przewidujesz, że agent zrobi.
>
> (s. 17, tamże)

Dennett zdaje się jednak twierdzić, że stanowisko intencjonalne jest tylko użytecznym poziomem abstrakcji. Można przyjąć stanowisko intencjonalne nie przypisując desygnatom pojęć psychologii potocznej żadnej głębszej realności. Davidson zdaje się sugerować, że stanowisko intencjonalne jest nie tylko użyteczne, ale również z zasady prawdziwe.

Davidsonowi, Dennetowi, Fodorowi i Pylyshynowi (systematyczność myślenia) chodzi raczej o racjonalność teoretyczną (logiczna poprawność rozumowań, poprawne i konsekwentne uwzględnianie świadectwa empirycznego), podczas gdy teorie racjonalne w psychologii poznawczej dotyczą najczęściej przede wszystkim optymalnych sposobów działania (racjonalność praktyczna), co sprawia, że związek między postulowanym przez Davidsona wymogiem racjonalności na poziomie całego systemu a psychologicznymi teoriami racjonalnymi nie jest zbyt jasny. Z braku powodów aby postąpić inaczej

zakładam dalej, że teorie psychologiczne mogą dotyczyć obu tych rodzajów racjonalności, a to, czy zaakcentowana jest bardziej racjonalność teoretyczna czy praktyczna, będzie zależało od podjętego problemu badawczego.

Uważam, że wymóg przybliżonej racjonalności jest nieusuwalną częścią teorii procesów i stanów poznawczych jako procesów i stanów poznawczych, a także, że przybliżona racjonalność musi obowiązywać na poziomie całego systemu. Ta część argumentacji Davidsona wydaje mi się wystarczająco przekonująca. Za nieuzasadniony uważam natomiast wniosek o niemożliwości praw psychofizycznych. Podatność hipotez dotyczących występowania i przebiegu procesów lub stanów mentalnych na rewizję ze względu na przyszłe świadectwa nie odróżnia tych hipotez od hipotez fizykalnych. Na podstawie znanych mi tekstów nie udało mi się ustalić, co dokładnie ma na myśli autor pisząc o „zasadniczo odmiennym uwikłaniu mentalnego i fizykalnego schematu pojęciowego". Jeżeli tą odmienność można sprowadzić do racjonalnego charakteru tego, co mentalne, to opisane przeze mnie w rozdziale trzecim racjonalne teorie generalizacji i oceny kauzalnej byłyby przykładem, że uwzględnienie racjonalności nie tylko nie przeszkadza, ale nawet czasem pomaga w odkryciu praw psychofizycznych tak ścisłych, jak to jest tylko w psychologii poznawczej możliwe. Jeżeli odmienność tego „uwikłania" polega na intencjonalnym charakterze tego, co mentalne, na owym „byciu o czymś", a przypuszczalnie właśnie o to chodzi Davidsonowi, sprawa się komplikuje. Nie zgadzam się jednak z poglądem, że stwierdzenie, co robi jakiś organizm, albo że w ogóle coś robi, na przykład w znaczeniu, że jakoś się zachowuje albo do czegoś dąży, z konieczności tworzy kontekst intensjonalny. Uzasadnienie tej tezy przedstawię w ostatnim rozdziale pracy. Zanim to nastąpi, zajmę się kwestią brakującego elementu systemu, na poziomie którego racjonalność powinna obowiązywać.

7.4 Środowisko w psychologii poznawczej

W moim odczuciu sukcesy aplikacyjne psychologii poznawczej nie są szczególnie imponujące. Do wiarygodnych powodów, dla których te sukcesy są nikłe, albo nie są tak imponujące, jak mogłyby być, należy zdumiewająco uporczywe ignorowanie znaczenia środowiska naturalnego. Nie chodzi przy tym wcale o to, czy badania są wystarczająco często przeprowadzane w warunkach naturalnych, tylko czy środowisko naturalne ze względu na jego związek z zachowaniem i procesami psychicznymi stanowi w wystarczającym stopniu przedmiot badań *teoretycznych* w psychologii. Nie stanowi.

Tak zwana psychologia środowiskowa (Gifford, 2007) jest nurtem, albo inaczej, dziedziną interdyscyplinarną, wyraźnie zorientowaną na badania aplikacyjne, ale nie na badania podstawowe. Trzeba się cieszyć, że jest, ale nie ma żadnego powodu, aby na tym poprzestać. Podobny brak większego zainteresowania formułowaniem teorii środowiska można dostrzec w „poznaniu usytuowanym" czy „ucieleśnionym" (Clancey, 1997; Barsalou, 2008), co pewnie częściowo wynika z nieodstającego od normy stosunku do

badań teoretycznych jako takich. Jak wobec tego w ogóle ma być możliwe stworzenie pozbawionej elementarnych wad teorii zachowania, albo pełniejsze zrozumienie funkcji psychicznych?

W artykule dotyczącym niedocenionej spuścizny Egona Brunswika Dunwoody (2006) dostarcza zawstydzającej diagnozy stopnia, w jakim środowisko jest uwzględniane w teoriach psychologicznych. Tylko w jednym z trzydziestu sześciu sprawdzonych podręczników udało mu się znaleźć definicję psychologii zawierającą termin „środowisko". Zdaniem Dunwoody'ego, do niedawna istniały tylko *dwie* eksplicitne psychologiczne teorie środowiska, to jest teorie Brunswika i Gibsona. Nawet gdyby Dunwoody się mylił i takich teorii było dziesięć razy więcej, liczba psychologicznych teorii środowiska budziłaby zdumienie. Uporczywe ignorowanie środowiska jako przedmiotu teorii jest w moim odczuciu zarazem najbardziej zadziwiającym i niezaprzeczalnym faktem dotyczącym dotychczasowej praktyki badawczej w psychologii nie tylko poznawczej.

Brunswik jest między innymi autorem pojęcia trafności ekologicznej, rozumianej nie jako przeprowadzanie badań w warunkach naturalnych, tylko jako uwzględnienie środowiska na poziomie teorii. Według tego autora, trafny ekologicznie plan badawczy powinien polegać na stworzeniu warunków eksperymentalnych *reprezentatywnych* dla zależności występujących w środowisku, a nie na przeprowadzaniu badań w warunkach naturalnych (Brunswik, 1955). Zdaniem Brunswika, kumulacja wiedzy w psychologii nie będzie następowała w zadowalającym tempie, dopóki zagadnienie interakcji organizmu ze środowiskiem nie stanie się centralnym przedmiotem tej dyscypliny. Według Dunwoody'ego, pomimo, że zachowanie wydaje się być centralnym przedmiotem zainteresowania przedstawicieli rozmaitych behawioryzmów, tego samego nie można powiedzieć o zredukowanym czasem do minimum („bodźce") środowisku, a więc również o interakcji organizmu ze środowiskiem. Nie można dobrze zrozumieć procesów psychicznych i zachowania przyglądając się jedynie bodźcom i reakcjom, albo bodźcom, reakcjom i temu, co dzieje się między nimi „w głowie", dlatego, że środowisko nie jest zbiorem bodźców.

Żeby dobrze zrozumieć, na czym polega interakcja organizmu ze środowiskiem, należy według Brunswika w pierwszej kolejności zwrócić uwagę na skuteczność adaptacji, a nie, tak jak to ma najczęściej miejsce w psychologii poznawczej, na popełniane błędy. Błędy mogą powiedzieć coś o ograniczeniach systemu poznawczego tylko, jeżeli wiadomo, że faktycznie są błędami. Żeby to stwierdzić, trzeba najpierw wiedzieć, jakie właściwie zadanie rozwiązuje organizm, co zwykle wcale nie jest oczywiste.

Dlaczego sposób, w jaki zadanie opisuje instrukcja przeznaczona dla osób badanych, miałby wprost wyznaczać zadanie, tak jak jest ono potraktowane przez te osoby? Dlaczego na przykład użyteczność przypisana przez eksperymentatora alternatywom w pewnym eksperymencie dotyczącym podejmowania decyzji miałaby być użytecznością z perspektywy osób badanych? Dlaczego prawdopodobieństwa przypisywane rozmaitym zdarzeniom miałyby być przez te osoby traktowane tak, jak prawdopodobieństwa rzeczywiste, a nie jak elementy pewnego (niekoniecznie wciągającego) zadania matematyczne-

go? Dlaczego osoby badane wykonujące zadanie Sternberga miałyby pamiętać prezentowane symbole, a nie coś innego, co z wystąpieniem tych symboli jest jakoś związane, na przykład utworzone z tych symboli wyrazy lub zlepki, albo symbole jako elementy szerszego kontekstu? Jak w ogóle mają się zadania wykonywane w warunkach laboratoryjnych do zadań w warunkach naturalnych? Przypuszczenie, że zadanie z perspektywy osób badanych jest tym samym zadaniem, które opisuje eksperymentator, jest nie tylko wątpliwe, ale często w oczywisty sposób błędne. Prosząc o wykonanie określonego zadania badacz nie tworzy zamkniętej w czasie i przestrzeni sytuacji. Opierając się jedynie na tym elementarnym spostrzeżeniu można poddać w wątpliwość wnioski oparte na licznych eksperymentach, ujawniających rzekomo nieracjonalność osób badanych, takich jak te, które w latach 70-tych i 80-tych przeprowadzili Tversky i Kahneman. Aby ustalić, na ile osoby badane są racjonalne, należy najpierw ustalić, co próbują zrobić i do jakiego stopnia granice tego, co mogą zrobić, są wyznaczone przez sam charakter zadania. To, czego eksperymentator wymaga od osób badanych jest tylko jednym z wielu elementów problemu, który osoby badane próbują rozwiązać.

Może się wydawać, że uczynienie z funkcji rozumianej jako (częściowe) rozwiązanie zadania centralnego przedmiotu psychologii wymaga odwołania się do badań ewolucyjnych. Faktycznie, współcześni przedstawiciele podejścia racjonalnego, do których należy w pierwszej kolejności zaliczyć zwolenników analizy racjonalnej i autorów określających swoje teorie jako probabilistyczne modele poznawcze, często powołują się na argumenty ewolucyjne. Teoria Sheparda zdaje się opierać właśnie na takich uzasadnieniach. Jednocześnie dokładniejsza analiza tych propozycji nie pozostawia cienia wątpliwości, że przywoływane przesłanki ewolucyjne nie są ani szczególnie jasne, ani tym bardziej solidnie wsparte przez wyniki odpowiednich, to znaczy ewolucyjnych badań. Autorzy Ci zresztą sami przyznają, że argumenty ewolucyjne nie odgrywają wcale kluczowej roli na etapie uzasadnienia formułowanych problemów i hipotez, chociaż mogą odgrywać niebagatelną rolę na etapie ich odkrycia.

Najważniejsze zagadnienia związane z celową interakcją ze środowiskiem można z powodzeniem badać, abstrahując nie tylko od rozważań ewolucyjnych, ale również, co mniej oczywiste, od konkretnego środowiska. Organizmy żywe rozwiązują nieustannie *aktualne* problemy i to właśnie te problemy mają decydujące znaczenie dla tego, co robią. Nie trzeba badać historii gatunku, żeby stwierdzić, na czym one polegają. Celowa interakcja organizmów ze środowiskiem rozgrywa się cały czas, przed naszymi oczami. Potrzeba *ogólnej* teorii środowisk wynika z kolei stąd, że te aktualne problemy są niezliczone i bez odpowiedniej, to znaczy między innymi ogólnej teorii, nie można ich zidentyfikować. Wymóg posługiwania się modelami rzeczywistego środowiska, w połączeniu z brakiem teorii tego, czym w ogóle jest środowisko jako przedmiot psychologii, należy do ważniejszych ograniczeń analizy racjonalnej w ujęciu Andersona. Psychologiczna klasyfikacja pewnych ważnych przypadków rzeczywistych środowisk jest tym, czym na przykład biologiczna klasyfikacja znanych gatunków. Taka klasyfikacja nie zastąpi teorii określającej to, co ma być klasyfikowane. W ostatnim rozdziale postaram się uzasad-

nić, że warunkiem istnienia ogólnej psychologicznej teorii celowej interakcji ze środowiskiem nie jest również, aby cele o rzeczywistym znaczeniu dla organizmu były a priori znane.

Trzeba w końcu oderwać się od konkretu. Jak doskonale ujął to Wittgenstein (1953/2001): „Te aspekty rzeczy, które są dla nas najważniejsze, są jednocześnie ukryte przez swoją prostotę i przez to, że są dobrze znane. (Nie da się zauważyć czegoś - dlatego, że cały czas znajduje się przed oczami)" (s. 129, tamże). Żeby „zacząć wszystko od nowa", potrzebny jest zarazem solidny i ostrożnie sformułowany punkt zaczepienia. Ten punkt zaczepienia musi pozwalać na formułowanie teorii zachowania i procesów poznawczych nie na obraz i podobieństwo maszyny Turinga, automatu probabilistycznego, kalkulatora, komputera, biologicznie scharakteryzowanego organizmu żywego, mózgu, systemu dynamicznego, robota, niezwykle zasłużonego dla badań psychologicznych studenta psychologii, ani nawet człowieka. Wszystkie te sposoby ujęcia mają niezaprzeczalne zalety, ale są tylko użytecznymi dodatkami do ewentualnej podstawowej, ogólnej teorii psychologicznej.

Byłoby najlepiej, gdyby teoria zachowania i procesów psychicznych ludzi i zwierząt była oparta na wyidealizowanym, uzasadnionym modelu zachowania i procesów psychicznych abstrakcyjnie rozumianego agenta, działającego w abstrakcyjnie ujętym środowisku. Parafrazując Pylyshyna, (wczesnych) Fodora i Putnama i wielu innych zwolenników funkcjonalizmu obliczeniowego, wydaje mi się, że obecnie istnieje tylko jeden, pod ważnymi względami nie przypominający niczego, co do niedawna zaproponowano w psychologii, sztucznej inteligencji i filozofii umysłu, obiecujący kandydat na taką teorię. Co nie powinno dziwić, autorami tej teorii nie są psychologowie.

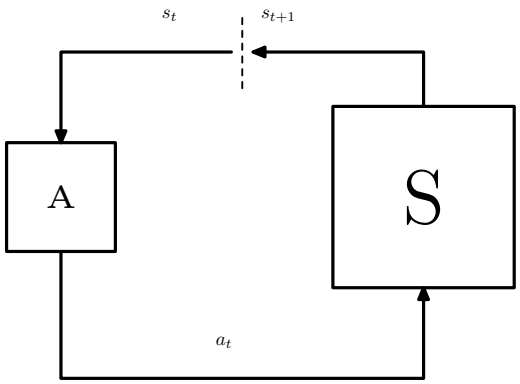
Rozdział 8

Funkcjonalizm racjonalny w praktyce

Główna konstruktywna propozycja tej pracy jest stanowiskiem metateoretycznym, dotyczącym właściwego charakteru ogólnej teorii zachowania, stanów i procesów poznawczych. Najważniejsze intuicje, na których jest ono oparte, zawdzięczam przede wszystkim lekturze pracy zawierającej opis pewnej szczególnej ramy pojęciowej i przykłady zastosowania tej ramy do analizy problemów wymagających inteligencji. Spróbuję przedstawić tą ramę pojęciową w sposób ujawniający jej psychologiczny charakter, przez co będę zmuszony do wprowadzenia kilku zmian w przyjętym w języku polskim nazewnictwie.

8.1 Najważniejsze elementy

Żeby możliwe było zachowanie, musi istnieć środowisko i coś lub ktoś, kto się w nim zachowuje. Dogodnie będzie używać terminu nieobarczonego nadmiernie bagażem przesądów. Niech tym kimś lub czymś będzie agent. Zachowanie rozgrywa się w czasie, więc trzeba dodać czas, na początek dla uproszczenia dyskretny. Nie obejdzie się też bez wejścia i wyjścia po stronie agenta *i środowiska*:

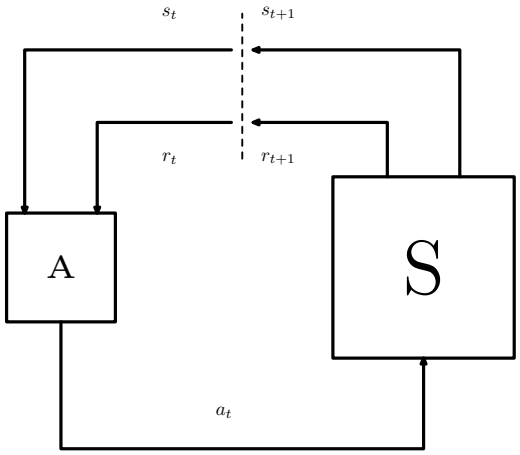


Rysunek 8.1: Przepływ sygnałów między agentem i środowiskiem

Proporcje między agentem i środowiskiem są nieadekwatne. Zawartość pudełka S jest niepomiernie większa niż zawartość pudełka A. Z trudną do ogarnięcia zawartością pudełka S najlepiej będzie sobie poradzić abstrahując od czego tylko się da. Z zawartością pudełka A najlepiej będzie sobie poradzić powstrzymując się od zaglądania do środka tak długo, jak to tylko możliwe.

Na wejściu agent otrzymuje od środowiska sygnały, których charakter nie powinien nas na tym etapie interesować. Im więcej będziemy w stanie powiedzieć na możliwie wysokim poziomie abstrakcji, tym większy będzie zakres potencjalnie opisywalnych i wyjaśnialnych przez teorię zjawisk. W każdym kroku interakcji agent wykonuje pewne działanie, ale działanie nie jest tym samym co zachowanie. Gdyby zachowanie dawało się sprowadzić do wyjścia pudełka otrzymującego sygnały ze środowiska (funkcjonalizm obliczeniowy uzupełniony o środowisko), kalkulator również by się zachowywał. Zakładam, że Czytelnik nie będzie skłonny się na to zgodzić.

W schemacie brakuje czegoś, co pozwalałoby zacząć mówić o celowości działań. Najmniej zobowiązującym i najbardziej elastycznym sposobem, umożliwiającym pojawienie się celowości jest prawdopodobnie wprowadzenie pojęcia użyteczności. Bez jakiegoś elementarnego sygnału, informującego agenta o tym, czy jest źle czy dobrze, celowości nie można sobie wyobrazić. Na początek będzie to użyteczność natychmiastowa, którą przedstawimy za pomocą dodatkowej strzałki od środowiska do agenta.



Rysunek 8.2: Przepływ sygnałów między agentem i środowiskiem uzupełniony o natychmiastową nagrodę

Ten „sygnał nagrody" musi pochodzić ze środowiska, inaczej agent miałby nad nim kontrolę i zamiast robić coś lepiej lub gorzej, sam decydowałby o tym, czy robi coś lepiej lub gorzej. W tym momencie warto skorzystać z trudnego do przecenienia spostrzeżenia Perry'ego (1918), do którego tak nawiązuje Tolman (1995):

> Koncepcja, za którą się tu opowiadamy, głosi, ujmując rzecz pokrótce, że zawsze, gdy jakąś reakcję cechuje wyuczalność w odniesieniu do pewnego końcowego rezultatu - gdy reakcja ta ma skłonność do: a) przejawiania się w próbach i błędach oraz b) wybierania stopniowo lub nagle z tych prób i błędów posunięć bardziej efektywnych pod względem dążenia do tego końcowego rezultatu - reakcja ta wyraża i określa coś, co dla wygody nazywamy celem (*purpose*). Gdziekolwiek pojawia się taki zbiór faktów (a gdzie, z wyjątkiem najprostszych i najbardziej sztywnych tropizmów i odruchów, nie pojawia się on?), tam mamy obiektywnie wyrażone i określone to, co dogodnie jest nazwać celem.
>
> (s. 32, tamże)

Potrzebna jest systematyczna zmiana sposobu reagowania, czyli działania zależnego od sygnału środowiska, odpowiadająca coraz bardziej skutecznemu dążeniu do końcowego rezultatu i może pojawi się coś (zachowanie?) wyuczalne ze względu na cel. Sposób reagowania potraktujemy najogólniej jak tylko się da, jako funkcję określoną na zbiorze możliwych sygnałów stanu, nazywaną dalej polisą. Zmiana sposobu reagowania będzie więc zmianą polisy. To samo pojęcie występuje w teorii decyzji pod nazwą „reguły", ale termin „reguła" może się kiedyś przydać do opisu tego, co dzieje się w pudełku A.

Potrzebna jest jeszcze definicja końcowego rezultatu. Jedną z możliwości jest przypisanie końcowości wybranemu stanowi środowiska, który byłby wtedy ostatecznym celem interakcji. Takie rozwiązanie nie jest jednak wystarczająco elastyczne. Dlaczego agent miałby zawsze dążyć do jakiegoś pojedynczego stanu, albo nawet elementu należącego do jakiegoś wyróżnionego zbioru stanów?

Skoro już przyjęliśmy, że agent otrzymuje sygnał informujący go o tym, na ile jest (lokalnie) dobrze lub źle, najprostszym i zarazem naturalnym rozwiązaniem jest zdefiniowanie celu w oparciu o sumę nagród uzyskiwanych w trakcie interakcji. Ostatecznym celem agenta będzie więc maksymalizacja skumulowanej sumy nagród. Mamy już wszystkie elementy potrzebne do zdefiniowania wyuczalności ze względu na cel. Wyuczalność ze względu na cel to *poszukiwanie polisy maksymalizującej skumulowaną sumę nagród*. Jak dotąd nie przekraczamy granic teorii decyzji. Jak długo ta interakcja miałaby trwać?

Rozwiązaniem najbardziej niezobowiązującym okazuje się, być może paradoksalnie, założenie, że nieskończenie długo. Wynika to stąd, że proces trwający nieskończenie długo może nieźle udawać proces trwający skończenie długo, ale nie odwrotnie. Wystarczy uznać, że pewien zbiór stanów trwającego nieskończenie długo procesu ma to do siebie, że jak już raz się tam trafi, nie można wyjść i nic ciekawego w tych stanach się nie dzieje, to znaczy, nie ma tam żadnych nagród do odebrania. Takie stany nazwiemy absorbcyjnymi. Jeżeli zadanie nie może trwać w nieskończoność, nazwiemy je epizodycznym, w innym wypadku będzie to zadanie ciągłe.

Wypada w końcu zająć się zawartością pudełka S. Na początek najwygodniej będzie pracować ze stanami dyskretnymi, powiedzmy, że ze skończoną liczbą takich stanów. Dynamika środowiska będzie więc następstwem stanów w czasie. Co warto na początek założyć na temat tego następstwa? Żeby nie rozstrzygać za wcześnie zbyt wiele przyjmiemy, że dynamika środowiska ma charakter probabilistyczny. Wtedy środowisko deterministyczne będzie szczególnym przypadkiem tak rozumianego środowiska, przy założeniu, że rozkład prawdopodobieństwa określony na przyszłych stanach ze względu na aktualny stan i działanie, albo ze względu na krótszą lub dłuższą historię interakcji, będzie przyjmował tylko wartości zero lub jeden.

Środowisko, a nie agent, jest tutaj automatem probabilistycznym Putnama. Gdyby Putnam upierał się przy założeniu, że prawdopodobieństwo następnego stanu automatu zależy tylko od stanu poprzedniego, musiałby posługiwać się pojęciem stanu niezbyt użytecznym do opisu „mentalnych mechanizmów obliczeniowych", ponieważ działanie tych mechanizmów na pewno zależy od historii interakcji w sposób, który znacznie wygodniej jest analizować posługując się pojęciem historii interakcji. Putnam zdawał sobie doskonale sprawę z niewygodnego charakteru pojęć automatu probabilistycznego i maszyny Turinga i zastrzegał się między innymi (Putnam, 1967/1980), że na przykład nie chodzi o jeden stan, tylko o coś bardziej złożonego, że jego automat nie ma czegoś takiego jak taśma nieskończonej długości (maszyna Turinga ma), i tak dalej. Tak czy inaczej, automat probabilistyczny Putnama, a razem z nim inne podobne struktury, znajduje się teraz w pudełku S. Automat probabilistyczny czuje się lepiej w pudełku S niż w pudełku

A, ponieważ w pudełku S nikt nie wymaga od niego, żeby działał celowo.

Nawet gdybyśmy włożyli automat probabilistyczny do pudełka A na siłę, nie uda się nam go wyjąć z pudełka S. Automat Putnama różni się pod tym względem od klasycznej maszyny Turinga, która nie jest dobrym modelem ani środowiska, ani umysłu. Maszyna Turinga nadaje się natomiast doskonale do rozstrzygania co jest, a co nie jest efektywnie obliczalne. Automat Putnama jest z kolei pojęciem tak ogólnym, że jak sam Putnam zauważył, wszystko można potraktować jako automat probabilistyczny, tylko po co?

Żeby powiedzieć cokolwiek interesującego o czymkolwiek, w tym także o środowisku, trzeba przyjąć model lepiej określony i znacznie mniej ogólny od automatu probabilistycznego. Najczęściej przyjmowanym w sztucznej inteligencji założeniem[1] na temat procesów stochastycznych, z którymi takie lub inne rozwiązanie ma sobie radzić, jest założenie, zgodnie z którym aktualny stan jest wszystkim, od czego zależy prawdopodobieństwo pojawienia się stanów następnych. O takich procesach mówi się, że mają własność Markowa. Wielką zaletą własności Markowa jest znaczne uproszczenie obliczeń. Poważną wadą tego założenia jako założenia na temat środowiska jest to, że środowisko posiada własność Markowa tylko w niezwykle ograniczonym stopniu, w każdym razie środowisko z punktu widzenia innego niż ten, którym może się cieszyć demon Laplace'a. Kwestię konsekwencji porzucenia założenia o własności Markowa podejmę jeszcze w tej pracy, na razie jednak okaże się ono niezwykle przydatne. Mamy już wszystkie elementy potrzebne do zademonstrowania za pomocą środków czysto teoretycznych kilku przykładowych, nietrywialnych własności takich połączonych pudełek.

8.2 Uczenie się ze wzmocnieniem

Przedstawiona do tej pory w zarysie rama pojęciowa to zaproponowany przez Suttona i Barto (1998), tak zwany paradygmat uczenia się ze wzmocnieniem, określany dalej za pomocą skrótu RL (ang. *Reinforcement Learning*). Zdaniem tych autorów wiele, jeżeli nie wszystkie postaci inteligentnego działania (czy to organizmów żywych, czy też urządzeń albo programów komputerowych) można z powodzeniem rozłożyć na następujące części:

Agent to cokolwiek, co może wykonywać pewne działania i otrzymywać sygnały ze środowiska.

Środowisko to cokolwiek, co w odpowiedzi na działania agenta emituje pewien natychmiastowy sygnał stanu i natychmiastową nagrodę.

Polisa to funkcja, określająca prawdopodobieństwo wykonania działań ze względu na spostrzegane stany.

[1] Sutton i Barto (1998) twierdzą, że założenie to jest przyjmowane „w 90 procentach przypadków".

Funkcja nagrody określa nagrody otrzymywane za możliwe przejścia między stanami, albo za samo trafienie do określonych stanów.

Funkcja wartości dla każdego stanu określa oczekiwaną sumę nagród od momentu wejścia do tego stanu.

Zadanie to kompletna specyfikacja środowiska, wraz z sygnałem stanu i nagrody.

Sutton i Barto twierdzą, że rama pojęciowa uczenia się ze wzmocnieniem jest prawdopodobnie pierwszą w historii sztucznej inteligencji próbą formalizacji pojęcia systemu, który czegoś chce. Zaletą tej ramy pojęciowej jest jej (zachwycająca) ogólność, której nie należy pochopnie psuć przez bezrefleksyjne wprowadzanie dodatkowych założeń. Wyniki analiz teoretycznych dotyczących tak rozumianej interakcji będą się stosowały do dowolnej sytuacji, którą można za jej pomocą w przybliżeniu opisać. Zakres na razie tylko potencjalnie wyjaśnianych zjawisk jest więc co najmniej atrakcyjny.

Tym, co odróżnia agenta jako poszukiwacza polisy maksymalizującej wartość stanów lub działań od termostatu albo spadającego jabłka jest to, że w pewnych warunkach tak się składa, że termostat utrzymuje odpowiednią temperaturę a jabłko spada, nie można jednak powiedzieć, żeby termostaty i jabłka się o coś starały, ponieważ termostaty i jabłka nie zmieniają swojego sposobu działania zależnie od spostrzeganych konsekwencji działań. W przeciwieństwie do spadającego jabłka, termostat może sprawiać wrażenie jakby działał celowo, dlatego że został skonstruowany przez działającego celowo człowieka. Agent działa celowo, chociaż na razie nie wydaje się działać intencjonalnie. Można powiedzieć, że sygnały stanu i nagrody są spostrzeganymi przez agenta konsekwencjami działań dlatego, że poszukiwanie coraz lepszej polisy opiera się na wykorzystaniu tych sygnałów. To właśnie należy rozumieć przez bycie spostrzeganymi konsekwencjami działań. Podstawowe pojęcia teorii muszą być zdefiniowane teleologicznie i kontekstowo, a nie redukcyjnie.

Działania, którymi dysponuje agent, mogą być względnie proste i niskopoziomowe, albo względnie złożone. Mogą być również operacjami niewidocznymi dla obserwatora z zewnątrz, na przykład, skoro już przy tym jesteśmy, mogą być operacjami o charakterze „obliczeniowym", działaniami służącymi regulacji wewnętrznego środowiska agenta (na przykład stanu jego ciała), i tak dalej. Wydaje się, że na tym etapie analizy nie warto przesądzać za wiele o czasie, jaki zajmuje wykonanie określonych działan. Sygnał stanu można w pewnych sytuacjach potraktować jako rodzaj informacji percepcyjnej, *wewnętrznym* sygnałem stanu może być rezultat wcześniejszego przetwarzania informacji płynącej ze środowiska.

Sygnał nagrody to natychmiastowa informacja mówiąca o tym, jak dobrze lub źle w danej chwili jest znajdować się w danym stanie, albo zastosować dane działanie w danym stanie, albo dokonać przejścia między stanami za pomocą określonych działań. Wygodnie jest zdefiniować funkcję nagrody dla przejść między stanami. Za pomocą tak zdefiniowanej funkcji można wyrazić użyteczność samych tylko stanów, przypisując jed-

nakowe nagrody wszystkim działaniom prowadzącym do tych stanów, ale można też wyrazić zróżnicowany koszt działań. Zwykle dla uproszczenia zakłada się, że funkcja nagrody przyjmuje wartości ze zbioru liczb rzeczywistych. Chociaż pewne wnioski wydają się być w znacznym stopniu niezależne od tego, jak wyrażona będzie funkcja nagrody, czasem może zajść potrzeba wprowadzenia bardziej złożonego sygnału, na przykład wektora rzeczywistego, jednak nawet skalarna nagroda pozwala na uzyskanie pouczających rezultatów. Często natomiast nieodzowne może być przedstawienie w bardziej złożonej postaci sygnału stanu - jako wektora, albo jako innej struktury danych (na przykład, wielomodalny sygnał stanu, odpowiadający różnym receptorom).

Nic nie stoi na przeszkodzie, aby agent, na podstawie wcześniejszej historii interakcji, tworzył własne reprezentacje nagród i stanów środowiska. Faktycznie, coś, co spełnia funkcję reprezentowania informacji na temat historii wcześniejszej interakcji staje się konieczne, żeby w ogóle było możliwe względnie skuteczne maksymalizowanie oczekiwanej sumy nagród w środowiskach o pewnym poziomie złożoności, między innymi w częściowo obserwowalnych procesach decyzyjnych Markowa. Takie elementy znajdowałyby się już jednak nie po stronie środowiska, ale agenta, ponieważ granica między agentem i środowiskiem to granica kontroli, a nie ciała. Na poziomie komputacyjnym można powiedzieć, że ciało agenta jest tą częścią środowiska, która jest zawsze stosunkowo łatwo dostępna.

Stan, w którym aktualnie znajduje się agent, będzie w ogólnym przypadku funkcją jego wcześniejszych działań, ale też dynamiki samego środowiska, częściowo niezależnej od działań agenta. W zależności od problemu badawczego granica między agentem i środowiskiem może czasem przebiegać w innym miejscu niż granica między ciałem i jego otoczeniem („bliżej samego agenta", jeżeli zgodzimy się na użycie tak czarująco wieloznacznego zwrotu). Na przykład, sygnał na podstawie którego podejmowana jest decyzja w modelu dyfuzyjnym jest częścią wewnętrznego środowiska agenta ze względu na pewien opis w kategoriach celowej interakcji. Zakłada się, że agent sprawuje nad tym sygnałem kontrolę najwyżej pośrednio, przez to, że może gdzie indziej skierować swoje receptory, albo na coś innego zwrócić uwagę. Przyjmując tą interpretację, psychologowie poznawczy od początku zajmowali się jednak teorią pewnej skromnej części środowiska, tyle że nazywali ją teorią ograniczeń systemu poznawczego.

Polisa jest funkcją określająca, jakie działania wybierze agent zależnie od otrzymywanych sygnałów stanu. Zwykle polisa będzie probabilistyczna. Jeżeli agent tworzy na własny użytek jakiś sygnał stanu, na przykład zawierający informacje o dotychczasowej historii interakcji, i takiej reprezentacji stanu środowiska używa do selekcji działań, polisa będzie funkcją określoną na historiach otrzymanych sygnałów stanu.

Funkcja wartości mówi o tym, jaka jest oczekiwana skumulowana nagroda dla danego stanu, albo przejścia między danymi stanami na skutek wybrania określonych działań. Celem agenta jest maksymalizacja sumy nagród uzyskanej w trakcie interakcji ze środowiskiem, funkcja wartości jest więc z definicji jedynym istotnym kryterium oceny stanów, działań i polis. Ponieważ to, jakie agent otrzyma nagrody w przyszłości, zależy

zarówno od samego środowiska, jak i od tego, jakie agent będzie wybierał dalej działania, wartość może być określona tylko ze względu na polisę. Inaczej mówiąc, stany lub przejścia można uznać za dobre lub złe w znaczeniu funkcji wartości tylko ze względu na to, co agent będzie robił w dalszym przebiegu interakcji, rozpoczynającym się od danego stanu lub przejścia. Podczas gdy wartość funkcji nagrody dla każdego odwiedzanego stanu jest bezpośrednio dostępna, funkcja wartości musi najczęściej dopiero zostać, wprost lub nie wprost, oszacowana. Sutton i Barto dostrzegli głęboki psychologiczny sens pojęć nagrody i wartości pisząc, że to pierwsze jest jak przyjemność i ból, a to drugie jak zadowolenie i niezadowolenie (s. 7, Sutton i Barto, 1998).

Funkcja wartości będzie szacowana wprost, jeżeli jakiś mechanizm obliczeniowy będzie szacował funkcję wartości. Problem szacowania funkcji wartości może być jednak rozwiązywany nawet wtedy, gdy nie da się wyróżnić takiego mechanizmu, tak jak można powiedzieć, że agent posiada reprezentację czegoś, nawet gdy ta reprezentacja nie jest nigdzie „zlokalizowana".

Granice między funkcjami wyodrębnionymi komputacyjnie nie muszą się pokrywać z gracami między mechanizmami lub podsystemami obliczeniowymi. Co więcej, na pewno istnieje więcej niż jeden opis komputacyjnej struktury rozwiązania, tak jak istnieje więcej niż jedna (na przykład, oparta na bayesowskiej teorii decyzji, klasycznej teorii decyzji, teorii złożoności algorytmicznej, albo jeszcze innej) racjonalna teoria rozwiązania problemu uczenia się ze wzmocnieniem. W każdym momencie za bliższą prawdy wypada uznać tą teorię, która ma większą moc wyjaśniającą. Jak ujął to Pylyshyn, „tak postępuje każda dojrzała nauka".

Wyobraźmy sobie, że natrafiliśmy na takie dwa, wchodzące ze sobą w interakcję systemy, że ze względu na pewien opis w kategoriach działań, funkcji nagrody i funkcji sygnału stanu jeden z tych systemów działa tak, że w przybliżeniu optymalnie maksymalizuje wartość odwiedzanych stanów. Będziemy wtedy zmuszeni przyznać, że ten system działa celowo, chociaż niekoniecznie intencjonalnie. Parafrazując Tolmana, gdziekolwiek pojawia się taki zbiór faktów, tam mamy obiektywnie wyrażone i określone to, co dogodnie jest nazwać zachowywaniem się agenta w środowisku. Jednocześnie będziemy mogli stwierdzić, że funkcje nagrody i wartości w przybliżeniu poprawnie reprezentują cel tego agenta.

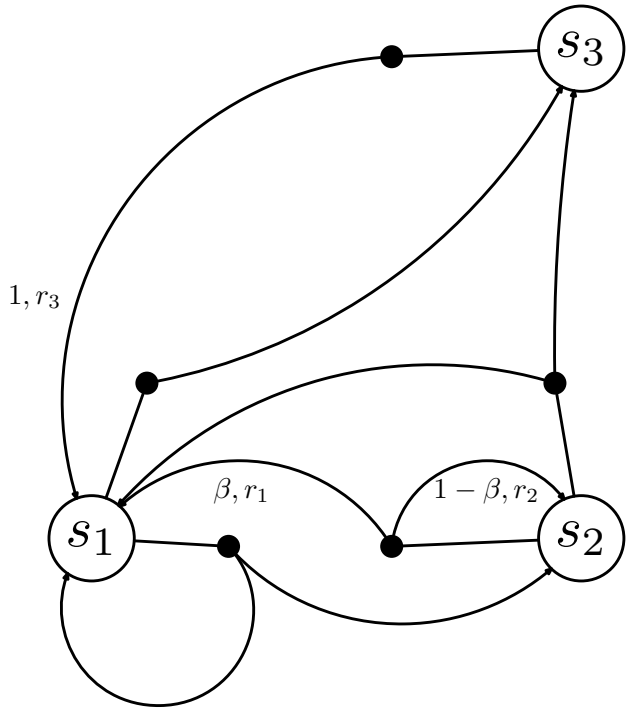
Przynajmniej dla niektórych wersji problemu uczenia się ze wzmocnieniem problem identyfikacji funkcji nagrody na podstawie obserwowanego zachowania się optymalnego agenta okazuje się rozwiązywalny (Inverse Reinforcement Learning, Ng i Russell, 2000). W ten sposób kryterium przybliżonej racjonalności pozwala częściowo rozwiązać problem nieznanego rzeczywistego celu, który słusznie zaniepokoił Simona. Im bardziej nieoczywiste ze względu na zakładany model środowiska, sygnałów i działań będą konsekwencje założenia racjonalności, tym bardziej te wnioski będą uzasadnione, dlatego że tym trudniej będzie wtedy o alternatywne wyjaśnienie. Im bardziej odległe od optymalnego będzie rzeczywiste działanie, tym bardziej wątpliwe będą wnioski. Tego rodzaju niepewności nie sposób uniknąć w żadnej nauce empirycznej. Z zarzutem arbitralności

opisu wchodzących w interakcję systemów w kategoriach funkcji nagrody, sygnału stanu i działań można sobie dodatkowo radzić uzasadniając przyjęty opis na przykład względami biologicznymi. Nic nas zresztą nie zmusza, aby doszukiwać się zachowania tam, gdzie to jest mało wiarygodne.

Jeżeli teraz umieścimy tego agenta w innym środowisku i okaże się, że agent tym razem nawet w przybliżeniu nie maksymalizuje wartości odwiedzanych stanów, będziemy zmuszeni stwierdzić, że system ten przestał być agentem i przestał się zachowywać. Nie ma czegoś takiego, jak zachowanie się niezależnie od środowiska, w którym się rozgrywa. Jeżeli ciało agenta przetrwa ten eksperyment, system nadal będzie sprawny albo żywy, ale w momencie zmiany przejdzie w stan, który można by nazwać psychologiczną hibernacją.

Przyjęta definicja zachowania nie określa czym jest zachowanie w kategoriach mechanizmu obliczeniowego agenta, ale w kategoriach warunków, jakie musi spełniać bliżej nieokreślony mechanizm obliczeniowy w bliżej nieokreślonym środowisku. Poszukiwanie polisy maksymalizującej oczekiwaną sumę nagród nie jest definicją żadnego konkretnego algorytmu ani mechanizmu obliczeniowego. W szczególności, optymalne rozwiązanie tego problemu nie musi być nawet efektywnie obliczalne, żeby dało się zdefiniować i było nietrywialne (Hutter, 2004). Na poziomie komputacyjnym, czy też, jak to nazywa Pylyshyn, semantycznym, obowiązują inne zasady, niż na poziomie obliczeniowym.

Zanim jeszcze wprowadzone zostaną formalne definicje podstawowych pojęć, wypada przedstawić przykład ilustrujący niektóre kluczowe kwestie. Poniżej znajduje się diagram konkretnego problemu uczenia się ze wzmocnieniem.

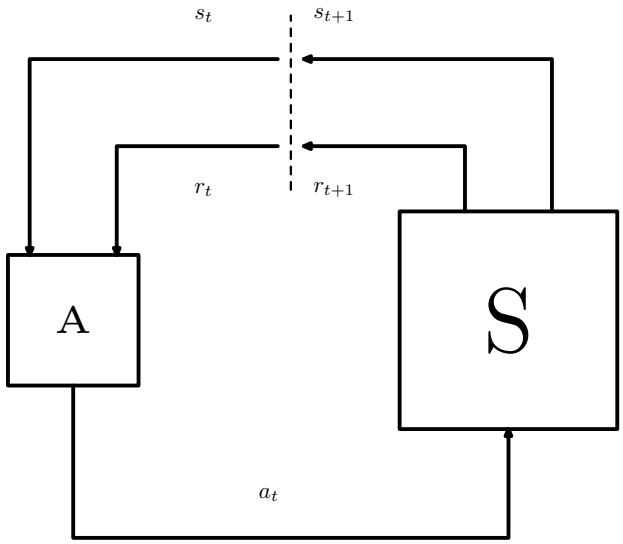


Rysunek 8.3: Graf przykładowego problemu uczenia się ze wzmocnieniem

Przyjęło się oznaczać węzły stanów za pomocą okręgów, węzły działań za pomocą mniejszych kółek, a przejścia, wynikające z wybrania określonych działań w określonych stanach, za pomocą strzałek. W tym przykładzie funkcja nagrody jest określona dla przejść. Na środowisko składają się trzy stany, s_1, s_2 i s_3. Znajdując się w stanie s_2 agent może wybrać jedno z dwóch działań, prowadzące do stanów s_1, s_3 lub do stanu wyjściowego. Prawdopodobieństwa związane z krawędziami grafu wychodzącymi od każdego z węzłów działań muszą się oczywiście sumować do 1. Przedstawione na diagramie zadanie składa się tylko z trzech stanów i czterech działań, ale nawet dla tak prostego i doskonale widocznego środowiska nie jest łatwo powiedzieć, co trzeba robić, żeby osiągnąć cel.

8.2.1 Formalna charakterystyka ramy pojęciowej uczenia się ze wzmocnieniem

Formalna charakterystyka ramy pojęciowej RL pozwoli mi w dalszej części pracy zachować zwięzłość wywodu i przyczyni się do uniknięcia nieporozumień, które mogłyby się pojawić, gdybym próbował wyrazić wszystko za pomocą mniej formalnego, a przez to być może bardziej intuicyjnego języka. Przypomnę raz jeszcze ogólny schemat interakcji agenta ze środowiskiem.



Rysunek 8.4: Diagram uczenia się ze wzmocnieniem

Formalnie, problem uczenia się ze wzmocnieniem można rozłożyć na następujące elementy:

S	zbiór stanów
$A(s)$	zbiór działań dostępnych w stanie $s \in S$
$\pi_t \colon S \times A \mapsto [0, 1]$	polisa
$R_t \colon S \times S \times A \mapsto R$	oczekiwana skumulowana nagroda od momentu t, czyli $r_{t+1} + r_{t+2} + \ldots$
$V^\pi(s)$	oczekiwana wartość stanu s ze względu na polisę π

Tabela 8.1: Formalna charakterystyka problemu uczenia się ze wzmocnieniem

Funkcję $V^\pi(s)$ zastępuje się czasem dla wygody równoważną funkcją $Q^\pi(s, a)$, która mówi o wartości podjęcia działania a w stanie s, a następnie postępowania zgodnie z polisą π.

8.3 Przetarg między eksploracją i eksploatacją

W RL zakłada się zwykle, że w trakcie interakcji ze środowiskiem agent nie dowiaduje się, które działanie jest optymalne w danym momencie, tylko na ile wybrane przez niego wcześniej działanie okazało się lokalnie dobre. Nie otrzymuje więc instrukcji, jak należy działać, czyli uczenie się nie ma charakteru nadzorowanego (przebiega bez nauczyciela).

Ponieważ sama informacja o tym, na ile dane działanie okazało się lokalnie dobre, nie mówi nic na temat tego, czy jakieś inne działanie nie byłoby lepsze na dłuższą metę, w każdym kroku interakcji agent zmuszony jest rozwiązać pewien dylemat.

Jednym z najprostszych przypadków problemu sekwencyjnego podejmowania decyzji, w którym pojawiają się już poważne konsekwencje niepewności wynikającej z probabilistycznego charakteru zadania, jest tak zwane zadanie n-rękiego bandyty. Zadanie to przypomina znaną z kasyn grę z automatem nazywanym jednorękim bandytą, z tą różnicą, że liczba ramion może być większa niż jeden.

W uproszczeniu, grę w jednorękiego bandytę można scharakteryzować następująco: każde pociągnięcie za ramię automatu prowadzi albo do wygranej, albo do przegranej. Jeżeli gracz wygrywa, otrzymuje pewną sumę pieniędzy. Jeżeli przegrywa, nie otrzymuje nic, traci więc sumę, którą musiał wrzucić do automatu żeby w ogóle zagrać. Możemy założyć, że każda taka pojedyncza rozgrywka jest rodzajem powtarzalnego eksperymentu, którego rezultat jest zmienną losową o określonym, stałym dla wszystkich rozgrywek rozkładzie.

Zadanie n-rękiego bandyty jest znacznie prostsze niż to, które przedstawiłem wcześniej na diagramie, ponieważ konsekwencje działań w tym zadaniu nie zależą od stanu środowiska. Zadania tego rodzaju nazywa się nieasocjacyjnymi. Przez cały czas trwania interakcji agent dysponuje n różnymi działaniami, a wybór każdego z tych działań prowadzi do uzyskania natychmiastowej nagrody. Niech dla każdego działania i momentu interakcji taka nagroda będzie wartością zmiennej losowej o rozkładzie normalnym, wariancji równej 1 i średniej μ_i, zależnej od wybranego działania. Po otrzymaniu nagrody agent pozostaje w stanie początkowym i może wykonać kolejne działanie.

Załóżmy dalej, że agent ma do dyspozycji trzy różne działania i na początku interakcji, w pierwszych trzech krokach, z sobie tylko znanych powodów wybrał kolejno każde z tych działań. W efekcie agent otrzymał następujące nagrody:

$$r_1 = 1,\ r_2 = 2,\ r_3 = 3$$

gdzie r_i oznacza nagrodę otrzymaną po wykonaniu działania a_i. W tym momencie działaniem najbardziej obiecującym wydaje się działanie a_3, załóżmy więc, że agent wybiera ponownie a_3 i otrzymuje nagrodę równą $2{,}5$. I co teraz? Można by przypuszczać, że strategia wyboru takiego działania, które do tej pory dostarczało największej nagrody, jest najlepszym albo co najmniej dobrym rozwiązaniem tego problemu.

Żeby zbadać tą kwestię dokładniej należy ustalić, co to właściwie znaczy, że jakieś działanie dostarczało do tej pory największą nagrodę. Agent otrzymuje jedynie informację natychmiastową, więc żeby w ogóle mógł korzystać z historii interakcji, musi tą historię w jakiś sposób pamiętać. Mimo pojawienia się tego swojsko brzmiącego terminu, nie ma jeszcze najmniejszych powodów, aby zaglądać do pudełka A, ponieważ nie wiemy co dokładnie pudełko A ma robić. Jako że zadanie ma charakter probabilistyczny, celem agenta jest maksymalizacja *oczekiwanej* sumy nagród. Żeby rozwiązać problem

n-rękiego bandyty, agent musi więc za pomocą takich lub innych środków szacować wartość działań. Szacowanie można tu rozumieć jako „czystą funkcję", która odgrywa „czysto funkcjonalną rolę" w scharakteryzowanym „czysto funkcjonalnie" procesie celowej interakcji ze środowiskiem. Zgoda na prymat funkcji nad mechanizmem obliczeniowym wymusza włożenie mechanizmu do pudełka S i pozostawienie zawartości pudełka A w spokoju tak długo, jak tylko się da. Parafrazując cytowanego w rozdziale trzecim Luce'a, takie modelowanie matematyczne byłoby zastosowaniem strategii ostrożnego otwierania czarnych skrzynek.

Szacowanie rozumiane jako element rozwiązania *konkretnego* zadania nie jest już do końca „czystą funkcją". Na przykład, dla określonego zadania, takiego jak konkretne zadanie trójrękiego bandyty, stwierdzenie, że agent w ten lub inny sposób szacuje wartość działań nakłada ograniczenia na zbiór dopuszczalnych mechanizmów obliczeniowych. Tylko niezwykle mały (co nie znaczy, że skończony) podzbiór możliwych mechanizmów działających w pudełku A może w przybliżeniu realizować tą funkcję dla określonego zadania, a jeszcze mniejszy podzbiór mechanizmów może realizować tą funkcję optymalnie. Nadal jednak szacowanie jako takie jest czymś, co rozgrywa się ponad wejściami, wyjściami, mechanizmem wewnątrz agenta i środowiskiem. Nie rozważamy przecież tylko konkretnych środowisk, ale ich klasy.

Możemy uogólniać założenia dotyczące zadania, sprawiając tym samym, że to, czego dowiemy się o rozmaitych funkcjach, będzie mniej zależne od specyficznych własności konkretnych zadań, ale nawet analiza konkretnych, prostych przypadków pozwala wyprowadzić wnioski zarazem ogólne i nietrywialne. Udało się to Shepardowi, Griffithsowi i Tenenbaumowi. Poszukiwanie rozwiązań optymalnych przydaje się często do wyprowadzania nietrywialnych i ogólnych wniosków na temat jakichkolwiek rozwiązań ogólnych klas zadań, dzięki temu, że pozwala dowiedzieć się czegoś o warunkach, jakie muszą być spełnione, aby rozwiązanie było optymalne choćby w przybliżeniu. Poszukiwanie rozwiązań optymalnych w znacznie większym stopniu wymusza zrozumienie funkcji niż poszukiwanie rozwiązań sprawiających wrażenie „wystarczająco skutecznych".

Strategię polegającą na wyborze działania o największej wartości określa się jako zachłanną, a działanie, dla którego szacowana wartość jest największa, działaniem zachłannym. Ponieważ interesuje nas maksymalizacja oczekiwanej sumy nagród na dłuższą, a nie krótszą metę, natychmiastowa nagroda jako taka nie ma zbyt dużego znaczenia. To, co ma dla nas znaczenie kluczowe, to oczekiwana suma nagród dla każdego z działań, czyli jego wartość. Jeżeli nagrody dla każdego z działań pochodzą ze stacjonarnych, czyli nie ulegających zmianie w trakcie przebiegu interakcji rozkładów normalnych, najlepszym oszacowaniem wartości działań będzie średnia z nagród, otrzymanych natychmiast po wybraniu danego działania, w trakcie całej wcześniejszej interakcji. Gdyby zadanie było asocjacyjne, najlepszym oszacowaniem wartości byłaby średnia ze wszystkich nagród otrzymanych od momentu wybrania danego działania w danym stanie i podążania zgodnie z daną polisą.

Takie zachłanne rozwiązanie nie jest optymalne dla probabilistycznego zadania n-

rękiego bandyty. Nawet jeżeli rozkład nagród dla każdego działania pozostaje niezmienny, najlepsze oszacowanie wartości działania jakim może dysponować agent (średnia) zawsze będzie obarczone pewnym błędem. Inaczej mówiąc, agent nigdy nie może mieć pewności, że jakieś inne działanie, które na podstawie dotychczasowej interakcji może wydawać się nieoptymalne, nie jest faktycznie lepsze niż to, dla którego oszacowanie wartości jest największe. Zawsze istnieje ryzyko, że ocena relatywnej wartości działań wynikająca z dotychczasowych oszacowań jest rezultatem błędu próby. Bez konieczności dokładnej analizy ogólnych klas zadań od razu jasne jest, że ten sam problem dotyczy nie tylko n-rękiego bandyty, w szczególności na pewno dotyczy przeważającej większości rzeczywistych problemów, które nieustannie rozwiązują ludzie.

Jednym z prostszych rozwiązań tego problemu jest przyjęcie strategii, zgodnie z którą zwykle wybiera się działanie zachłanne, ale od czasu do czasu wybiera się jakieś inne, dowolne działanie. Wybór działania innego niż zachłanne można nazwać „eksploracją", a wybór działania zachłannego „eksploatacją". W ten sposób uzyskujemy bardzo ogólną (choć błędną) definicję dwóch funkcji, odgrywających zasadniczą rolę w procesie celowej interakcji ze środowiskiem. Zanim przedstawię zarysy nieco udoskonalonej teorii eksploracji i eksploatacji, pozwolę sobie, ze względów tylko i wyłącznie retorycznych, przytoczyć kilka fragmentów z rozprawy doktorskiej Pisuli (2003), dotyczącej zachowań eksploracyjnych u zwierząt.

8.3.1 Eksploracja jako przedmiot badań teoretycznych i empirycznych w psychologii - studium przypadku

> (...) nie próbowaliśmy przyjmować definicji ciekawości i zachowań eksploracyjnych, co do której zgodziliby się wszyscy współautorzy. Biorąc pod uwagę stan badań oraz teorie ciekawości i zachowań eksploracyjnych, sądziliśmy, że byłoby to niewłaściwe.
>
> (s. 3, Keller, Schneider i Henderson, 1994)

Dlaczego Keller, Schneider i Henderson uznali za niewłaściwe próby sformułowania przekonującej definicji eksploracji (i ciekawości) i dlaczego stwierdzili, że nie pozwala na to stan *badań* pozostanie dla mnie tajemnicą. Sam Pisula, nie ułatwiając sobie zadania, pisze dalej tak:

> W istocie, zachowania eksploracyjne to dziedzina wymykająca się prostym klasyfikacjom. Z jednej strony mamy tu do czynienia z zachowaniami elementarnymi, będącymi prostym przedłużeniem reakcji orientacyjnej, z drugiej zaś - jak chciał D. E. Berlyne (1963) - z ciekawością poznawczą, właściwą człowiekowi. Tak złożony obszar badawczy musi nastręczać wiele trudności definicyjnych i klasyfikacyjnych.

(s. 27, tamże)

Tak marnie zdefiniowany przedmiot badań musi nastręczać wiele wszelkiego rodzaju trudności. Skąd wiadomo, które zachowania są eksploracyjne, a które nie? Jeżeli nie wiadomo, coś trzeba założyć, żeby w ogóle wyodrębnić jakiś obszar badawczy. Co to znaczy, że niektóre z tych zachowań są „prostymi przedłużeniami reakcji orientacyjnej"? Dlaczego od razu wrzucać do jednego worka zachowania eksploracyjne, ciekawość poznawczą i zabawę, jak czyni to wielu autorów cytowanych przez Pisulę?

Dalej Pisula twierdzi, że „biorąc za punkt wyjścia teorię poziomów integracji, analizę zachowań eksploracyjnych należy zacząć od poziomu opisu samego zachowania" (s. 27, tamże). Okazuje się, że zachowania takie (nadal nie wiadomo skąd Pisula wie, które to są zachowania) „cechuje duża różnorodność". Na następnej stronie mamy okazję zapoznać się z rezultatem zastosowania jednego z częściej stosowanych narzędzi analizy teoretycznej w psychologii, to jest z zestawieniem „możliwych ujęć, w jakich poszczególne dyscypliny naukowe traktują zachowania eksploracyjne" w postaci tabelki. Z perspektywy psychologii motywacji problemem badawczym będzie między innymi „mechanizm motywowania zachowań eksploracyjnych, rola stymulacji zewnętrznej, procesów wewnętrznych, ich interakcji i tym podobnych", z perspektywy psychologii różnic indywidualnych będzie to „ekspresja postulowanych cech temperamentu: aktywności, poszukiwania wrażeń, zapotrzebowania na stymulację sensoryczną, a także pośrednio: reaktywności, wrażliwości i tym podobnych". Pojawiają się również problemy badawcze ujęte z perspektywy psychofarmakologii i etologii stosowanej.

Jednym ze źródeł trudności, z jakimi musiał zmierzyć się autor, mogła być imponująca znajomość literatury psychologicznej. Berlyne'owi (1960) zawdzięcza Pisula pojęcie „ogólnej eksploracji ruchowej, traktowanej jako zachowanie nieukierunkowane na określony obiekt", od Birke'go i Archera (1983) dowiedział się, że odmianą tejże ogólnej eksploracji ruchowej jest „patrolowanie, polegające na przemierzaniu przestrzeni dobrze znanej", w pismach Butlera (1953) odnalazł pojęcie „eksploracji percepcyjnej", za pomocą którego autor ten opisywał „zachowanie rezusów, polegające na wpatrywaniu się w obiekt", przy czym „stwierdził on, iż możliwość patrzenia na określone obiekty ma znaczną wartość nagradzającą". Znowu Berlyne'owi zawdzięcza Pisula „reakcje badawcze, określane (...) jako eksploracja ukierunkowana", które są „rozszerzeniem eksploracji ruchowej i percepcyjnej". Bycie „rozszerzeniem", „podstawą czegoś" i „bycie związanym z" to popularne definicyjne własności desygnatów pojawiających się w niepokojącym tempie psychologicznych „terminów technicznych".

Dalej jest jeszcze wiele zagadkowych klasyfikacji, tabelek, wykresów i diagramów. Autor dokonuje heroicznych wysiłków, aby scalić to wszystko za pomocą Feiblemana (1954) koncepcji poziomów integracji. Zrozumienie tego „sposobu analizy rzeczywistości" okazuje się wymagać „odwołania do ogólnych stwierdzeń teorii poziomów integracji z filozofii nauki" (s. 22, tamże). Jeden z recenzentów tej pracy nie mylił się pisząc, że reprezentuje ona „wysoki poziom naukowy", dokonał jednak skrótu myślowego. Wysoki

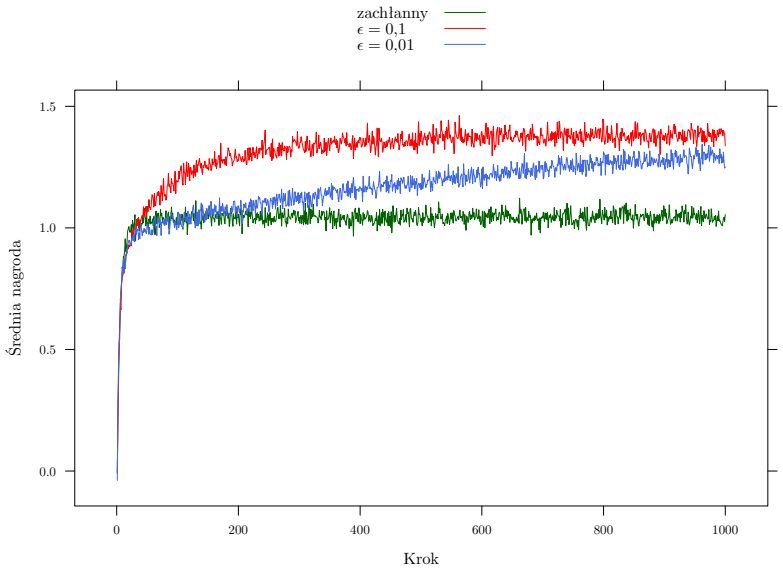
jak na psychologię.

8.3.2 Analizy rozwiązań zachłannych i epsilon-zachłannych ciąg dalszy

Przyjmijmy, że w każdej próbie prawdopodobieństwo wybrania działania eksploracyjnego jest stałe i oznaczmy je przez ϵ. W ten sposób uzyskujemy klasę rozwiązań ϵ-zachłannych. Jak wiadomo, wraz ze wzrostem liczby prób średnia z próby dąży do średniej w populacji. Jeżeli tylko $\epsilon > 0$, liczba przypadków wykonania każdego działania będzie dążyła do nieskończoności wraz ze wzrostem liczby prób, o ile oczywiście zadanie jest ciągłe. Takie rozwiązanie musi więc prędzej czy później doprowadzić do uzyskania oszacowań wartości działań wystarczająco dobrych, aby na ich podstawie możliwe było dokonywanie wyborów optymalnych.

Poniżej przedstawiłem wyniki symulacji zadania 10-rękiego bandyty dla agentów zachłannego i dwóch różnych agentów ϵ-zachłannych, z wartościami ϵ równymi $0{,}1$ i $0{,}01$. Przebieg symulacji był następujący[2]: na początku każdej serii rozgrywek losowano z rozkładu normalnego o średniej 0 i odchyleniu standardowym 1 średnie rozkładu nagród dla każdego z ramion (rozkład normalny o wariancji równej 1). Każdy z trzech symulowanych agentów wykonał 1000 kroków w ramach każdego z 2000 wylosowanych zadań. Wartości działań były szacowane za pomocą średniej z uzyskanych nagród, przy czym początkowa wartość oszacowania dla każdego z działań wynosiła 0. Na wykresie poniżej widać uśrednioną po zadaniach nagrodę uzyskaną w kolejnych krokach interakcji. Wyniki dla proporcji wyborów optymalnych są jakościowo takie same.

[2]Wyniki dokładnie takiej samej symulacji omawiają Sutton i Barto (s. 28-29, Sutton i Barto, 1998)



Rysunek 8.5: Rezultaty symulacji zadania 10-rękiego bandyty dla agentów zachłannego i ϵ-zachłannych

Rozwiązanie zachłanne okazuje się tutaj rozwiązaniem najgorszym. Mimo, że bardzo prymitywna, strategia eksploracji pozwala obu agentom ϵ-zachłannym uzyskiwać systematycznie lepsze wyniki. Jasne jest również znaczenie prawdopodobieństwa wyboru działania eksploracyjnego. Im to prawdopodobieństwo jest wyższe, tym szybciej oszacowania są korygowane i szybciej rośnie prawdopodobieństwo wybrania działania optymalnego. Z drugiej strony, prawdopodobieństwo wyboru działania eksploracyjnego decyduje też o tym, jakie będzie maksymalne osiągalne przez danego agenta prawdopodobieństwo wyboru działania optymalnego. Pierwszy ogólny, choć może jeszcze niezbyt pouczający wniosek brzmi - generalnie opłaca się eksplorować na początku interakcji, ale z czasem warto eksplorować coraz mniej. Coś takiego niemal nieustannie robią ludzie.

Na podstawie przedstawionej wcześniej charakterystyki zadania i uzyskanych wyników możemy natychmiast sformułować pewne wnioski na temat tego, jak przetarg między eksploracją i eksploatacją powinien być rozwiązywany. Eksploracja będzie tym bardziej potrzebna, im większa będzie niepewność związana z oszacowaniami wartości.

Ta niepewność jest funkcją zarówno rozkładu uzyskiwanych nagród jak i długości trwania interakcji. Jeżeli rozkład nagród związanych z każdym działaniem będzie miał małą wariancję, nawet początkowe oszacowania będą względnie dobrym przybliżeniem prawdziwej wartości i eksploracja nie będzie tak bardzo potrzebna. Jeżeli wariancja nagród będzie duża, nawet znaczna liczba kroków może nie wystarczyć do uzyskania rozsądnych przybliżeń wartości. Korzyści płynące z eksploracji będą też zależały od czasu trwania interakcji. Eksploracja będzie szczególnie ważna na początku, ponieważ agent będzie wtedy dysponował stosunkowo małą liczbą obserwacji. Jeżeli interakcja ma trwać krótko, eksploracja może nie zdążyć się zwrócić, wobec czego, nawet w sytuacji dużej zmienności obserwacji, nacisk być może powinien być w większym stopniu położony na eksploatację. W skrajnym przypadku, gdy interakcja ma się zakończyć w następnym kroku, nie pozostaje nic innego jak eksploatować. Im dłuższy będzie czas interakcji, tym większe będą koszty błędnego oszacowania wartości. Wszystkie te ogólne wnioski można wyprowadzić poddając nieformalnej analizie teoretycznej skrajnie wyidealizowany model celowej interakcji ze środowiskiem, jakim jest zadanie n-rękiego bandyty.

Uzyskane w ten sposób wnioski można zastosować w psychologii do dowolnego procesu, który podejrzewamy o celowość, a który zdaje się charakteryzować malejącą zmiennością. Każdy taki proces będzie szczególnym przypadkiem uczenia się w warunkach niepewności. Taki charakter ma na przykład proces automatyzacji, albo proces twórczego rozwiązywania problemów, ale nie można też z góry wykluczyć, że na podobnej zasadzie działają bardziej niskopoziomowe procesy, takie jak uwaga selektywna, percepcja, albo wybór wartości progu decyzyjnego w procesie wyboru ze skończonej liczby alternatyw. Trzeba tylko pamiętać, że granica między agentem i środowiskiem jest granicą kontroli i można ją przesuwać zależnie od podjętego problemu badawczego.

Związek z twórczością narzuca się ze względu na oczywiste znaczenie działań o charakterze eksploracyjnym, związek z automatyzacją nasuwa się ze względu na element uczenia się w warunkach niepewności. Dzięki temu, że proces celowej interakcji ze środowiskiem został sformułowany ogólnie, możliwe jest zastosowanie tej samej ramy pojęciowej do wielu, powierzchownie różnych procesów, które zaczynają się teraz jawić jako szczególne przypadki podstawowej funkcji kontroli, co wygląda na realizację jednego z postulatów zawartych w cytowanym wcześniej fragmencie pochodzącym od Angella. Jednocześnie można żywić uzasadnioną nadzieję, że ta ogólność pozwoli na transfer wniosków uzyskanych w badaniach nad jedną klasą procesów czy zjawisk na inne.

Mogłoby się wydawać, że teraz należałoby po prostu „sprawdzić, jak jest naprawdę". Nic nie stoi na przeszkodzie, aby postawić bardziej uszczegółowione pytania i i zaryzykować hipotezy o charakterze empirycznym. Prawdopodobnie tak postąpiłby psycholog poznawczy, dla którego całe dotychczasowe rozumowanie brzmi względnie przekonująco. Próby empirycznej oceny takich lub innych uszczegółowionych hipotez byłyby jednak przedwczesne, *właśnie dlatego*, że nie wiemy zbyt wiele na temat przetargu między eksploracją i eksploatacją. W zaprezentowanej wersji rozwiązania tego problemu znajdują się poważne luki. Można to stwierdzić nie dlatego, że wiemy gdzie te luki się znaj-

dują, ale dlatego, że jak dotąd nie znamy rozwiązania optymalnego i nie wiemy gdzie go szukać. Dopiero, gdy uda się ustalić coś na temat rozwiązania optymalnego, funkcja eksploracji i eksploatacji ujawni się w całej okazałości, chociaż tylko ze względu na przyjętą metateorię zadania (klasyczną teorię decyzji, bayesowską, lub inną).

8.3.3 Poszukiwanie rozwiązania optymalnego

Aby zdiagnozować, na czym polegają braki teorii w jej dotychczasowej postaci, warto spróbować ustalić, w jakim właściwie znaczeniu uzyskaliśmy częściowe choćby rozwiązanie omawianego problemu. Analizę tą dobrze jest zacząć od uwag na temat najłatwiej dostrzegalnych właściwości rozwiązań ϵ-zachłannych.

Zauważmy najpierw, że proponowane rozwiązanie można przedstawić jako model z jednym wolnym parametrem, to jest prawdopodobieństwem ϵ wyboru działania innego niż zachłanne. Żeby taki model z jednym wolnym parametrem można było zastosować wprost do analizy wyników badań, musimy założyć, że pomijając wartość tego parametru, wszystkie pozostałe elementy (środowisko, dostępne działania i sposób uczenia się) są znane. Założenie to, choć absurdalne, przyjmę na użytek rozważań.

W kontekście tego, co napisałem w poprzednich rozdziałach na temat ilościowej i jakościowej oceny modeli matematycznych, hipoteza zawierająca tylko jeden wolny parametr może się wydawać mało elastyczna, a sam parametr przypuszczalnie w miarę czytelnie interpretowalny. Ten jeden wolny parametr ma jednak dosyć szczególny status metateoretyczny. Wprowadzenie go pozwala mianowicie przynajmniej częściowo rozwiązać problem przetargu *z punktu widzenia badacza, który dysponuje wiedzą na temat zadania*, bez konieczności podania przekonującego uzasadnienia teoretycznego. Można na przykład obserwować konsekwencje przyjęcia określonych wartości tego parametru dla określonych zadań. Można próbować ustalić, w jaki sposób wartość tego parametru powinna się zmieniać dla danego zadania, żeby optymalne rozwiązanie było wybierane możliwie najszybciej i najczęściej. Może nawet udałoby się odkryć jakąś funkcję, pozwalającą wyznaczyć względnie efektywny sposób manipulowania parametrem ϵ dla całej wybranej klasy zadań. W ten sposób można się zbliżyć do zadowalającego rozwiązania niektórych wersji problemu w praktyce, albo do dopasowania modelu do danych, ale nie do zrozumienia, o co w nim chodzi.

Omówione wyżej metody rozwiązania zadania n-rękiego bandyty opierają się na oszacowaniach wartości poszczególnych działań. Ponieważ zadanie jest nieasocjacyjne (istnieje tylko jeden stan), wartości samych działań są wszystkim, czego agent potrzebuje, aby wybrać działanie optymalne. Gdyby zadanie zawierało więcej niż jeden stan, w ogólnym przypadku wartości działań byłyby zależne od tego, w jakim stanie te działania są wybierane. Nieasocjacyjność pozwala więc uprościć rozumowanie, niemniej zakładany model środowiska nie będzie adekwatny dla przeważającej większości realnych sytuacji sekwencyjnego podejmowania decyzji. Nie należy się tym przejmować, nie szukamy teraz bowiem konkretnego modelu rzeczywistego środowiska, tylko próbujemy uchwycić

sens funkcji.

Zakładając, że zadanie jest stacjonarne, a więc jego dynamika nie ulega zmianie w trakcie interakcji, każde dostępne dla agenta działanie będzie miało pewną prawdziwą, choć nieznaną agentowi wartość. W przebiegu interakcji agent może szacować wartość działań, na przykład obliczając średnią ze wszystkich dotychczasowych nagród, otrzymanych po wybraniu tego działania. Jeżeli prawdziwą wartość działania a oznaczymy przez $Q(a)$, a oszacowanie w momencie t, obliczone jako średnia z k_a nagród, uzyskanych po wybraniu działania a do momentu t, przez $\hat{Q}_t(a)$, mamy następującą gwarancję asymptotyczną:

$$\lim_{k_a \to \infty} \hat{Q}_t(a) = Q(a) \tag{8.1}$$

Aby spełnić ten warunek dla danego działania, prawdopodobieństwo wyboru tego działania w trakcie interakcji musi pozostawać niezerowe. Odpowiednio, aby uzyskać gwarancję, że oszacowania wartości wszystkich działań będą w granicy dążyły do wartości prawdziwych, warunek ten musi być spełniony dla wszystkich działań. Rozwiązanie ϵ-zachłanne, gdzie $\epsilon > 0$, jest jednym ze sposób spełnienia tego warunku, ponieważ zapewnia stałą eksplorację dla wszystkich działań.

Gwarancje asymptotyczne są jednak tylko gwarancjami asymptotycznymi. W szczególności, (8.1) nie mówi nic na temat tego, w jakim tempie oszacowania wartości działań będą się zbliżać do wartości prawdziwych. Warunek (8.1) można spełnić na nieskończenie wiele sposobów, nawet jeżeli ograniczymy się tylko do rozwiązań ϵ-zachłannych, przyjmując dowolną wartość ϵ taką, że $0 < \epsilon <= 1$. Zależnie od tego prawdopodobieństwa i środowiska rozwiązanie może być dobre, albo złe.

Ewidentną wadą rozwiązań ϵ-zachłannych jest traktowanie wszystkich działań niezachłannych jednakowo. Rozwiązanie takie będzie bardzo nierozsądne na przykład wtedy, gdy agent po dłuższym okresie interakcji ze środowiskiem uzyska dosyć dobre oszacowania wartości. Agent będzie wtedy dysponował ważną informacją nie tylko na temat tego, które działanie jest prawdopodobnie optymalne, ale również tego, które spośród pozostałych działań są prawdopodobnie lepsze a które (być może znacznie) gorsze. Jednym ze standardowych sposobów wykorzystania informacji uzyskanej w trakcie interakcji do poprawy mechanizmu eksploracji jest metoda selekcji działań typu softmax[3]:

$$P_t(a) = \frac{e^{Q_t(a)/\tau}}{\sum_{b=1}^{n} e^{Q_t(b)/\tau}},\ 0 < \tau$$

gdzie n to liczba dostępnych działań, a τ to parametr temperatury. Wysokie wartości parametru τ sprawiają, że działania stają się niemal jednakowo prawdopodobne, a im wartość tego parametru jest niższa, tym bardziej prawdopodobieństwo wyboru każdego

[3]Nazywana też często metodą selekcji opartą na rozkładzie Boltzmanna.

działania będzie zależne od różnicy między szacowaną wartością tego działania, a wartościami pozostałych działań. Im bardziej temperatura zbliża się do zera, tym bardziej reguła softmax przypomina regułę zachłanną. W ten sposób uzyskujemy kontinuum rozwiązań - w miarę jak rośnie wartość τ, reguła zachowuje się coraz mniej zachłannie, wybierając działania z prawdopodobieństwami proporcjonalnymi do ich szacowanej wartości. Dalszy wzrost temperatury powoduje w końcu coraz silniejsze wyrównanie wszystkich prawdopodobieństw. Podobny efekt można uzyskać na wiele sposobów. Można dodać wartość zmiennej losowej do oszacowanych wartości każdego działania i wybierać działanie, dla którego obliczona w ten sposób suma jest największa. Manipulując wariancją takiej zmiennej losowej możemy uzyskać zbliżone spektrum rozwiązań, od zachłannego do coraz bardziej eksploracyjnych.

Zarówno w przypadku reguły ϵ-zachłannej, jak i w przypadku reguły softmax musimy poradzić sobie jakoś z jedynym wolnym parametrem. Trudno oprzeć się wrażeniu, że przynajmniej dla tak prostych zadań jak n-ręki bandyta istnieją jakieś zasady określające, jaka jest optymalna wartość tego wolnego parametru i w jaki sposób parametr powinien się optymalnie zmieniać w trakcie interakcji. Mimo, że reguła softmax wydaje się lepsza niż ϵ-zachłanna, nadal nie wiemy jak bardzo, ani czy nie istnieją rozwiązania znacznie od niej lepsze.

Tak naprawdę cały czas zmierzaliśmy w niewłaściwym kierunku. Udoskonalanie rozwiązań zaledwie obiecujących przez dodawanie mechanizmów korygujących elastyczne elementy jest rozkładaniem na części niewłaściwego homunculusa. Żadne z opisanych rozwiązań nie uwzględnia *niepewności* związanej z oszacowaniami, zapominamy też o horyzoncie czasowym interakcji. Uwzględnienie tych dwóch elementów wymaga, aby agent korzystał wprost lub nie wprost z modelu środowiska.

Jeżeli natrafimy na agenta działającego w przybliżeniu optymalnie w nietrywialnym środowisku niedeterministycznym i uprawnione będzie założenie, że tego środowiska na początku dokładnie nie znał, najlepszym wyjaśnieniem tego faktu będzie stwierdzenie, że agent dysponował odpowiednim modelem. To wyjaśnienie nie przestanie być dobre, gdy okaże się, że trudno w pudełku A znaleźć cokolwiek, co wyglądałoby na „strukturę reprezentującą środowisko", albo mechanizm szacowania wartości.

Optymalne, ale nazbyt kosztowne obliczeniowo rozwiązanie problemu n-rękiego bandyty z perspektywy bayesowskiej podał Bellman (1957). W wydanej niedawno, niezwykłej pracy teoretycznej Hutter (2004) przedstawił optymalne, ale nieobliczalne rozwiązanie problemu uczenia się ze wzmocnieniem dla dowolnego środowiska o dowolnym nieznanym, efektywnie obliczalnym rozkładzie prawdopodobieństwa, na podstawie znajomości rozwiązania optymalnego stworzył też szereg efektywnie obliczalnych rozwiązań w przybliżeniu optymalnych. Psychologiczny sens tych rezultatów nie jest jednak dla mnie jasny, jako że opierają się na zastosowaniu aparatu pojęciowego teorii złożoności algorytmicznej.

Dearden, Friedman i Russell (1998) uzyskali przybliżone rozwiązanie bayesowskie ogólnego problemu uczenia się ze wzmocnieniem posługując się pojęciem wartości in-

formacji, czyli oczekiwanej poprawy przyszłych decyzji, która mogłaby wynikać z informacji uzyskanej dzięki eksploracji. Za pomocą zbliżonego pojęcia oczekiwanego zysku informacyjnego, Oaksford i Chater (1994) stworzyli interesujący model racjonalny, zawierający trzy wolne parametry i wyjaśniający wiele, czasami zaskakujących regularności obserwowanych w sytuacji wykonywania rozmaitych wersji nieśmiertelnego zadania Wasona (Wason, 1968; Johnson-Laird i Wason, 1970).

Żeby nie odrywać się zanadto od głównego wątku, pominę tutaj techniczne szczegóły rozwiązania optymalnego. Sens tego rozwiązania dla najprostszego przypadku zadania dwurękiego bandyty zwięźle opisał Dimitrakakis (2006) - gdy dane są dwa możliwe działania takie, że jedno z nich ma niższą oczekiwaną wartość niż drugie, stosowanie przypuszczalnie gorszego działania może być uzasadnione, jeżeli: istnieje niepewność co do tego, czy gorsze działanie nie jest faktycznie lepsze, chcemy maksymalizować coś więcej niż tylko nagrodę w następnym kroku (oczekiwaną sumę nagród) i będzie można dokładnie ustalić, które działanie jest lepsze na tyle szybko, że straty wynikające z eksploracji nie będą zbyt duże.

Jak zauważył Dimitrakakis, podjęty w latach 40-tych przez Walda (1947) problem wnioskowania statystycznego na podstawie wyników warunkowo przerwanego eksperymentu jest szczególnym przypadkiem ogólnego problemu przetargu między eksploracją i eksploatacją. Gromadząc kolejne obserwacje w trakcie przeprowadzania badania uzyskujemy coraz silniejsze wsparcie dla jednej z hipotez, na przykład zerowej lub alternatywnej. Działaniem zachłannym będzie wtedy podjęcie decyzji o odrzuceniu jednej z hipotez, a działaniem eksploracyjnym kontynuowanie badania. Szczególnym przypadkiem problemu warunkowego przerwania eksperymentu jest z kolei proces dyfuzyjny, który nie jest niczym innym, jak optymalnym rozwiązaniem problemu wyboru z dwóch alternatyw dla określonego środowiska wewnętrznego. Środowiskiem w modelu dyfuzyjnym jest źródło świadectw (abstrakcyjnie ujęty element fizjologii mózgu), a świadectwa są wewnętrznymi sygnałami stanu środowiska. Model dyfuzyjny jest optymalny w tym znaczeniu, że minimalizuje ilość czasu do podjęcia decyzji przy założonym prawdopodobieństwie błędu i minimalizuje prawdopodobieństwo błędu przy założonej ilości czasu (Bogacz i in., 2006). Między innymi to miałem na myśli pisząc o cechach szczególnych tego modelu w podsumowaniu rozdziału drugiego. Bogacz (2006) zaproponował kilka mocniejszych hipotez dotyczących optymalności tego procesu. Gdyby wynikające z nich predykcje miały się okazać w przybliżeniu trafne, uzasadniona byłaby odpowiednio mocniejsza interpretacja parametrów, na przykład, gdyby początkowa wartość świadectwa zmieniała się tak, jak powinna się zmieniać subiektywna siła przekonania co do prawdopodobieństwa wystąpienia jednej z alternatyw, można by twierdzić, że interpretacja wartości tego parametru w kategoriach subiektywnego prawdopodobieństwa jest uzasadniona.

W ten sposób racjonalna teoria dostarcza interpretacji komputacyjnej dla stosunkowo elementarnego procesu, pozwala modelować rozkłady czasów reakcji i poprawności w wielu prostych zadaniach i częściowo wyjaśnia obserwowane w tych zadaniach efek-

ty zarówno w kategoriach procesów poznawczych jako procesów poznawczych jak i w kategoriach mechanizmu obliczeniowego. Nawet najprostsze zadania laboratoryjne wymagają wyrafinowanego wnioskowania i podejmowania decyzji w warunkach niepewności. Na komputacyjnym poziomie opisu, na którym mechanizmy obliczeniowe stają się rozwiązaniami tego rodzaju problemów, współczesne modele zintegrowane i przypominające modele zintegrowane współczesne modele pamięci roboczej są bardzo słabo określone. Ważnym źródłem tego niedookreślenia jest przypuszczalnie pragnienie poznania mechanizmu, zanim dobrze zrozumie się funkcję.

Mniej więcej tak, jak to ujął Dimitrakakis, mogłaby moim zdaniem brzmieć rozsądna *psychologiczna* robocza definicja eksploracji. Stosując taką definicję można próbować stwierdzić, kiedy i w jakim stopniu rzeczywiste działające celowo systemy dokonują eksploracji, kiedy eksploracja jest, a kiedy nie jest konieczna, od czego może zależeć jej skuteczność i jakie są granice tej skuteczności. Można też zaproponować ewentualną typologię zachowań eksploracyjnych. Przejdę teraz do omówienia kilku zagadnień związanych z niektórymi, podstawowymi ze względu na komputacyjne wymagania skutecznej celowej interakcji, własnościami środowisk.

8.4 Własność Markowa i jej psychologiczny sens

Nieformalnie, własność Markowa to inaczej niezależność od ścieżki. Własność ta występuje wtedy, gdy jakiś stan rzeczy zawiera pełną informację na temat swoich następstw. Na przykład, znając położenie i wektor ruchu jakiegoś obiektu w przestrzeni wiemy mniej więcej wszystko, co trzeba wiedzieć, aby przewidzieć położenie obiektu w chwili następnej. Podobnie, znowu upraszczając, jeżeli znamy układ figur na szachownicy, wiemy w przybliżeniu wszystko co potrzebujemy wiedzieć, aby podjąć decyzję o następnym posunięciu. Takie procesy mają własność Markowa.

Sygnał stanu ma własność Markowa, a całe zadanie jest szczególnym przypadkiem procesu decyzyjnego Markowa (MDP), jeżeli dynamika środowiska daje się całkowicie opisać przez funkcje:

$$P^a_{s,s'} = Pr(s_{t+1} = s' | s_t = s, a_t = a)$$
$$R^a_{s,s'} = E_\pi(r_{t+1} | s_t = s, s_{t+1} = s', a_t = a)$$

gdzie $P^a_{s,s'}$ to prawdopodobieństwo przejścia ze stanu s do stanu s', po wykonaniu w stanie s działania a, a $R^a_{s,s'}$ to nagroda za przejście ze stanu s do stanu s', po wykonaniu w stanie s działania a. Prawdopodobieństwo każdego możliwego przejścia i oczekiwana nagroda za przejście zależą jedynie od stanu wyjściowego i działania wykonanego w tym stanie. Nie zależą w ogóle od tego, którędy agent tam dotarł, albo jakie wybierał wcześniej działania. Jeżeli sygnał stanu ma własność Markowa i zbiory stanów i działań są skończone, to zadanie jest skończonym procesem decyzyjnym Markowa (fMDP).

Ponieważ liczba stanów i działań jest skończona, zadanie takie można zwizualizować za pomocą grafu albo tabelki.

Jak już wspomniałem wcześniej, własność Markowa znacznie upraszcza wiele obliczeń, szczególnie gdy liczba stanów i działań jest skończona. Okazuje się też, że rozwiązania stworzone przy założeniu tej własności dają się często z niezłym skutkiem stosować do rozwiązywania zadań, które tak naprawdę są procesami Markowa tylko w przybliżeniu (Sutton i Barto, 1998). Na podstawie analizy teoretycznej komputacyjnych wymagań skończonych procesów decyzyjnych Markowa Suttonowi i Barto udało się odkryć pewien algorytm, który jak dotąd wydaje się być najlepszym (choć nadal dalekim od doskonałości) modelem procesu warunkowania klasycznego u ludzi i zwierząt (Dayan i Niv, 2008; Montague, Dayan i Sejnowski, 1996; O'Reilly, Frank, Hazy i Watz, 2007; Sutton i Barto, 1990). Ten tak zwany algorytm różnic czasowych (w skrócie TD, ang. *Temporal Difference learning*), jest w przybliżeniu optymalnym rozwiązaniem[4] pewnej ogólnej klasy zadań uczenia się ze wzmocnieniem.

8.4.1 Optymalne i w przybliżeniu optymalne rozwiązania zadań typu fMDP

Przebieg interakcji w nieepizodycznym zadaniu typu fMDP można przedstawić za pomocą rozrastającego się w nieskończoność drzewa. Na przykład, środowisko może się składać z dwóch stanów s_1 i s_2 takich, że w każdym z tych stanów dostępne są dwa działania, które z niezerowym prawdopodobieństwem przeprowadzają agenta albo do drugiego stanu, albo do stanu wyjściowego. Zakładając, że jeden ze stanów jest zawsze stanem początkowym, przestrzeń możliwych interakcji można sobie wyobrazić jako drzewo rozrastające się w nieskończoność od tego stanu początkowego. Konsekwencje każdego działania w każdym ze stanów też będą takim rozgałęziającym się w nieskończoność drzewem, a to, która konkretnie trajektoria zostanie wybrana zależy zarówno od polisy jak i od środowiska.

Załóżmy, że agent wędrował przez pewien czas po tym środowisku i właśnie za pomocą działania a_{12} przeszedł ze stanu s_1 do stanu s_2. Uzyskał w ten sposób pewną nagrodę i powinien wykorzystać tą informację do ewentualnej korekcji polisy. To, czym agent jest przede wszystkim zainteresowany, to poszukiwanie polisy maksymalizującej wartość podejmowanych działań. Jakiekolwiek by nie były wcześniejsze oszacowania wartości działań, oszacowania te powinny być teraz uaktualnione w oparciu o uzyskaną nagrodę. Nie można tego zrobić po prostu uaktualniając oszacowanie średniej nagrody za przejście, ponieważ wartość działań zależy nie tylko od nagrody uzyskanej za odpowiednie przejście, ale także od tego, jakie nagrody będą uzyskane w dalszym przebiegu interakcji.

[4]Zawiera jednak kilka elementów wprowadzonych ad hoc i na pewno nie jest rozwiązaniem optymalnym, tylko właśnie w przybliżeniu optymalnym. Nie wiadomo, jak dobre jest to przybliżenie w ogólnym przypadku, wiadomo natomiast, że czasami jest bardzo złe nawet dla zadań typu fMDP (np. Bertsekas, 1995)

Dla zadań nieepizodycznych sumę nagród definiuje się zwykle jako sumę dyskontowaną, co gwarantuje, że suma ta będzie skończona, o ile ciąg r_k jest ograniczony:

$$R = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \cdots = \sum_{k=0}^{\infty} (\gamma^k r_{t+k+1}) \tag{8.2}$$

gdzie $0 \leqslant \gamma \leqslant 1$ to parametr dyskontowania. Im większa wartość γ, tym bardziej (obiektywna) ocena wartości działań jest dalekowzroczna. Zależności między wartościami działań w zadaniu typu fMDP wyraża tak zwane równanie Bellmana:

$$\begin{aligned} V^{\pi}(s) =& E_{\pi}(R|s_t = s) \\ =& E_{\pi}\left[\sum_{k=0}^{\infty} (\gamma^k r_{t+k+1})|s_t = s\right] \\ =& E_{\pi}(r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \cdots |s_t = s) \\ =& E_{\pi}\left[r_{t+1} + \gamma \sum_{k=0}^{\infty} (\gamma^k r_{t+k+2})|s_t = s\right] \\ =& \sum_{a \in A(s)} \pi(s,a) \sum_{s' \in S} P^a_{s,s'}(R^a_{s,s'} + \gamma V^{\pi}(s')) \end{aligned} \tag{8.3}$$

Na elementach zbioru polis można zdefiniować relację częściowego porządku:

$$\pi' \geqslant \pi \leftrightarrow \forall_{s \in S} V^{\pi'}(s) \geqslant V^{\pi}(s)$$

Można wykazać, że dla fMDP zawsze będzie istniała co najmniej jedna optymalna polisa π, to jest taka, że $\forall_{\pi'} \pi \geqslant \pi'$. Z definicji tej polisy wynika natychmiast, że wszystkie polisy optymalne mają tą samą (optymalną) funkcję wartości $V^*(s) = \max_{\pi} V^{\pi}(s)$. Gdy optymalna funkcja wartości jest znana, wyznaczenie optymalnej polisy staje się trywialne - będzie to dowolna polisa maksymalizująca wartość tej funkcji. Wynika stąd między innymi, co być może zaskakujące, że optymalne rozwiązanie zadania typu fMDP jest zachłanne, odroczone konsekwencje działań są bowiem uwzględniane przez polisę optymalną poprzez lokalne oszacowania wartości. Dla fMDP ta optymalna funkcja wartości daje się obliczyć analitycznie, wymaga to jednak rozwiązania układu tylu równań, ile jest stanów, co w przypadku nietrywialnego zadania jest nazbyt kosztowne. Dodatkowo, żeby agent mógł w ten sposób (analitycznie) rozwiązywać problem uczenia się ze wzmocnieniem, musiałby z góry znać dynamikę środowiska, a w psychologii prawie nigdy nie można tego zakładać.

Wiele klasycznych rozwiązań problemu uczenia się ze wzmocnieniem dla zadań typu fMDP opiera się na wykorzystaniu rekurencyjnej natury równania Bellmana. Znajomość rozwiązania optymalnego przydaje się do tworzenia rozwiązań przybliżonych, podobnie jak teoria racjonalna przydaje się do formułowania uzasadnionych hipotez na temat mechanizmu obliczeniowego. Zgodnie z równaniem (8.3), wartość danego stanu

(lub równoważnie, działania w danym stanie) jest sumą nagród za dostępne w tym stanie przejścia i wartości stanów osiągalnych z tego stanu, ważoną przez prawdopodobieństwa odpowiednich przejść. Dla uproszczenia będę odtąd zakładał, że polisa jest deterministyczna.

Z równania (8.3) wynika coś, co można było już dostrzec rozważając przestrzeń trajektorii interakcji jako rozrastające się w nieskończoność drzewo. Oszacowania wartości stanów (lub działań) mogą być korygowane w oparciu o oszacowania wartości innych stanów i uzyskane nagrody. Niezależnie od tego, jakimi oszacowaniami agent akurat dysponuje, po każdym przejściu można je poprawić propagując wstecz informację o uzyskanej nagrodzie. W ten sposób agent może skorygować oszacowanie wartości stanu, z którego właśnie wyszedł, stanów, które do niego prowadzą i tak dalej, z dowolną wybraną głębokością.

Na początek załóżmy, że agent zna dynamikę środowiska, ale nie zna funkcji wartości. O ile $\gamma < 1$, lub dla każdego stanu polisa prowadzi do stanu terminalnego, dla dowolnej polisy będzie istniała unikalna funkcja wartości. Aby ją obliczyć, można rozpocząć od dowolnego oszacowania wartości V_0 i korygować je iteracyjnie zgodnie z równaniem (8.3). Powstaje w ten sposób ciąg kolejnych przybliżeń:

$$\begin{aligned} V_{k+1}(s) =& E_\pi \left[r_{t+1} + \gamma \sum_{k=0}^{\infty} (\gamma^k r_{t+k+2}) | s_t = s \right] \\ =& \sum_{a \in A(s)} \pi(s,a) \sum_{s' \in S} P^a_{s,s'} (R^a_{s,s'} + \gamma V^\pi_k(s')) \end{aligned}$$

Zależnie od tego, czy w każdym kroku tej procedury uwzględniane są wszystkie, czy tylko część stanów i czy uwzględniane są wszystkie, czy tylko część stanów osiągalnych ze stanów, dla których oszacowanie jest korygowane, powstaną różne wersje w istocie tego samego, ogólnego rozwiązania. Można wykazać, że taki ciąg przybliżeń będzie zmierzał do wartości prawdziwej.

Skoro dokonaliśmy już *ewaluacji* polisy, przyszedł czas aby ją *poprawić*. Z twierdzenia o poprawie polisy Bellmana wynika, że można to zrobić tworząc polisę π' zachłanną ze względu na funkcję wartości dla polisy wyjściowej:

$$\pi'(s) = \max_a Q^\pi(s,a)$$

Z definicji polisy optymalnej wynika natomiast, że jeżeli uzyskana w ten sposób polisa π' będzie taka, że $V^\pi = V^{\pi'}$, czyli nie nastąpi żadna dalsza poprawa, polisa π' będzie optymalna. Proces poszukiwania polisy optymalnej można więc przedstawić jako dokonywanie na przemian ewaluacji i poprawy polisy do momentu, aż dalsza poprawa nie będzie możliwa. Dokonując za każdym razem ewaluacji w oparciu o funkcję wartości dla poprzedniej polisy, korzystamy z rozsądnych początkowych oszacowań, zmniejszając tym samym koszty ewaluacji. Dla zadania typu fMDP zbiór polis deterministycznych jest

skończony, a więc wiemy, że polisa optymalna zostanie odnaleziona w skończonej liczbie kroków. Tego rodzaju rozwiązania często zbiegają się w zaskakująco krótkim czasie (Sutton i Barto, 1998).

Opierając się na tej samej zasadzie można stworzyć wiele rozwiązań praktycznych. Gdy zbiór stanów jest duży, redukcję kosztów ewaluacji można uzyskać zmniejszając liczbę przejść przez zbiór stanów, dokonując ewaluacji tylko dla stanów wybranych, lub uwzględniając tylko część stanów następnych. Na przykład, stany dla których dokonywana jest korekcja oszacowań można wybierać ze względu na to, jak bardzo są obiecujące. Skuteczną redukcję kosztów obliczeniowych uzyskuje się dzięki temu, że rozwiązanie jest ukierunkowane.

Zdaniem Suttona i Barto, przeważającą większość rozwiązań problemu uczenia się ze wzmocnieniem można z pożytkiem przedstawić jako szczególne przypadki iteracyjnej ewaluacji i poprawy polisy (nazywanej przez nich uogólnioną iteracją polisy). Kroki ewaluacji i poprawy mogą być dowolnie przekładane i mniej lub bardziej wyczerpująco realizowane. Jeżeli zagwarantujemy, że prawdopodobieństwo ewaluacji każdego stanu pozostanie niezerowe, metoda zbiegnie się ostatecznie do polisy optymalnej. Na tej samej ogólnej zasadzie opiera się jedno z klasycznych rozwiązań problemu uczenia się ze wzmocnieniem w sytuacji, gdy dynamika środowiska nie jest znana, to jest metoda różnic czasowych.

8.4.2 Metoda różnic czasowych

Najważniejsza różnica między opisanymi wcześniej metodami opartymi na założeniu znajomości dynamiki środowiska a metodą TD dotyczy kroku ewaluacji polisy. Załóżmy, że w momencie t agent wykonał pewne działanie, na skutek czego uzyskał nagrodę r_{t+1}. Prawdziwa wartość stanu s w momencie t wynosi $V^{\pi}(s_t) = r_{t+1} + \gamma V^{\pi}(s_{t+1})$, nie zakładamy jednak, że agent zna wartość stanów, tylko że dysponuje niedoskonałymi oszacowaniami. Przyjmijmy, że $V(s_t)$ jest oszacowaniem, a nie wartością prawdziwą i dla uproszczenia pomińmy na razie oznaczenie zależności wartości stanów od polisy.

Wielkość $r_{t+1} + \gamma V(s_{t+1})$, którą będę odtąd nazywał próbką wartości stanu, dostarcza tylko częściowej, niedoskonałej informacji na temat prawdziwej wartości stanu s_t. Agent nie wie, czy podejmując to samo działanie w tym samym stanie uzyska tą samą nagrodę. Przejścia między stanami nie mają charakteru deterministycznego, polisa nie musi być deterministyczna, agent nie może również zakładać, że $V(s_{t+1})$ jest prawdziwą wartością stanu s_{t+1}. Próbka wartości stanu jest oszacowaniem wartości stanu s_t, opartym na pojedynczej próbie z rozkładu nagród, zależnego zarówno od polisy jak i od dynamiki środowiska, i na niedoskonałym oszacowaniu wartości tylko jednego ze stanów osiągalnych z s_t[5]. Oszacowanie $V(s_t)$ trzeba do tej wartości zbliżyć, ale nie wiadomo jak bardzo.

[5]Korygowanie oszacowań na podstawie innych oszacowań określa się zwykle jako „bootstrapowanie" (Efron i Gong, 1983).

W najprostszej wersji metoda TD polega na dodaniu różnicy między aktualnym oszacowaniem a uzyskaną próbką wartości stanu, ważonej przez pewien parametr, pozwalający właśnie na taką ostrożną korekcję oszacowania. Konkretnie, w każdym kroku wartość opuszczanych stanów jest uaktualniana zgodnie z następującą regułą:

$$V(s_t) = V(s_t) + \alpha[(r_{t+1} + \gamma V^{\pi}(s_{t+1})) - V(s_t)]$$

Dla dowolnej ustalonej polisy π, metoda zbiega się do wartości prawdziwej, o ile tylko wartość α jest odpowiednio mała, lub maleje zgodnie z warunkami aproksymacji stochastycznej (Robbins i Monro, 1951). Ponieważ uaktualnianie dotyczy tylko odwiedzanych stanów, zbieżność do polisy optymalnej wymaga stałej eksploracji. Stosując niezupełnie eksploracyjną metodę selekcji działań uzyskujemy rozwiązanie polegające na mniej lub bardziej ukierunkowanym uczeniu się. Odwiedzane będą wtedy przede wszystkim stany obiecujące i uaktualnianie będzie dotyczyło przede wszystkim takich właśnie stanów.Dayan i Sejnowski (1994) wykazali, że w wielu wypadkach algorytm TD zbiegnie się do polisy optymalnej, o ile oczywiście prawdopodobieństwo wszystkich działań dostępnych w każdym ze stanów pozostaje niezerowe i polisa zbiega się w granicy do zachłannej.

8.4.3 Algorytm różnic czasowych jako mechanizm warunkowania klasycznego

W typowym eksperymencie dotyczącym warunkowania klasycznego prezentowane są jeden po drugim dwa bodźce, to jest bodziec warunkowy (BW) i bodziec bezwarunkowy (BB). Systematyczna prezentacja tych bodźców powoduje, że reakcja (RW) wywoływana początkowo tylko przez bodziec bezwarunkowy, zaczyna się pojawiać również w odpowiedzi na bodziec warunkowy. Jeżeli BB będzie podmuchem powietrza skierowanym na gałkę oczną, a BW dźwiękiem dzwonka, z czasem dźwięk dzwonka zacznie wywoływać mrużenie (RW). Zgodnie z dwoma klasycznymi interpretacjami, proces warunkowania klasycznego może polegać na uczeniu się zależności przyczynowo-skutkowych (Dickinson, 1980), albo na przewidywaniu wystąpienia BB na podstawie wystąpienia BW (Rescorla i Wagner, 1972).

Teorie warunkowania klasycznego można podzielić ze względu na to, jak traktują dynamikę wydarzeń wewnątrz próby i ze względu na to, czy wyjaśniają rozmaite obserwowane efekty w kategoriach przetwarzania BB, BW, czy przetwarzania obu tych rodzajów bodźców. Teorie czasu rzeczywistego opisują zmiany w asocjacjach jako proces rozgrywający się w trakcie próby, zależny od tego, w którym momencie bodźce zaczynają i przestają oddziaływać. Teorie poziomu próby ignorują całkowicie szczegóły dynamiki zdarzeń rozgrywających się w trakcie prób, za ich pomocą można więc ewentualnie wyjaśnić wyniki tylko takich eksperymentów, w których dynamika zdarzeń wewnątrz prób pozostaje stała.

Jak zauważyli Sutton i Barto, bodaj wszystkie teorie warunkowania traktują zmiany w sile asocjacji jako multiplikatywny efekt „przetwarzania" BB i BW. Określając za

autorami poziom przetwarzania BB jako „wzmocnienie", a poziom przetwarzania BW jako „relewancję" (ang. *eligibility*)[6], teorie te można sprowadzić do schematu:

$$\Delta V = \text{Wzmocnienie} \times \text{Relewancja} \tag{8.4}$$

Postulowany wpływ czynników związanych z BB i BW jest multiplikatywny, dlatego wiele znanych efektów można zwykle wyjaśnić odwołując się albo do zmian wzmocnienia, albo relewancji. Zastosowanie terminu „relewancja" wydaje mi się częściowo uzasadnione. Obserwowane zmiany w sile asocjacji zależne od właściwości BW związane są między innymi z jego wyrazistością, intensywnością i dystansem czasowym, jaki dzieli BW i BB, a efekty te można wyjaśnić jako sensowne konsekwencje spostrzeganej siły związku między BW i BB, czyli konsekwencje tego, na ile BW wydaje się odpowiedni jako kandydat na predyktora BB. Przy tej interpretacji większa intensywność BW oznaczałaby, że to raczej BW a nie coś innego pozwala przewidzieć BB. Ta interpretacja jest jednak daleka od doskonałości, skoro przynajmniej teoretycznie możliwe jest, żeby warunkowanie klasyczne było złożoną konsekwencją oceny występowania, siły i charakteru związku jednocześnie, tak jak to ma miejsce w modelu oceny kauzalnej Griffithsa i Tenenbauma.

Do niedawna bodaj najbardziej udaną, w znaczeniu prostoty i zakresu wyjaśnianych zjawisk, teorią warunkowania klasycznego był model Rescorli-Wagnera (1972), dalej w skrócie R-W. Można go przedstawić jako szczególny przypadek równania (8.4), jeżeli tylko uzupełnimy równanie modelu R-W o zmienną X_i, przyjmującą wartość 1 gdy BW_i wystąpił w danej próbie i 0, gdy nie wystąpił:

$$\Delta V_i = \beta(\lambda - \bar{V})\alpha_i X_i \tag{8.5}$$

Bez szkody dla dalszych wniosków, dla uproszczenia założyłem, że bodziec bezwarunkowy może być tylko jeden. Wartości parametrów β i α zależą od własności BB (β) i BW (α), przy czym α odzwierciedla między innymi wyrazistość BW. Zwykle zakłada się, że α i β nie mogą być mniejsze niż 0 ani większe niż 1. Według autorów tego modelu warunkowanie polega na przewidywaniu pojawienia się i intensywności BB na podstawie spostrzeganego oddziaływania BW.

Zmiana asocjacji miałaby zachodzić wtedy, gdy zdarza się coś nieoczekiwanego, to znaczy niezgodnego z przewidywaniami. Przewidywany poziom BB reprezentowany jest w modelu przez $\bar{V} = \sum_i V_i X_i$, czyli sumaryczną „siłę oczekiwania" BB, zależną od

[6]W literaturze polskiej to, co nazywam w tej pracy „działaniem", „metodą selekcji", „polisą", „siłą obecności" i „relewancją" określa się odpowiednio jako „akcję", „strategię" (zarówno polisa jak i metoda selekcji), „współczynnik świeżości" i „aktywność". W teorii decyzji zamiast terminu „polisa" używa się często terminu „reguła", albo „reguła działania". Przyjąłem własną terminologię, żeby uzgodnić język aparatów pojęciowych stosowanych w tej pracy i ułatwić Czytelnikowi uchwycenie psychologicznego sensu tych pojęć.

obecności bodźców warunkowych i sił asocjacji między tymi bodźcami a bodźcem bezwarunkowym. Parametr λ reprezentuje z kolei „siłę obecności", albo „stopień przetwarzania" BB w danej próbie. Jeżeli BB jest nieobecny, $\lambda = 0$, jeżeli jest obecny, λ przyjmuje wartość dodatnią, zależną w bliżej nieokreślony sposób od wyrazistości i intensywności BB i długości interwału (ISI) między pojawieniem się BW i BB. Gdy te czynniki są stałe i BB nie ulega zmianie między próbami, często zakłada się, że $\lambda = 1$. Jeżeli systematycznie po tym samym BW będzie następował BB, siła asocjacji będzie się coraz bardziej zbliżała do λ, czyli obserwowanego „poziomu" BB.

Model R-W z $\lambda = 1$ wydaje się wymagać wszystkich tych wolnych parametrów. Jako że są one związane z fizycznymi charakterystykami bodźców i ISI, nie może istnieć teoria racjonalna, pozwalająca na całkowite usunięcie tej elastyczności. Akurat co do tego rodzaju zależności Cummins miał rację. Można sobie jednak wyobrazić rozwiniętą wersję teorii, określającą jak *zmiany* w wartościach tych parametrów powinny zależeć od zmian fizycznych własności bodźców i ISI, co pozwoliłoby na częściowe zredukowanie statystycznej elastyczności w sytuacji, gdy wymienione czynniki są kontrolowane (nie trzeba by było dopasowywać tych parametrów osobno dla każdego warunku). Wersja modelu, w której dopuszcza się, aby parametry λ i β oba zależały od własności BB (w tym ISI) jest już bardziej problematyczna, dlatego że oba te parametry mają teoretycznie reprezentować wpływ tych samych czynników. Dla ustalonych ISI i własności fizycznych BB wydaje się, że taki model jest zbyt elastyczny, stąd też konwencja, aby w takim przypadku stosować $\lambda = 1$ sprawia wrażenie uzasadnionej.

Z modelu R-W wynika, że bodźce warunkowe konkurują ze sobą o maksymalną możliwą sumę asocjacji λ, przy czym uczenie się dotyczy tylko tych bodźców warunkowych, które wystąpiły w danej próbie. Model pozwala wyjaśnić wiele regularności warunkowania klasycznego, a historycznie ważnym źródłem wsparcia empirycznego dla tej teorii było między innymi wynikające z niego, stosunkowo proste i intuicyjne wyjaśnienie blokowania. Efekt blokowania polega na tym, że uwarunkowanie jednego bodźca blokuje proces późniejszego warunkowania bodźca prezentowanego razem z bodźcem wcześniej uwarunkowanym. W teorii R-W blokowanie jest oczywistą konsekwencją konkurowania bodźców o sumaryczną siłę asocjacji, na poziomie komputacyjnym efekt wyjaśniany jest więc przez stwierdzenie, że drugi BW nie przyczynia się do dalszej redukcji niepewności co do wystąpienia BB. Istnieje jednak wiele znanych regularności, do których model nie pasuje, a teoria nie wyjaśnia.

Model R-W nie radzi sobie między innymi z poprawnym przewidywaniem efektów warunkowania drugiego rzędu. Jeżeli po pewnym BW (A) systematycznie następuje pewien BB, a w trakcie kolejnej serii prób po innym BW (B) również systematycznie następuje A, ale nie BB, obserwuje się wyraźny efekt wyuczonej asocjacji między B i A (Rescorla, 1980). Oba te bodźce są warunkowe, a więc zgodnie z modelem $\lambda = 0$ i żadne warunkowanie nie powinno nastąpić. Chociaż warunkowanie drugiego rzędu nie jest przewidywane przez model, można ten efekt wyjaśnić za pomocą nieznacznie rozwiniętej wersji teorii, zachowując jej komputacyjny sens. Wystarczy zauważyć, że uwa-

runkowany BW dostarcza informację na temat wystąpienia BB nawet wtedy, gdy BB faktycznie po nim nie następuje. W formalnym modelu uwarunkowany BW powinien generować własną wartość λ, jednak teoria nie określa jaką.

Bardziej podstawowa wada modelu R-W polega na nieuwzględnieniu dynamiki zdarzeń rozgrywających się w trakcie próby. Segundo, Galeano, Sommer-Smith i Roig (1961) zastosowali stosunkowo rozciągnięty w czasie BB, poprzedzany przez dwa BW, z których jeden (A) występował zaraz przed pojawieniem się BB, a drugi (B) zaraz przed jego zakończeniem. Bodziec A okazał się być później pozytywnie, a bodziec B negatywnie powiązany z BB. Narzucającym się wyjaśnieniem tej obserwacji w kategoriach przewidywania BB jest stwierdzenie, że na podstawie sygnałów ze środowiska organizm uczy się przewidywać nie tylko moment rozpoczęcia, ale także moment zakończenia oddziaływania BB.

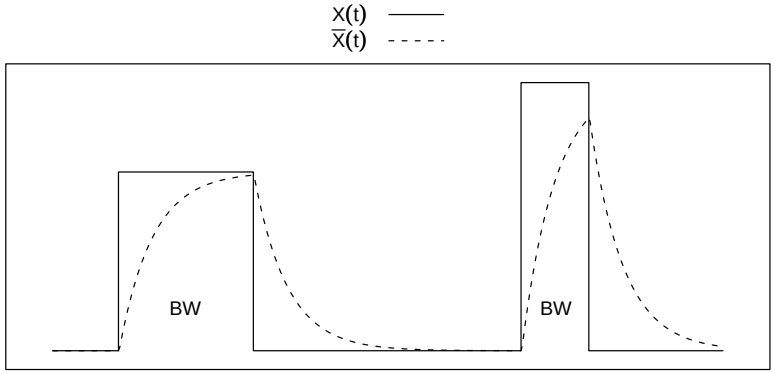
Komputacyjna teoria warunkowania Suttona i Barto należy do teorii czasu rzeczywistego i jest oparta na założeniu, że wzmocnienie determinują rozgrywające się w trakcie trwania prób *zmiany* w poziomie BB, a nie sama jego obecność[7]. Zgodnie z tą teorią, pojawienie się BB generuje dodatnią, a ustąpienie ujemną wartość λ. Zakłada się też, że wszystkie bodźce, a nie tylko bezwarunkowe, generują w ten sposób wzmocnienie, tyle że w przypadku bodźców warunkowych ulega ono zmianie w wyniku uczenia się. Oparty na początkowej wersja tej teorii (Sutton i Barto, 1981) model $\dot{Y}$ redukuje się do modelu R-W dla odpowiednich przypadków, pozwala jednak poprawnie przewidzieć między innymi efekty warunkowania drugiego rzędu. Jako model czasu rzeczywistego jest też znacznie ogólniejszy.

Z powodu elementarnych ograniczeń, teoria nadal nie pozwala *wyjaśnić* między innymi efektów ISI, chociaż pozwala je *modelować*. Wiadomo, że warunkowanie zachodzi również wtedy, gdy pojawienie się BB następuje z pewnym odroczeniem (ISI) po zakończeniu BW. Tempo uczenia się i asymptotyczny poziom reagowania zwykle okazują się być systematycznie związane z tym odroczeniem. Efekty zależne od ISI można modelować za pomocą parametru α, zakładając, że zmiany we wzmocnieniu wywołane przez rozpoczęcie i zakończenie oddziaływania bodźców utrzymują się przez czas wystarczający do wywołania zmian asocjacji. Pozbawiona założenia, że zdarzenia w środowisku pozostawiają jakiś „ślad" w organizmie, teoria warunkowania nie może wyjaśnić efektów ISI większego od zera.

Zgodnie z teorią Suttona i Barto, zaraz po pojawieniu się BW powstaje pewna wewnętrzna reprezentacja, która pozostaje stała i trwa aż do momentu zakończenia oddziaływania BW. Wraz z pojawieniem się tej reprezentacji powstaje ślad relewancji, rosnący stopniowo do asymptoty dopóki BW jest reprezentowany. W momencie, gdy BW przestaje być reprezentowany, ślad zaczyna stopniowo zanikać. Wykrycie wystąpienia i cha-

[7] O ile mi wiadomo, pierwszą sformalizowaną teorią warunkowania czasu rzeczywistego kładącą nacisk na znaczenie rozgrywających się wewnątrz prób zmian we wzmocnieniu była teoria redukcji popędu Hulla (Hull, 1943).

rakteru zmian w stymulacji musi zająć trochę czasu, a więc ślad nie może osiągać maksymalnego poziomu od razu. Gdyby ślad zanikał natychmiast po ustąpieniu oddziaływania BW, organizm byłby bezradny wobec zależności między zdarzeniami następującymi po sobie z pewnym odroczeniem. Zależność między wystąpieniem BW, jego reprezentacją i śladem przedstawia schematycznie poniższy wykres:



Rysunek 8.6: Reprezentacja wewnętrzna i ślad bodźca warunkowego w teorii warunkowania Suttona i Barto

Reprezentację wewnętrzną $(X(t))$ zaznaczyłem linią ciągłą, a ślad $(\bar{X}(t))$ przerywaną. Wartość śladu obliczyłem tak samo jak Sutton i Barto, posługując się wzorem $\bar{X}(t+1) = \bar{X}(t) + \delta(X(t) - \bar{X}(t))$. Zgodnie z tym, co piszą autorzy, na poziomie komputacyjnym ślad można moim zdaniem rozumieć jako świadectwo na rzecz rzeczywistego, niekoniecznie przyczynowo-skutkowego związku między BW, a zdarzeniami następującymi od momentu jego pojawienia się.

Drugim, obok śladu, kluczowym elementem omawianej teorii jest założenie, że tak samo jak w przypadku BW, istotną ze względu na przyjęty sens warunkowania informację dotyczącą poziomu BB można przedstawić za pomocą pewnej krzywej, konkretnie krzywej przewidywanego poziomu BB. Krzywa ta powinna zdaniem autorów stopniowo

rosnąć w miarę zbliżania się do momentu pojawienia się BB i stopniowo maleć w miarę zbliżania się do momentu jego ustąpienia. Przewidywania organizmu powinny więc dotyczyć powierzchni pod tą krzywą, liczoną od momentu dokonywania predykcji. Obszar pod krzywą będzie zależał od intensywności, czasu trwania i odroczenia BB, dzięki czemu pojawia się możliwość wyjaśnienia efektów wszystkich tych czynników. Obszar ten byłby zatem dokładnie wartością wzmocnienia pierwotnego.

Formalnie, dla uproszczenia zakładając, że czas jest dyskretny, prawdziwa wartość nadchodzącego wzmocnienia w momencie t ma być sumą wszystkich przyszłych momentarnych wartości λ, czyli:

$$\bar{V}_t = \lambda_{t+1} + \lambda_{t+2} + \lambda_{t+3} + \cdots$$

Sutton i Barto trafnie spostrzegli, że nie można zakładać, jakoby uczenie się dotyczyło tylko wydarzeń w danej próbie. Podział na próby jest przecież sztuczny i istnieje jedynie w umyśle eksperymentatora. Co dokładnie miałby przewidywać organizm? Autorzy twierdzą, że organizm nie może po prostu przewidywać sumy wszystkich przyszłych wartości BB. Ich zdaniem wartości te muszą być jakoś zdyskontowane, przewidywana wartość powinna być więc dana przez:

$$\bar{V}_t = \lambda_{t+1} + \gamma\lambda_{t+2} + \gamma^2\lambda_{t+3} + \cdots$$

Jak łatwo zauważyć, wyrażenie to przypomina definicję wartości w ramie pojęciowej uczenia się ze wzmocnieniem. Ta idealna predykcja wygląda podobnie do śladu relewancji[8], z tą różnicą, że na wykresie 8.6 trzeba odwrócić strzałkę czasu. Należy przy tym pamiętać, że wagi dyskontowe (γ^k) będą się przesuwały wraz z przesuwaniem się momentu, dla którego predykcja jest określana, przez co krzywa ulegnie odpowiednio „spłaszczeniu" lub „uwypukleniu".

Organizm nie zna idealnej wartości predykcji, ale może ją szacować. Błąd predykcji w momencie t, który można wykorzystać do skorygowania siły asocjacji, dany jest teraz przez $\lambda_{t+1} + \gamma\bar{V}_{t+1} - \bar{V}_t$. Zastępując parametr λ w modelu R-W przez $\lambda_{t+1} + \gamma\bar{V}_{t+1}$ uzyskujemy uogólnioną wersję modelu R-W. Jednostką czasu nie jest pojedyncza próba, a więc można modelować i wyjaśniać dynamikę zmian w asocjacjach rozgrywającą się w trakcie każdej próby. Wprost wyrażone jest również znaczenie szerszego niż zakres pojedynczej próby horyzontu czasowego, bez czego rozsądna teoria warunkowania obyć się nie może. W połączeniu z modelem śladów relewancji daje to razem różnicowo-czasowy model warunkowania Suttona i Barto:

$$\Delta V_i = \beta(\lambda_{t+1} + \gamma\bar{V}_{t+1} - \bar{V}_t) \times \alpha_i\bar{X}_i$$

Model przewiduje poprawnie wszystkie znane efekty, które poprawnie przewiduje model R-W. Nie może być inaczej, skoro model R-W można uzyskać z modelu Suttona i

[8]Równania opisujące ślad i wzmocnienie różnią się zasadniczo, dlatego dokładny kształt krzywej nie będzie ten sam, chociaż jakościowo, po odwróceniu kierunku, oba rodzaje krzywych są bardzo podobne.

Barto jako przypadek szczególny. O ile mi wiadomo, pozwala też wyjaśnić wszystkie albo prawie wszystkie efekty, z którymi radzą sobie co bardziej udane, alternatywne modele czasu rzeczywistego (Dayan i Niv, 2008; Sutton i Barto, 1990; Redish, Jensen, Johnson i Kurt-Nehlson, 2007; O'Reilly i in., 2007), radzi sobie dobrze z niektórymi znanymi efektami, których nie przewidują poprawnie żadne z tych modeli, a specjalnie w tym celu przeprowadzone eksperymenty dostarczyły rezultatów zgodnych z wynikającymi z tego modelu, nowymi i nieoczywistymi predykcjami. Wyniki niektórych badań (Montague i in., 1996; Dayan i Sejnowski, 1994; Dayan i Niv, 2008) zdają się świadczyć, że algorytm różnic czasowych jest realizowany bezpośrednio przez pewne mechanizmy neuronalne, takie jak aktywność fazowa neuronów dopaminowych, co może być jedną z ważniejszych przyczyn dającego się od pewnego czasu zaobserwować wzrostu popularności modelu. Komputacyjna interpretacja i prostota modelu pozwala traktować predykcje jak przyzwoite wyjaśnienia. Teoria nie wyjaśnia wszystkiego, co w ogóle wiadomo na temat warunkowania, niemniej jak dotąd jest to najbardziej udany model warunkowania klasycznego, zatem teoria komputacyjna, z której ten model wynika, jest najlepszym dostępnym psychologicznym wyjaśnieniem warunkowania. Teoria warunkowania Suttona i Barto posiada jednak pewne poważne ograniczenia.

8.4.4 Ograniczenia teorii warunkowania Suttona i Barto

Nie ulega wątpliwości, że teoria Suttona i Barto jest komputacyjna w przyjętym w tej pracy znaczeniu. Model wyprowadzony jest w oparciu o założenia dotyczące tego, jakie zadanie organizm rozwiązuje i co powinien robić, żeby rozwiązać je w przybliżeniu optymalnie, niektóre ważne elementy teorii wprowadzone są jednak ad hoc. Nie chodzi o to, że jak na razie nie są w niej uwzględnione wyraźnie czynniki, wymienione zresztą przez Suttona i Barto, o których wiadomo, że odgrywają ważną rolę w procesie warunkowania, takie jak wyrazistość, uwaga, efekty konfiguracji bodźców, czy uczenie się, jak się uczyć. Teoria jest nadal mocno problematyczna jako teoria komputacyjna, nawet gdy pominiemy te ograniczenia.

Cel nie jest wyrażony w modelu jawnie, przez co proces warunkowania nie jest wprost powiązany z zachowaniem. Wydaje się, że niewiele potrzeba, aby usunąć ten mankament. Trudno nie zgodzić się na interpretację odruchów biologicznych, takich jak mrużenie oczu (działanie) w odpowiedzi na podmuch powietrza (stan), zgodnie z którą reakcja bezwarunkowa powinna dostarczać wyższą nagrodę niż brak takiej reakcji. Zamiast mówić o *obserwowanych* zmianach w łącznej sile asocjacji, albo „sile oczekiwania" wzmocnienia w danym momencie, powiedzielibyśmy, że zmianie ulega obserwowana polisa probabilistyczna. Wyjaśnieniem tych zmian byłby proces celowego uczenia się, czyli poszukiwania polisy maksymalizującej oczekiwaną wartość. Organizm uczyłby się minimalizować koszty (co jest równoważne z maksymalizowaniem wartości), wynikające z nieadekwatnego reagowania na bodźce warunkowe (polisa). Nieadekwatne znaczyłoby tutaj przedwczesne (koszty samej reakcji), spóźnione, lub niedostatecznie intensywne

(koszty niewystarczającego lub spóźnionego zmrużenia oczu).

Reinterpretacja teorii Suttona i Barto z perspektywy zaproponowanej przez tych samych autorów ramy pojęciowej uczenia się ze wzmocnieniem wydaje się jak dotąd nie wymagać żadnych formalnych zmian w modelu. Teraz dopiero, gdy ustalony został związek ze zdefiniowanym wcześniej zachowaniem, teoria zaczyna być jawnie psychologiczna. Zmiana terminologii nie jest do końca kosmetyczna - to co wcześniej było zakładane niejawnie (cel) można teraz poddać krytycznej analizie i zapytać na przykład, czy funkcja nagrody powinna być stała, czy może powinna się zmieniać zależnie od okoliczności zewnętrznych, zmęczenia organizmu i tak dalej.

Dodatkową zaletą tej reinterpretacji jest możliwość wyprowadzenia wniosków na temat ograniczeń teorii warunkowania na podstawie ograniczeń ogólniejszej ramy pojęciowej i opisanych dotąd rozwiązań problemu uczenia się ze wzmocnieniem. Taka właśnie analiza jest najbardziej pożądana, ponieważ dotyczy możliwie abstrakcyjnego poziomu teraz już jawnie psychologicznej teorii, wnioski będą więc dotyczyły nie tylko teorii warunkowania. Jakie są zatem wady omówionej teorii warunkowania, wymienionych dotychczas rozwiązań problemu uczenia się ze wzmocnieniem, a być może nawet samej ramy pojęciowej?

8.4.5 Ograniczenia ramy pojęciowej uczenia się ze wzmocnieniem

Na początek warto zauważyć, że teoria Suttona i Barto nie określa rozwiązania optymalnego, co czyni ją teorią komputacyjną, ale nie racjonalną. Dowiadujemy się tylko, co przypuszczalnie należy zrobić, aby zbliżyć się do bliżej nieokreślonego, optymalnego rozwiązania. Nie wiemy co należy zrobić, aby zbliżać się do niego możliwie najszybciej. Nie znamy rozwiązania optymalnego, a więc nie wiemy jeszcze, na czym dokładnie polega zadanie, a co za tym idzie nie rozumiemy zbyt dobrze funkcji, tak samo jak nie zrozumieliśmy zbyt dobrze funkcji eksploracji, zanim nie poznaliśmy (w zarysie) optymalnego rozwiązania przetargu między eksploracją i eksploatacją.

W ramie pojęciowej uczenia się ze wzmocnieniem parametr siły dyskontowania γ służy jedynie do obejścia technicznego problemu, pojawiającego się w sytuacji nieograniczonego horyzontu czasowego. Trzeba po prostu coś zrobić, żeby uniknąć nieskończonej sumy nagród i parametr γ faktycznie jest pewnym rozwiązaniem tej trudności, ale tylko rozwiązaniem ad hoc, na co zwrócili uwagę niektórzy autorzy (np. Bertsekas, 1995; Hutter, 2004). Rozwiązanie to nie jest bynajmniej niewinne. W zależności od siły dyskontowania agent będzie albo dalekowzroczny, albo krótkowzroczny, a przecież uwzględniany horyzont czasowy powinien zależeć od czegoś innego niż arbitralna decyzja teoretyka. Kiedy Sutton i Barto piszą, że organizm nie może dosłownie uwzględniać całej swojej przyszłości, nie podają żadnego przekonującego uzasadnienia tego twierdzenia. Każdy organizm żywy kiedyś umiera i teoretycznie może i powinien uwzględniać cały (nieznany) horyzont czasowy, dlatego że ten horyzont nie jest nieskończony. Potrzebne jest jakieś inne uzasadnienie dla dyskontowania, jeżeli w ogóle mamy się zgodzić

na dyskontowanie na poziomie teorii komputacyjnej, o teorii racjonalnej nie wspominając.

Naturalnym częściowym uzasadnieniem dyskontowania u organizmów żywych jest związek między horyzontem czasowym a niepewnością. Przebieg interakcji nie jest doskonale przewidywalny i niepewność co do konsekwencji działań rośnie wraz z ich odroczeniem, ponadto wiele dóbr jest nietrwałych. Dodatkowo, jeżeli oczekiwany czas interakcji jest krótki, nie warto dbać o odroczone nagrody. Jeżeli oczekiwany czas interakcji jest długi, dyskontowanie powinno być odpowiednio słabsze. Ludzie są oczywiście wrażliwi na tego rodzaju czynniki zarówno w warunkach laboratoryjnych jak i naturalnych (Ostaszewski, 1997). Wraz z dyskontowaniem pojawia się wiele interesujących zagadnień psychologicznych, takich jak samokontrola i postawy moralne (Rachlin, 2004; Ainslie, 2001), co sprawia, że pojęcie domaga się dokładniejszej analizy. Żeby uzasadnić dyskontowanie, musimy wyposażyć agenta między innymi w zdolność do szacowania czasu interakcji, żeby z kolei uzasadnić racjonalnie konieczność posiadania tej zdolności, musimy tak zdefiniować zadanie, aby czas interakcji w ogóle dawał się oszacować.

Rozwiązaniem ad hoc są też ślady relewancji. Sutton i Barto nie podają żadnego racjonalnego uzasadnienia dla wyrażenia tych śladów w postaci średniej kroczącej, bo też nie jest łatwo sobie takie uzasadnienie wyobrazić. Wydaje się wysoce prawdopodobne, że takie proste ślady relewancji są użyteczne tylko w bardzo ograniczonym zakresie. Racjonalne rozwiązanie powinno polegać na ocenie, które bodźce warunkowe i w jakiej sytuacji są wiarygodnymi kandydatami na predyktory bodźca bezwarunkowego. Wiadomo, że rozmaite bodźce warunkowe różnią się znacznie pod względem tego, jak łatwo można je uwarunkować za pomocą określonych rodzajów bodźców bezwarunkowych, na przykład awersja pokarmowa jest niezwykle podatna na powiązanie z bodźcami zapachowymi, ale już znacznie mniej podatna na powiązanie z czystymi tonami. Główną przyczyną tego i innych powiązanych problemów teoretycznych (takich jak wady początkowych rozwiązań przetargu między eksploracją i eksploatacją) jest nieuwzględnienie niepewności związanej z oszacowaniami, co wynika z nieuwzględnienia w podstawowej ramie pojęciowej modelu środowiska, którym dysponuje agent.

Tak jak dyskontowanie ma rozwiązywać pewien problem związany z przyszłym horyzontem czasowym, tak ślady relewancji mają rozwiązywać problem związany z horyzontem czasowym rozciągającym się wstecz. Ten ostatni problem pojawia się wyraźnie wraz z porzuceniem założenia o własności Markowa. Wyobraźmy sobie, że sygnał stanu jest wielowymiarowym wektorem. Jeżeli zadanie nie ma własności Markowa, konsekwencje działań mogą zależeć od zdarzeń rozgrywających się w krótszym lub dłuższym fragmencie historii interakcji. Załóżmy, że agent próbuje szacować wartości działań lub stanów ze względu na każdy z wymiarów sygnału stanu środowiska. Im większa będzie liczba uwzględnianych wymiarów i im dłuższy będzie rozciągający się wstecz horyzont czasowy interakcji, tym większa będzie niepewność związana z tymi oszacowaniami i tym wolniej ta niepewność będzie redukowana. Problemem jest tutaj statystyczna złożoność stosowanego modelu, związana z liczbą wymiarów sygnału stanu i długością roz-

ciągającego się wstecz horyzontu czasowego. Zgodnie z tym, co napisałem w rozdziale trzecim, agent powinien zaczynać od modelu prostszego, uzyskując w ten sposób lepszą uogólnialność, i powinien rozwijać ten model stopniowo, w miarę otrzymywania kolejnych informacji. To jest jedno z możliwych racjonalnych wyjaśnień ograniczonej pojemności pamięci przemijającej, nie odwołujące się do ograniczeń systemu poznawczego. Wydaje się, że zmieniając kierunek osi czasu i przeprowadzając zbliżone rozumowanie w podobny sposób można próbować częściowo wyjaśnić ograniczenia uwagi selektywnej. Zwracanie uwagi na to, co dzieje się z lewej strony (wymiar sygnału stanu) oznacza, że właśnie z lewej, a nie z prawej strony agent spodziewa się ważniejszych informacji.

Metoda różnic czasowych ma być skutecznym rozwiązaniem dla zadań typu fMDP, ale okazuje się, że z perspektywy agenta takie zadania nie mają własności Markowa. Aktualny stan nie jest wszystkim, co agent potrzebuje wiedzieć, dlatego że konsekwencje działań zależą dodatkowo od polisy, a ta ulega cały czas zmianie. Co więcej, teoria Suttona i Barto wymaga wprowadzenia pojęcia reprezentacji bezpośrednio nieobserwowalnego stanu środowiska. Żeby poprawić i rozwinąć teorię trzeba rozważyć inną klasę środowisk. Analiza wymagań komputacyjnych tej klasy pozwala powiedzieć coś na temat tego, czym jest reprezentacja.

Pojęcie własności Markowa umożliwia także lepsze zrozumienie różnicy między procesami automatycznymi i kontrolowanymi. Jedną z podstawowych, odkrytych przez psychologów własności zadań, decydującej o możliwości automatyzacji ich wykonania jest coś, co niektórzy autorzy (np. Schneider i Schiffrin, 1977; Schiffrin i Schneider, 1977; Ackerman, 1992) nazywają spójnością, czy może raczej stałością wymagań obliczeniowych (ang. *consistency of information-processing demands*), albo relacji między bodźcami i reakcjami. Ackerman (s. 599, 1992) definiuje ją jako „sytuację, w której mapowanie między bodźcami i reakcjami jest takie, że gwarantuje całkowitą pewność co do zależności, gdy ta została już wyuczona". Im mniej spójna jest ta relacja, tym więcej „zasobów uwagi" wymaga zadanie. Korzystając z pojęć własności Markowa i wewnętrznej reprezentacji stanu można poprawić nieco tą definicję. Ani żaden rzeczywisty bodziec, ani żadna stymulacja percepcyjna nie występuje dwukrotnie dokładnie w tej samej postaci i nie ma czegoś takiego, jak całkowita pewność co do zależności między bodźcami i reakcjami, a raczej konsekwencjami reakcji. Możliwa jest jednak wystarczająco mała niepewność co do zależności między spostrzeganym stanem środowiska (a nie bodźcem) a niektórymi konsekwencjami reakcji, aby przeszukiwanie pamięci, rozumowanie, czy inne procesy angażujące „zasoby uwagi" nie były potrzebne ze względu na cele agenta i jego wiedzę o danej sytuacji. Jeżeli agent może nabyć taką reprezentację stanu, że zadanie będzie miało w przybliżeniu własność Markowa, automatyzacja będzie możliwa. W takim razie, znacznie mniej niejasna niż „spójność mapowania między bodźcami i reakcjami", własność Markowa może stanowić element komputacyjnej teorii procesów automatycznych i kontrolowanych.

8.5 Uwagi na temat reprezentacji i znaczenia

Zgodnie z zalecaną przez Pylyshyna (1989) wersją funkcjonalizmu obliczeniowego, o symbolach, na których wykonywane są operacje obliczeniowe można powiedzieć, że coś reprezentują i mają jakieś znaczenie ze względu na ich rolę w procesie obliczeniowym. Jeżeli da się systematycznie interpretować wejścia i wyjścia pewnego mechanizmu jako liczby w taki sposób, że ta relacja między wejściem i wyjściem będzie odpowiadała operacji dodawania, a operacje symboliczne rozgrywające się między wejściem i wyjściem będą zachowywały przyjęte znaczenie wejść i wyjść, to zdaniem Pylyshyna uzasadnione będzie stwierdzenie, że dany mechanizm czy system dodaje liczby. Argumentowałem już wcześniej, że taki system dodaje liczby tylko w znaczeniu metaforycznym, teraz jednak podejmę inny wątek.

O takim mechanizmie obliczeniowym, operującym na kodach symbolicznych, można zdaniem tego autora równoważnie powiedzieć, że operuje na reprezentacjach. Znaczenie wyrażeń złożonych, czyli formalnych struktur złożonych z takich symboli, powinno przy tym w sposób systematyczny zależeć od znaczenia elementów składowych tych struktur. Zgodnie z tym stanowiskiem, tym co czyni kody symboliczne reprezentacjami, a więc czymś, co ma znaczenie, jest ich rola w dającym się odpowiednio zinterpretować procesie obliczeniowym.

Jak trafnie zauważył Harnad (2002), z powyższym stanowiskiem związana jest pewna podstawowa trudność. O takich symbolach nie można jeszcze powiedzieć, że *mają* znaczenie, ale co najwyżej, że znaczenie może zostać im przypisane przez obserwatora. Inaczej mówiąc, znaczenie tych symboli znajduje się w głowach użytkowników mechanizmu obliczeniowego - kalkulator służy do liczenia. Moim zdaniem stwierdzenie, że kalkulator nie działa celowo albo argument Harnada w zupełności wystarczają, aby uznać utożsamienie znaczenia z rolą w odpowiednio zinterpretowanym mechanizmie obliczeniowym za błędne. Intuicyjnie, kody symboliczne powinny czerpać znaczenie z jakiegoś innego źródła, to znaczy, jak określa to Harnad, powinny być odpowiednio „ugruntowane".

Harnad rozważa dwa podstawowe rozwiązania problemy ugruntowania symbolu. Zgodnie z tak zwanym szerokim rozumieniem znaczenia, aby sprawić, że kody symboliczne lub słowa będą posiadały znaczenie, co według Harnada oznacza, że „są o czymś", należy poszerzyć granice systemu symbolicznego w taki sposób, że cały system będzie zawierał również środowisko, w którym znajdują się desygnaty wyrażeń. Harnad uważa to rozwiązanie za problematyczne i zamiast tego proponuje wąskie rozumienie znaczenia, zgodnie z którym symbole nie mają być połączone z obiektami w środowisku, ale z „proksymalnymi cieniami, które obiekty dystalne rzucają na sensomotoryczne powierzchnie systemu" (s. 146, Harnad, 2002). Symbole, których denotacją są klasy obiektów, muszą być ugruntowane w zdolności do sortowania, nazywania, a ogólnie interreagowania z proksymalnymi sensomotorycznymi projekcjami dystalnych egzemplarzy kategorii w sposób zgodny z ich semantyczną interpretacją, zarówno dla indywidualnych

symboli jak i dla wyrażeń złożonych.

Jak podkreśla autor, nie wszystkie symbole muszą być ugruntowane. Dzięki językowi możliwa jest „kradzież znaczenia", to znaczy nabycie znaczących reprezentacji nie na skutek uczenia się w ramach kosztownej (próby i błędy, ryzykowna eksploracja) interakcji ze środowiskiem, ale „tanio", dzięki możliwości tworzenia nowych znaczeń. Język umożliwia takie przyswajanie znaczeń wyrażeń, których nie dałoby się nabyć dzięki projekcjom sensomotorycznym, na przykład dlatego, że ich desygnat nie istnieje. Harnad twierdzi jednak, że aby język w ogóle mógł działać w ten sposób, pewne symbole muszą być wcześniej ugruntowane.

W moim odczuciu najważniejszą różnicą między stanowiskiem Harnada i pokrewnymi pomysłami a utożsamieniem znaczenia z rolą obliczeniową jest uwzględnienie celowej interakcji ze środowiskiem. Mimo, że autor nie zwraca nadmiernej uwagi na samą celowość, nie może się bez niej obejść pisząc o zdolności do sortowania albo nazywania. Czytelnik domyślił się już zapewne, że nie podzielam poglądów Harnada na temat szerokiej koncepcji znaczenia. Pozwolę sobie w tym miejscu przytoczyć sformułowaną przez tego autora argumentację przeciwko przyjęciu szerokiego rozumienia znaczenia *w psychologii*:

> Oto zaleta szerokiego podejścia: jako że szerokie znaczenie obejmuje zarazem wewnętrzną myśl i jej zewnętrzny obiekt, kompletne wyjaśnienie szerokiego znaczenia niczego by nie pominęło. Gdy w końcu uzyskasz udane wyjaśnienie, nie ma już żadnych więcej dziwnych pytań dotyczących tego, jak myśli są „połączone" z tym co znaczą, ponieważ to, co znaczą, jest już w pewnym sensie częścią tego, czym myśli są. Zaadaptowanie takiego podejścia przez psychologa ma jednak pewne niekorzystne strony, ponieważ wymagałoby to od niego, aby był znacznie więcej niż tylko psychologiem: aby pokryć cały obszar, na jaki rozciągają się myśli, musiałby być autorytetem nie tylko co do tego, co dzieje się w głowie, ale także tego, co dzieje się w świecie.
>
> (s. 144, Harnad, 2002)

Proszę zauważyć, że niejawnie przyjmuje się tu jako oczywiste założenie, zgodnie z którym psychologowie środowiskiem właściwie się nie zajmują. Dalej Harnad omawia niektóre trudności związane z szeroką koncepcją znaczenia:

> (...) gdyby próbować uprawiać „szeroką robotykę", określając nie tylko to, co dzieje się wewnątrz robota, ale także to, co dzieje się w świecie, konieczne byłoby modelowanie zarówno robota jak i jego środowiska. Należałoby zaprojektować wirtualny świat, a następnie scharakteryzować stany robota jako obejmujące jednocześnie jego wewnętrzne stany i [stany]

> wirtualnego świata, w którym jest usytuowany. W tym ekumenicznym obszarze robotyki „usytuowanej" (Hallam i Malcolm, 1994) nie musiałoby to wcale wyglądać na zły pomysł, jednak w istocie szeroka robotyka byłaby niezgodna z metodami robotyki usytuowanej, która podkreśla używanie realnego świata do testowania i kierowania procesem projektowania robota dokładnie dlatego, że tak trudno jest odgadnąć, jaki jest świat, za pomocą jakiegokolwiek wirtualnego modelu tego świata („świat jest swoim najlepszym modelem"). W pewnym momencie problem ramy odniesienia (ang. *frame problem*, Pylyshyn, 1987; Harnad, 1993) zawsze się pojawi; właśnie wtedy [okazuje się, że] źle odgadnięto wystarczająco wiele na temat rzeczywistego świata na etapie projektowania wirtualnego świata robota, na skutek czego robot działający doskonale w świecie wirtualnym okazuje się bezradny w świecie rzeczywistym.
>
> (s. 144-145, tamże)

Psycholog miałby się ograniczyć do tego, co dzieje się „w głowie" i pozostawić środowisko zewnętrzne kosmologom. Zadziwiające jest nie tylko to, że można o psychologach powiedzieć, że nie zajmują się środowiskiem (co tylko częściowo mija się z prawdą), ale że można im tego zabraniać.

Nie zgadzam się z argumentacją Harnada i wydaje mi się, że nie jest trudno ustalić, dlaczego jest błędna. Żeby w ogóle wysłowić pojęcie proksymalnego cienia obiektu dystalnego trzeba skorzystać z pojęcia obiektu dystalnego. Na tym można by spokojnie zakończyć krytyczną ocenę przytoczonych powyżej argumentów, ale z pewnych względów warto im się przyjrzeć nieco dokładniej. Harnad nieustannie próbuje ukryć gdzieś środowisko i zinternalizować znaczenie, jednak w żadnym miejscu nie udaje mu się wyrazić własnej propozycji ograniczając się jedynie do „cienia", zawsze jest to „cień obiektu", albo cień „pewnego rodzaju rzeczy desygnowanych przez symbole" (s. 146, tamże).

Nie jest jasne, czym dokładnie ta koncepcja różni się od koncepcji znaczenia jako roli obliczeniowej. Albo agent jest zredukowany do wejść, wyjść i tego, co się pomiędzy nimi rozgrywa, albo trzeba się zgodzić na znaczenie rozumiane szeroko. Zgodnie z życzliwszą choć naciąganą interpretacją, przypisanie znaczenia reprezentacjom wymaga zidentyfikowania pewnych własności procesu celowej interakcji ze środowiskiem, które miałyby świadczyć o tym, że „coś w głowie agenta" reprezentuje pewne obiekty. To z kolei wymaga ustalenia, co jest, a co nie jest celową interakcją ze środowiskiem, czego po prostu nie da się zrobić nie korzystając z modelu środowiska. Stąd, że modelowanie środowiska nie jest zadaniem trywialnym, w żaden sposób nie wynika jeszcze, że należy z niego zrezygnować. Stąd, że wirtualne modele środowiska często okazują się mylące, nie wynika jeszcze, że nie należy poszukiwać lepszych modeli. Wirtualne modele środowiska bywają mylące, ponieważ brakuje wystarczająco dobrej ogólnej teorii środowisk jako elementu celowej interakcji. Psycholog powinien chcieć, żeby problem ramy odniesienia się

pojawił, żeby móc chociaż spróbować go rozwiązać. Postępowanie psychologa zgodne z „duchem robotyki usytuowanej" odpowiadałoby przeprowadzaniu badań w warunkach naturalnych, najpierw jednak trzeba określić, jakie warunki są naturalne i dlaczego, co wymaga przyjęcia takiej lub innej teorii środowiska. Bez choćby implicitnej psychologicznej teorii środowiska nie wiadomo, jakie elementy środowiska naturalnego należy brać pod uwagę opisując proces celowej interakcji.

Wiele doniosłych zalet komputacyjna teoria warunkowania zawdzięcza temu, że jest oparta na abstrakcyjnej teorii zadania, a nie na konkretnym modelu środowiska. Dzięki temu na przykład możliwe jest zidentyfikowanie pewnych fundamentalnych własności (własność bycia środowiskiem stacjonarnym, dyskretnym lub ciągłym, deterministycznym lub niedeterministycznym, własność bycia procesem decyzyjnym Markowa, częściowo obserwowalnym procesem decyzyjnym Markowa i tak dalej), ze względu na które określone rozwiązanie (warunkowanie) jest jeszcze rozwiązaniem stosunkowo skutecznym. Teoria środowiska przydaje się nie tylko do konstruowania konkretnych modeli środowiska na potrzeby określonych badań (analiza racjonalna w ujęciu Andersona), ale przede wszystkim do poszukiwania takich właśnie podstawowych własności środowisk lub zadań, wyznaczających granice między ogólnymi klasami rozwiązań problemu celowej interakcji ze środowiskiem. Granice między funkcjami poznawczymi mogą być wyznaczane przez różnice w wymaganiach komputacyjnych ogólnie scharakteryzowanych klas zadań. Teoria środowiska dostarczana przez kosmologię nie odpowiada potrzebom poznawczym psychologii. Tym co jest potrzebne, bez czego w rzeczy samej nie można się ostatecznie obejść, jest psychologiczna teoria środowiska.

Opierając się na niektórych spostrzeżeniach Harnada można przynajmniej próbować stworzyć aksjomatyczno-dedukcyjną, ogólną psychologiczną teorię reprezentacji. Wypada zgodzić się z założeniem, że pewne reprezentacje są ugruntowane. Reprezentacje interesują psychologię o tyle, o ile ostatecznie pozwalają rozwiązać problemy pojawiające się w ramach celowej interakcji ze środowiskiem. Reprezentacje ugruntowane będą miały znaczenie podstawowe niezależnie od tego, czy „kradzież semantyczna" jest bez nich możliwa czy nie. Reprezentacje ugruntowane będą miały znaczenie podstawowe nawet niezależnie od tego, czy bez takich reprezentacji możliwy jest język, ponieważ nie wszystkie organizmy zdolne do celowej interakcji i reprezentowania czegoś muszą posiadać język. Reprezentacje pojawiają się już na poziomie warunkowania (najlepsze dostępne wyjaśnienie) i nie muszą być przyklejone do żadnych „symboli w głowie". Kolejne ważne spostrzeżenie Harnada dotyczy związku między obecną w pewnych systemach zdolnością do reprezentowania a kategoryzacją:

> Takie systemy hybrydowe kładą silny nacisk na szczególną zdolność (ang. *capacity*), którą wszyscy posiadamy, a jest to zdolność do kategoryzowania, do układania chaosu docierającego do naszych powierzchni sensomotorycznych w relatywnie uporządkowane rodzaje taksonomiczne, określone przez (ang. *marked out by*) nasze zróżnicowane *reakcje* - włączając w

> to wszystko, od reakcji instrumentalnych, takich jak jedzenie, uciekanie od, manipulowanie lub dobieranie się w pary z takimi a nie innymi rodzajami rzeczy, po przypisywanie unikalnych, arbitralnych nazw pewnym rodzajom rzeczy, a nie innym (Harnad, 1987).
>
> Łatwo zapomnieć, że nasza zdolność do kategoryzacji jest w rzeczy samej zdolnością sensomotoryczną. W przypadku reakcji instrumentalnych, w oparciu o Gibsonowskie niezmienniki „udostępnione" (ang. *„afforded"*) przez nasze interakcje sensomotoryczne ze światem, mamy tendencję do zapominania, że niearbitralne ale zróżnicowane reakcje są właściwie aktami kategoryzacji, dzielącymi wejścia na takie, z którymi robi się pewne rzeczy i takie, z którymi robi się coś innego. Również w przypadku jednoznacznie kategoryzacyjnych aktów nazywania zapominamy, że to jest także transakcja sensomotoryczna, chociaż taka, która jest zapośredniczona przez reakcję arbitralną (symbol) zamiast niearbitralnej.
>
> (s. 147, tamże)

Akty kategoryzacji nie dzielą wejść na takie, z którymi robi się pewne rzeczy i takie, z którymi robi się coś innego w znaczeniu o które przypuszczalnie chodzi autorowi. Z wartościami na wejściu nic się nie robi, tylko się je odbiera, ewentualnie później przetwarza, a robi się pewne rzeczy dopiero z obiektami w środowisku. Harnad zdaje się w tym fragmencie sugerować, że zdolność do reprezentowania i kategoryzowania nie wymaga zdolności do posługiwania się symbolami, a więc nie wymaga języka. Ogólnie rozumiane kategoryzowanie byłoby w takim razie zdolnością do wyróżnienia obiektów w świecie, umożliwiającą zachowywanie się w taki sposób, jakby te obiekty były traktowane jako odrębne elementy, z którymi organizm wchodzi w sensowne interakcje, co brzmi całkiem rozsądnie.

Jeżeli te spostrzeżenia są trafne, można sformułować teorię reprezentacji ugruntowanych bez odwoływania się wprost do mechanizmu obliczeniowego i symboli, na których operacje obliczeniowe są wykonywane, nie można jednak pominąć elementu celowej interakcji ze środowiskiem. Reprezentowanie byłoby wtedy własnością agenta działającego w środowisku, a nie symboli znajdujących się „w jego głowie". Dzięki takiej teorii reprezentacji znaczenie nie będzie pochodziło od obserwatora, będzie ugruntowane. Tego rodzaju teoria reprezentacji nie wymaga w ogóle otwierania pudełka *A*. Pudełko *A* trzeba natomiast otwierać, jeżeli mamy się czegoś dowiedzieć o mechanizmie obliczeniowym odpowiadającym za zdolność do reprezentowania, a niewątpliwie chcemy się tego dowiedzieć, a nawet wydaje się, że trochę już wiemy. Dotychczasowe wnioski wskazują jednak na to, że niebagatelna część psychologicznej teorii reprezentacji może powstać bez zaglądania do tego pudełka.

Jednym z powodów, dla których jak dotąd zagadnienie reprezentacji nie pojawiło się z całą mocą w trakcie omawiania ramy pojęciowej uczenia się ze wzmocnieniem było, odrzucone niejawnie na etapie omawiania komputacyjnej teorii warunkowania, założenie o

doskonałej obserwowalności środowiska. Do tej pory sygnały stanu otrzymywane przez agenta pozwalały na jednoznaczną identyfikację stanu świata. Na skutek porzucenia tego założenia, zdolność do reprezentowania zaczyna domagać się bliższych wyjaśnień.

8.5.1 Częściowo obserwowalny proces decyzyjny Markowa i jego psychologiczny sens

Częściowo obserwowalny proces decyzyjny Markowa (w skrócie POMDP, ang. *Partially Observable Markov Decision Process*) jest uogólnieniem procesu decyzyjnego Markowa. Zakłada się, że dynamika jest określona przez proces Markowa, jednak agent nie obserwuje bezpośrednio stanów środowiska, tylko próbki pochodzące z rozkładu sygnału stanów. Nagrody otrzymywane są za przejścia między stanami, a nie za przejścia między sygnałami stanów, optymalne rozwiązanie musi więc polegać na wnioskowaniu wprost lub nie wprost o nieobserwowalnym stanie środowiska. Ważna klasa optymalnych rozwiązań POMDP polega na poszukiwaniu polisy określonej nie na stanach środowiska, które nie są bezpośrednio obserwowalne, tylko na „przekonaniach" co do tych stanów. Dzięki rozróżnieniu na stan i sygnał stanu POMDP pozwala modelować znacznie więcej rzeczywistych procesów sekwencyjnego podejmowania decyzji niż MDP, w szczególności pozwala uchwycić kluczową dla psychologa różnicę między stymulacją percepcyjną a jej źródłem, albo obiektem i jego sensomotorycznym cieniem. Formalnie, skończony POMDP uzyskujemy z fMDP dodając skończony zbiór możliwych obserwacji O i rozkład prawdopodobieństwa $P(o|s,a),\ o \in O$, gdzie a jest działaniem, które doprowadziło do stanu s. Pozostałe elementy zadania nie ulegają zmianie.

W trosce o cierpliwość Czytelnika i będąc świadomym własnych skromnych kompetencji zdecydowałem się nie pisać o optymalnych albo obiecujących rozwiązaniach tego typu zadań. Stosunkowo przystępny przegląd znajduje się między innymi u Murphy'ego (2000) i Loveyoja (1991). Wspólną cechą znanych mi rozwiązań jest to, że chociaż wydają się interesujące jako punkt wyjścia dla komputacyjnej lub racjonalnej teorii reprezentacji ugruntowanej, z powodów, które wkrótce przedstawię, przypuszczalnie nie są wystarczające do uchwycenia specyfiki zjawisk intencjonalnych. Agent, o którym cały czas mówimy, czegoś chce, ale nie chce niczego konkretnego - stara się tylko żeby było lepiej i nie jest łatwo sprawić, aby podobnie jak celowość, konkretność tego czegoś nie pochodziła w znacznym stopniu od autora rozwiązania.

Założenie własności Markowa jest psychologicznie akceptowalne tylko w szczególnych okolicznościach. Dla psychologa bodziec ma, a w każdym razie powinien mieć, zawsze dwa końce, z których jeden to wartość na wejściu, a drugi to źródło tej wartości. Organizmy żywe nie posiadają zdolności bezpośredniego wglądu w stan środowiska, czymkolwiek by on nie był, przez co zmuszone są do nieustannego rozwiązywania problemów znacznie bardziej wymagających, niż problemy decyzyjne Markowa. Stan środowiska naturalnego ulega ciągłym zmianom i znajdujące się gdzieś pod powierzchnią sensomotorycznych cieni środowisko nie jest stacjonarne, co wprowadza kolejny jako-

ściowy wymiar komputacyjnych wymagań celowej interakcji. Opisane wcześniej obiecujące rozwiązania problemu uczenia się ze wzmocnieniem są wobec takich wymagań w ogólnym przypadku bezradne.

Łatwo można o tym wszystkim zapomnieć. Podkreślając znaczenie interakcji ze środowiskiem, autorzy tacy jak Brooks albo Van Gelder sugerowali niedawno, że język reprezentacji i procesów poznawczych nie jest potrzebny do wyjaśnienia inteligentnego zachowania. Brooks zaczął budować roboty, które na początku wykazywały tylko najprostsze „zdolności behawioralne", dodając później kolejne zdolności, na bazie tych wcześniej stworzonych. Sześcionożny robot (Brooks, 1989) może być najpierw wyposażony w „zdolność do wstawania" i pozostawania w pozycji stojącej, niezależnie od słabszych lub silniejszych wstrząsów. Potem dodaje się mechanizm kontroli, który pozwala przemieszczać się naprzód. Jeżeli robot upadnie, zadziała mechanizm wstawania. Dodając kolejne warstwy kontroli i wyposażając robota w coraz bardziej wyrafinowane zdolności, nie korzysta się wcale z obliczeniowych mechanizmów percepcji, pamięci, rozumowania czy planowania. Stworzenie takiego robota nie jest łatwe i nie pomagają w tym szczególnie narzędzia obliczeniowe sztucznej inteligencji albo teorie psychologiczne, nad czym na pewno warto się zastanowić. Wyjaśnieniem tych wyrafinowanych zdolności jest złożony mechanizm, na który składa się środowisko i hierarchicznie skonstruowane ciało. Mechanizm działa w znacznym stopniu dlatego, że wykorzystuje to, co ze względu na konkretne ciało robota umożliwia konkretne środowisko. Dokonania Brooksa są znakomitym przykładem realizacji idei robotyki usytuowanej.

Jedną z korzyści płynących z dysponowania dostatecznie ogólnym, komputacyjnym aparatem pojęciowym jest możliwość zakwestionowania takich rewolucyjnych poglądów w kilku krótkich zdaniach. Roboty Brooksa nie działają celowo, ponieważ są reaktywne. Sposób ich działania jest całkowicie zdeterminowany przez aktualny stan środowiska, celowo i inteligentnie działają tylko ich twórcy. Systemy reaktywne mogą poprawnie działać (ale nie mogą się zachowywać, bo nie działają celowo) tylko w środowisku posiadającym własność Markowa. Ludzie i zwierzęta potrafią rozwiązywać zadania znacznie bardziej wymagające niż procesy decyzyjne Markowa, bo nie są demonami Laplace'a. Żeby względnie skutecznie rozwiązać zadanie, które tej własności nie posiada, trzeba dysponować pamięcią i w jakiś sposób reprezentować historię interakcji. Pomysł, żeby zbudować humanoidalnego robota pozbawionego zdolności do reprezentowania środowiska, historii interakcji i do planowania (Brooks i Stein, 1993) wydaje się równie rewolucyjny co nierozsądny.

Inny interakcjonista, Van Gelder, również postuluje możliwość wyjaśnienia inteligentnego zachowania bez pomocy odwoływania się do reprezentacji i procesów poznawczych. Van Gelder (1995; 1998, 1995), podobnie jak Skarda i Freeman (1987), Thelen i Smith (1994) i Ward (2002) upodobał sobie formalną teorię systemów dynamicznych. System dynamiczny to coś, co posiada stan określony przez wektor rzeczywisty, będący punktem w pewnej przestrzeni stanów. Ustalona reguła ewolucji takiego systemu określa następstwo stanów w czasie, zachowanie jest więc określone przez trajektorię w

przestrzeni stanów. Za wyjątkiem najprostszych systemów dynamicznych, tej trajektorii nie da się dokładnie poznać, a możliwości jej przewidywania są bardzo ograniczone (Ekeland, 1988).

Do niezaprzeczalnych zalet ujęcia procesu interakcji z perspektywy teorii połączonych systemów dynamicznych należy zaakcentowanie nieprzewidywalności takich systemów i spojrzenie na rozgrywający się w czasie proces jako na subtelną trajektorię. Możliwe, że psychologom przydałby się nowy sposób myślenia o dynamice systemów, ale mówiąc delikatnie, nie jest prawdą, jak wprost stwierdza na przykład Ward (2002), że psychologowie czy kognitywiści ignorują dynamiczny wymiar zachowania i procesów psychicznych, a teorie psychologiczne są zwykle „statyczne". Zdaniem Van Geldera (1995), wyjaśnienia „inteligencji" w kategoriach połączonych systemów dynamicznych mogą zastąpić wyjaśnienia w kategoriach mechanizmów obliczeniowych i reprezentacji. Wyjaśnienia oparte na teorii połączonych systemów dynamicznych mogą być jednak tylko wyjaśnieniami dynamiki systemów. Na poziomie komputacyjnym połączone systemy dynamiczne nie wytrzymują porównania z zadaniem dwurękiego bandyty, bo w systemach dynamicznych jako takich nie ma nic, co nadawałoby się do wyrażenia celowości interakcji.

Wbrew temu, co napisał Cummins, nie da się wykluczyć możliwości odkrycia ogólnej, aksjomatyczno-dedukcyjnej, psychologicznej teorii zachowania, procesów i stanów poznawczych. Dotychczasowe rezultaty stosowania analizy racjonalnej wskazują na to, że taka teoria jest nie tylko możliwa, ale wręcz powstaje na naszych oczach dokładnie dzięki temu, że przedmiot psychologii jest specyficzny. Jak sądzę, dopóki teoria dotyczy mechanizmu, który właściwie rozumiany jest tak naprawdę abstrakcyjnie ujętą fizjologią mózgu, rozumowanie dedukcyjne nie na wiele się przyda. Cummins miał rację pisząc o wątrobie (ewolucja nie działa celowo w przyjętym tutaj znaczeniu), chociaż niekoniecznie o będących artefaktami silnikach spalinowych (twórca silnika działa celowo). Procesy i stany poznawcze nie są jednak mechanizmami i stanami obliczeniowymi, tylko czymś, co jest przez te mechanizmy zrealizowane. Unikalna w swoim charakterze teoria identyfikacji architektur Townsenda pozwala czasem względnie przekonująco rozstrzygnąć kilka elementarnych kwestii, ale jej moc wyjaśniająca nie wydaje się szczególnie imponująca. W moim odczuciu Anderson niemal trafił w sedno pisząc, że architektury są tylko systemami zapisu. Typowe teorie mechanistyczne są ulotnymi szkicami, w których można dostrzec wyblakłe i niewyraźne zarysy tego, co racjonalne i tego, co fizjologiczne.

Według Deutscha (1960) istnieją dwa poziomy teorii psychologicznej, które nazwał funkcjonalnym i fizjologicznym. Na podstawie tego, co do tej pory napisałem, nie pozostaje mi nic innego jak stwierdzić, że pomiędzy tymi poziomami nie ma właściwie nic, a teorie poziomu funkcjonalnego znajdują poprawny wyraz dopiero w teoriach racjonalnych. Niezależnie od poziomu abstrakcji opisu, fizjologia interesuje psychologa poznawczego tylko o tyle, o ile można jej nadać interpretację komputacyjną, a warunkiem trafności tej interpretacji jest dysponowanie najpierw opisem i zrozumieniem zadania. Do najważniejszych i w rzeczywistości cały czas wykorzystywanych źródeł przesłanek i

intuicji teoretycznych w psychologii poznawczej należy wymóg przybliżonej racjonalności, rozumianej jako optymalne rozwiązanie takich zadań. Specyficzną cechą przedmiotu psychologii, która umożliwia formułowanie uzasadnionych hipotez na drodze rozumowania dedukcyjnego jest celowość działania. Nie potrafię wskazać żadnego innego, równie solidnego punktu zaczepienia, od którego można by próbować rozpocząć analizę zagadnienia intencjonalności.

8.6 Uwagi na temat intencjonalności

Żeby podać choćby częściowe wyjaśnienie zjawisk intencjonalnych, czyli między innymi udzielić poprawną odpowiedzi na pytanie, dlaczego ktoś chce czegoś, a nie czegoś innego, dlaczego myśli o tym, o czym myśli, albo dlaczego zwraca uwagę na coś, a nie na coś innego, nie trzeba jeszcze dysponować teorią intencjonalności. Wyjaśnienie może wtedy polegać na podaniu przypuszczalnych motywów i przekonań, które zdecydowały o treści postaw propozycjonalnych. Ogólna teoria intencjonalności staje się moim zdaniem potrzebna wtedy, gdy chcemy odpowiedzieć na pytania typu dlaczego zjawiska intencjonalne są takie, a nie inne. Nie potrafię podać zbyt wielu dobrych przykładów takich pytań (dlaczego stany mentalne są o czymś, a nie o niczym albo o wszystkim? dlaczego tworzą kontekst intensjonalny, a nie ekstensjonalny?), stąd moje uwagi na temat intencjonalności będą jeszcze bardziej spekulatywne, pobieżne i niejednoznaczne, niż wszystko co do tej pory napisałem.

Pomijając takie niepokojące czarne dziury, które od pewnego czasu zaczynają się wypełniać treścią, jak brak społeczności agentów[9], omówione do tej pory klasy środowisk mają jeszcze jedną zasadniczą wadę. Stany albo sygnały stanów są określone przez ich miejsce w abstrakcyjnej, niekoniecznie deterministycznej strukturze przyczynowo-skutkowej, która jedynie w bardzo ograniczonym stopniu posiada coś, z czego korzystają organizmy żywe i roboty Brooksa, a właściwie ich konstruktorzy.

Jednym z pierwszych ważnych sukcesów aplikacyjnych metod rozwiązania problemu uczenia się ze wzmocnieniem był stworzony przez Tesaura (1995) algorytm grający na poziomie eksperckim w backgammona. W przypadku tej gry, założenie o własności Markowa jest w przybliżeniu spełnione (źródłem błędu przybliżenia jest obecność przeciwnika i sam proces uczenia się). Backgammon nie jest jednak arbitralnym skończonym procesem decyzyjnym Markowa, tylko bardzo szczególnym przypadkiem takiego procesu. Reguły tej gry rządzą się określoną logiką, wynikającą z jej struktury. Ta struktura umożliwia transfer tego, co zostało wyuczone, na sytuacje, które wcześniej się nie pojawiły. Rozwiązanie Tesaura działa przyzwoicie dzięki temu, że zastosowany przez niego algorytm różnic czasowych został uzupełniony ad hoc o stosunkowo dobrze uogólniającą się sztuczną sieć neuronową. O ile mi wiadomo, nie pojawiły się dotąd pomysły do-

[9]Krytyczny przegląd rozwiązań opartych na metodach uczenia się ze wzmocnieniem dla systemów wieloagentowych znajduje się między innymi u Shohama, Powersa i Grenagera (2003).

tyczące tego, jak można by sobie radzić ze strukturą zadań w ogólnym przypadku, które byłyby równie eleganckie jak opisana wcześniej rama pojęciowa uczenia się ze wzmocnieniem. Próby polegające na uzupełnieniu klasycznych metod o elementy hierarchicznej reprezentacji polis i abstrakcję czasową (Precup, Sutton i Singh, 1998), abstrakcję nad stanami (np. Hengst, 2003; Otterlo, 2004), albo mniej lub bardziej arbitralne metody wnioskowania na modelach w moim odczuciu nie są wystarczające[10].

Żeby takie rozwiązanie nie było rozwiązaniem ad hoc i żeby dostarczało wglądów w to, co umożliwia i czego wymaga specyficzna struktura zadań, trzeba moim zdaniem poszukiwać ogólnej teorii struktury zadań. Można powiedzieć, że rozwiązanie, takie jak to zaproponowane przez Tesaura, na podstawowym poziomie sprowadza się do skorzystania ze spostrzeganych podobieństw i różnic między „sytuacjami". Ponad opisem w kategoriach abstrakcyjnego, stosunkowo arbitralnego procesu decyzyjnego znajduje się być może jakiś opis w kategoriach abstrakcyjnej struktury podobieństw, który mógłby być użyteczny do analizy pojęcia „sytuacji", „obiektu", czy „własności środowisk", takich jak choćby własności przestrzenne.

Treściami aktów intencjonalnych są między innymi przedmioty, zjawiska i własności wyróżnione, jak mi się zdaje, jako elementy niearbitralnej, chciałoby się powiedzieć „sensownej" struktury środowiska. Zgodnie z przedstawionym w tej pracy w sposób rozproszony stanowiskiem funkcjonalizmu racjonalnego, podstaw dla ogólnej, psychologicznej teorii intencjonalności mogłaby przypuszczalnie dostarczać abstrakcyjna teoria struktury środowiska, ujętej z perspektywy komputacyjnych wymagań skutecznej celowej interakcji.

8.7 Wnioski końcowe

Co by się stało, gdyby *zdefiniować* funkcje poznawcze jako optymalne rozwiązania pewnych zadań, pojawiających się w ramach celowej interakcji abstrakcyjnie rozumianego agenta z abstrakcyjnie rozumianym środowiskiem? Zachowanie, percepcja, uwaga, pamięć, uczenie się, podejmowanie decyzji, rozwiązywanie problemów, kategoryzacja i inne nie byłyby procesami, które zachodzą u ludzi, zwierząt i być może niektórych sztucznie stworzonych systemów, tylko niektóre procesy, które zachodzą w ramach interakcji tych organizmów i systemów ze środowiskiem byłyby do pewnego stopnia procesami percepcji, uwagi, pamięci, uczenia się, podejmowania decyzji, rozwiązywania problemów lub kategoryzacji. Teoria określałaby ich wyidealizowany, optymalny przebieg. Nie oznaczałoby to wcale przypisywania tym rzeczywistym systemom doskonałej racjonalności, ani powtarzania błędów psychologizmu.

Takie podejście maksymalizowałoby szanse na identyfikację możliwie prostych zależności, dostarczając czasem modeli o minimalnej osiągalnej w psychologii złożoności, a więc też maksymalnej osiągalnej mocy predykcyjnej. Korzystalibyśmy w ten sposób

[10]Przeglądu zagadnień i metod dostarczają Kaelbling, Littman i Moore (1996) i Glorennec (2000).

jawnie z przesłanek teoretycznych, które i tak odgrywają kluczową, chociaż zwykle niejawną rolę na etapie wstępnego zawężania przestrzeni alternatywnych hipotez, dotyczących bezpośrednio nieobserwowalnego mechanizmu obliczeniowego. Próbując ustalić, które ograniczenia wynikają ze struktury i sposobu działania umysłu, a które są konsekwencją natury samych zadań, staralibyśmy się unikać formułowania mechanistycznych wyjaśnień tych regularności, które mają swoje źródło przede wszystkim w wymaganiach celowej interakcji.

Zarówno wyższe jak i niższe funkcje poznawcze, badane zwykle albo w izolacji, albo poprzez tworzenie bardziej wyczerpujących, skomplikowanych modeli obliczeniowych o trudnych do ustalenia właściwościach i opisywane za pomocą rozlicznych trudnych do uzgodnienia terminów można by rozpatrywać jako szczególne przypadki podstawowej funkcji kontroli, próbując przy tym odkryć najbardziej użyteczną postać teorii podejmowania decyzji w warunkach stale zmieniającego się, złożonego środowiska zewnętrznego i wewnętrznego, co dawałoby przynajmniej nadzieję na powstanie teorii możliwie prostej i ogólnej, obejmującej zarazem naturalną (psychologia) i sztuczną inteligencję. Poszukiwalibyśmy wtedy uniwersalnych i ścisłych praw tam, gdzie pojawiają się naturalnie, czyli na poziomie formalnej teorii aksjomatyczno-dedukcyjnej, a nie wśród obserwowalnych regularności w działaniu systemów, których jedną z cech charakterystycznych jest przecież niezwykła elastyczność i ograniczona przewidywalność.

I tak musimy stale zadawać pytanie o rozwiązanie optymalne. Interesuje nas nie tylko to, co faktycznie taki lub inny agent robi, ale też to, co w ogóle może zrobić i od czego zależy skuteczność realizacji rozmaitych zadań. Musimy zadawać to pytanie dlatego, że dzięki temu możemy lepiej zrozumieć, na czym te zadania polegają. Musimy zadawać pytanie o rozwiązanie optymalne, ponieważ sens wielu pojęć psychologicznych sprawia, że identyfikacja ich desygnatów wydaje się wymagać przyjęcia założenia o przybliżonej racjonalności na poziomie działania całego systemu. Potrzebujemy abstrakcyjnej teorii środowisk, ponieważ nasze teorie dotyczące procesów poznawczych są tylko tak dobre, jak zadania, które stosujemy, żeby je testować, ponieważ nie wiemy zbyt dobrze, czym jest środowisko z perspektywy działających celowo, rzeczywistych agentów, czyli środowisko opisane w języku nadającym się do udzielania odpowiedzi na część ważnych pytań stawianych przez psychologię poznawczą, nie wiemy też z góry, jakie rzeczywiście cele te systemy realizują.

Na poziomie racjonalnej teorii celowej interakcji ze środowiskiem możemy próbować udzielić odpowiedzi na pytanie o naturę zachowania, percepcji, uwagi, pamięci, uczenia się, podejmowania decyzji, rozwiązywania problemów, kategoryzacji i innych funkcji. Dopiero na tym poziomie możemy uchwycić ich sens i w tym znaczeniu poziom teorii racjonalnej jest wyjątkowy. Na tym poziomie możemy powiedzieć co agent robi, myśli i na co zwraca uwagę, na poziomie mechanizmu możemy powiedzieć jak to się dzieje, że coś robi, myśli, albo zwraca na coś uwagę. Nie da się określić mechanizmu realizującego funkcję nie rozstrzygając wcześniej, choćby niejawnie, na czym ona polega. Teleologiczna rola elementów składowych może być oczywista tylko w przypadku tych

systemów, które sami stworzyliśmy dla realizacji niektórych naszych celów.

Nie sposób na tym etapie przewidzieć, jakie byłyby konsekwencje przesunięcia na większą niż dotychczasowa skalę akcentu z poziomu obliczeniowego i implementacyjnego na komputacyjny, docenienia środowiska jako pełnoprawnego przedmiotu teorii psychologicznej, a badań teoretycznych jako równie ważnych jak empiryczne. Powody wydają mi się jednak na tyle trudne do zignorowania, a potencjalne korzyści tak atrakcyjne, że nie potrafię oprzeć się wrażeniu nieuchronności nadchodzących zmian. Wydaje mi się wręcz, że już się rozpoczęły.

Literatura cytowana

Ackerman, P. L. (1992). Predicting individual differences in complex skill acquisition: Dynamics of ability determinants. *Journal of Applied Psychology*, *77*, 598-614.

Aczél, J. (1966). *Lectures on functional equations and their applications*. New York: Academic Press.

Ainslie, G. (2001). *Breakdown of will*. Cambridge University Press.

Akaike, H. (1973). Information theory as an extension of the maximum likelihood principle. W: B. N. Petrov i N. Csaki (Red.), *Second international symposium on information theory*. Budapest: Akademiai Kiado.

Allan, L. G. (1980). A note on measurement of contingency between two binary variables in judgment tasks. *Bulletin of the Psychonomic Society*, *15*, 147-149.

Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Erlbaum.

Anderson, J. R. (1991a). The adaptive nature of human categorization. *Psychological Review*, *98*(3), 409-429.

Anderson, J. R. (1991b). Is human cognition adaptive? *Behavioral and Brain Sciences*, *14*, 471-484.

Anderson, J. R. (1991c). The place of cognitive architectures in rational analysis. W: K. V. Lehn (Red.), *Architectures for intelligence: The 22nd Carnagie Mellon symposium on cognition* (s. 1-24). Hillsdale, NJ: Erlbaum. (Oryginalna praca opublikowana w roku 1988)

Anderson, J. R., Bothell, D., Byrne, M. D. i Douglass, S. (2004). An integrated theory of the mind. *Psychological Review*, *111*, 1036-1060.

Anderson, J. R., i Lebiere, C. (1998). *The atomic components of thought*. Mahwah, NJ: Erlbaum.

Anderson, J. R., i Sheu, C.-F. (1995). Causal inferences as perceptual judgments. *Memory and Cognition*, *23*, 510-524.

Angell, J. R. (1907). The province of functional psychology. *Psychological Review*, *14*, 61-91.

Ashby, F. G. (1982). Deriving exact predictions from the cascade model. *Psychological Review*, *89*, 599-607.

Ashby, F. G., i Townsend, J. T. (1980). Decomposing the reaction time distribution: Pure insertion and selective influence revisited. *Journal of Mathematical Psychology*, *21*(2), 93-123.

Atkinson, R. C., Holmgren, J. E. i Juola, J. F. (1969). Processing time as influenced by the number of elements in visual display. *Perception & Psychophysics*, *6*, 321-326.

Balakrishnan, J. D., MacDonald, J. A., Busemeyer, J. R. i Lin, A. (2002). *Dynamic signal detection theory: The next logical step in the evolution of signal detection analysis* (Raport Techniczny Nr. 248). Pod adresem `http://www.cogs.indiana.edu/publications/techreps2002/248/`

Barsalou, L. W. (2008). Grounded cognition. *Annual Review of Psychology*, *59*, 617-45.

Bellman, R. E. (1957). *A problem in the sequential design of experiments*. Sankhya.
Berger, J. O. (1980). *Statistical decision theory and bayesian analysis* (2 Wyd.). New York: Springer-Verlag.
Berlyne, D. E. (1960). *Conflict, arousal and curiosity*. McGraw-Hill.
Berlyne, D. E. (1963). Motivation problems rised by exploratory and epistemic behavior. W: W. S. Koch (Red.), *Psychology: A study of science* (Tom. 5). New York: McGraw-Hill.
Bertsekas, D. P. (1995). A counterexample to temporal differences learning. *Neural Computation*, *7*, 270-279.
Birke, L. I. A., i Archer, J. (1983). Some issues and problems in the study of animal exploration. W: J. Archer i L. I. A. Birke (Red.), *Exploration in animals and humans*. New York: Van Nostrand Reinhold.
Block, N. (1980). Troubles with functionalism. W: N. Block (Red.), *Readings in philosophy of psychology*. Harvard University Press.
Block, N. (2006). What is functionalism? W: D. M. Borchert (Red.), *The Macmillan encyclopedia of philosophy* (2 Wyd., s. 756). Thomson Gale. (Oryginalna praca opublikowana w roku 1996)
Block, N., i Fodor, J. A. (1972). What psychological states are not. *Philosophical Review*, *81*(2), 159-181.
Bogacz, R., Brown, E., Moehlis, E., Holmes, P. i Cohen, J. D. (2006). The physics of optimal decision making: a formal analysis of models of performance in two-alternative forced choice tasks. *Psychological Review*, *113*, 700-765.
Bones, A. K., i Johnson, N. R. (2007). Measuring the immeasurable: Or „could Abraham Lincoln take the Implicit Association Test?". *Perspectives on Psychological Science*, *2*, 406-411.
Brentano, F. (1999). *Psychologia z empirycznego punktu widzenia*. Warszawa: Wydawnictwo Naukowe PWN.
Broadbent, D. E. (1984). Perception and communication. *Citation Classics*, *23*, 134.
Brooks, R. A. (1989). A robot that walks: emergent behaviors from a carefully evolved network. *Neural Computation*, *1*, 253-262.
Brooks, R. A., i Stein, L. A. (1993). *Building brains for bodies*. Cambridge: Memo 1439, MIT Artificial Intelligence Lab.
Browne, M. (2000). Cross-validation methods. *Journal of Mathematical Psychology*, *44*, 108-132.
Brunswik, E. (1955). Representative design and probabilistic theory in a functional psychology. *Psychological Review*, *62*, 193-217.
Buechner, J. (2007). *Gödel, Putnam, and Functionalism*. Cambridge: MIT Press.
Buehner, M. J., Cheng, P. W. i Clifford, D. (2003). From covariation to causation: A test of the assumption of causal power. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*, 1119-1140.
Busemeyer, J. R., i Townsend, J. T. (1993). Decision field theory: A dynamic-cognitive

approach to decision making in an uncertain environment. *Psychological Review*, *100*, 432-459.
Bush, R. R., i Mosteller, F. (1951). A model for stimulus generalization and discrimination. *Psychological Review*, *58*, 413-423.
Butler, R. A. (1953). Discrimination learning by rhesus monkeys to visual-exploration motivation. *Journal of Comparative and Physiological Psychology*, *46*, 95-98.
Byrne, M. D. (2007). Local theories versus comprehensive architectures: The cognitive science jigsaw puzzle. W: W. D. Gray (Red.), *Integrated models of cognitive systems* (s. 431-443). Oxford University Press.
Carella, J. (1985). Information processing rates in the elderly. *Psychological Bulletin*, *98*, 67-83.
Chamberlin, T. C. (1965). The method of multiple working hypotheses. *Science*, *148*, 754-759.
Chase, V. M., Hertwig, R. i Gigerenzer, G. (1998). Visions of rationality. *Trends in Cognitive Sciences*, *2*, 206-213.
Chater, N., i Oaksford, M. (1999). Ten years of the rational analysis of cognition. *Trends*, *3*, 57-65.
Chater, N., i Oaksford, M. (2000). The rational analysis of mind and behavior. *Synthese*, *122*, 93-131.
Chater, N., Tenenbaum, J. B. i Yuille, A. (2006a). Probabilistic models of cognition: Conceptual foundations. *Trends in Cognitive Science*, *10*(7), 287-291.
Chater, N., Tenenbaum, J. B. i Yuille, A. (2006b). Special issue on „probabilistic models of cognition". *Trends in Cognitive Sciences*, *10*.
Chater, N., i Vitányi, P. M. B. (2003). The generalized universal law of generalization. *Journal of Mathematical Psychology*, *47*, 346-369.
Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological Review*, *104*, 367-405.
Clancey, W. (1997). *Situated cognition: On human knowledge and computer representation*. New York: Cambridge University Press.
Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, *49*(12), 997-1003.
Cummins, R. E. (1975). Functional analysis. *The Journal of Philosophy*, *72*, 741-760.
Cummins, R. E. (2000). „How Does It Work" versus „What Are the Laws?": Two conceptions of psychological explanation. W: F. Keil i R. A. Wilson (Red.), *Explanation and cognition* (s. 117-145). Cambridge: MIT Press.
Davidson, D. (1992). Zdarzenia mentalne. W: B. Schwarcman-Czarnota (Red.), *Eseje o prawdzie, języku i umyśle* (s. 163-193). Wydawnictwo Naukowe PWN. (Oryginalna praca opublikowana w roku 1984)
Dawid, A. P. (1984). Statistical theory: The prequential approach. *Journal of the Royal Statistical Society*, *147*, 278-292.

Dayan, P., i Niv, Y. (2008). Reinforcement learning: The good, the bad and the ugly. *Current Opinion in Neurobiology*, *18*, 1-12.

Dayan, P., i Sejnowski, T. J. (1994). TD(λ) converges with probability 1. *Machine Learning*, *14*, 295-301.

Dearden, R., Friedman, N. i Russell, S. (1998). Bayesian Q-learning. W: *AAAI/IAAI* (s. 761-768). Madison, Wisconsin: AAAI Press.

Dennett, D. C. (1978). *Brainstorms: Philosophical essays on mind and philosophy*. Brighton: Harvester.

Dennett, D. C. (1989). *The intentional stance*. Cambridge: The MIT Press.

Deutsch, J. A. (1960). *The structural basis of behavior*. Chicago: University of Chicago Press.

Dickinson, A. (1980). *Contemporary animal learning theory*. Cambridge: Cambridge University Press.

Dimitrakakis, C. (2006). Nearly optimal exploration-exploitation decision thresholds. W: *ICANN (1)* (s. 850-859).

Donders, F. C. (1969). On the speed of mental processes. W: W. G. Koster (Red.), *Attention and performance II.* North-Holland. (Oryginalna praca opublikowana w roku 1869)

Dunwoody, P. T. (2006). The neglect of the environment by cognitive psychology. *Journal of Theoretical and Philosophical Psychology*, *26*, 139-153.

Dutilh, G., Wagenmakers, E.-J., Vandekerckhove, J. i Tuerlinckx, F. (2008). A diffusion model account of practice. *Manuskrypt w przygotowaniu*.

Dzhafarov, E. N. (1997). Process representation and decompositions of response times. W: A. A. J. Marley (Red.), *Choice, decision and measurement: Essays in honor of R. Duncan Luce* (s. 225-277). Hillsdale, NJ: Erlbaum.

Dzhafarov, E. N. (1999). Conditionally selective dependence of random variables on external factors. *Journal of Mathematical Psychology*, *43*, 123-157.

Dzhafarov, E. N. (2001). Unconditionally selective dependence of random variables on external factors. *Journal of Mathematical Psychology*, *45*, 421-451.

Dzhafarov, E. N. (2003). Selective influence through conditional independence. *Psychometrika*, *68*, 7-25.

Dzhafarov, E. N., i Cortese, J. M. (1996). Empirical recovery of response time decomposition rules I: Sample-level decomposition tests. *Journal of Mathematical Psychology*, *40*, 185-202.

Dzhafarov, E. N., i Schweickert, R. (1995). Decompositions of response times: An almost general theory. *Journal of Mathematical Psychology*, *39*, 285-314.

Dzhafarov, E. N., Schweickert, R. i Sung, K. (2004). Mental architectures with selectively influenced but stochastically interdependent components. *Journal of Mathematical Psychology*, *48*, 51-64.

Ebbinghaus, H. (1987). *Memory: A contribution to experimental psychology*. Dover. (Oryginalna praca opublikowana w roku 1885)

Efron, B., i Gong, G. (1983). A leisurely look at the bootstrap, the jackknife and cross-svalidation. *American Statistician*, *37*, 36-48.

Ekeland, I. (1988). *Mathematics and the unexpected*. Chicago: University of Chicago Press.

Estes, W. K. (2002). Traps in the route to models of memory and decision. *Psychonomic Bulletin & Review*, *9*, 3-25.

Feibleman, J. K. (1954). Theory of integrative levels. *The british journal for the philosophy of science*, *5*, 59-66.

Feller, W. (2006). *Wstęp do rachunku prawdopodobieństwa* (6 Wyd.). Warszawa: Wydawnictwo Naukowe PWN.

Fink, D. (1995). *A compendium of conjugate priors* (Rap. Tech.). Pod adresem `http://www.people.cornell.edu/pages/df36`

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London*, *A-222*, 309-368.

Fodor, J. A. (1965). Explanations in psychology. W: M.Black (Red.), *Philosophy in America*. London: Routledge and Kegan Paul.

Fodor, J. A. (1983). *Modularity of mind: An essay on faculty psychology*. Cambridge: The MIT Press.

Fodor, J. A., i Lepore, E. (1992). *Holism: A shopper's guide*. New York: Wiley-Blackwell.

Fodor, J. A., i Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, *28*, 3-71.

Frassen, B. V. (1980). *The scientific image*. Oxford: Clarendon Press.

Gelder, T. V. (1995). What might cognition be if not computation? *Journal of Philosophy*, *91*, 345-381.

Gelder, T. V. (1998). The dynamical hypothesis in cognitive science. *Behavioral and Brain Sciences*, *21*, 615-665.

Gelman, A., Carlin, J. B., Stern, H. S. i Rubin, D. B. (1995). *Bayesian data analysis*. New York: Chapman & Hall.

Ghirlanda, S., i Enquist, M. (2003). A century of generalization. *Animal Behaviour*, *66*, 15-36.

Gifford, R. (2007). *Environmental psychology: Principles and practice*. Colville, WA: Optimal Books.

Gigerenzer, G., i Goldstein, D. G. (1996). Reasoning fast and frugal way: Models of bounded rationality. *Psychological Review*, *103*(4), 650-669.

Glass, L. A. (1984). Effects of memory set on reaction time. W: J. R. Anderson i S. M. Kosslyn (Red.), *Tutorials in learning and memory* (s. 119-136). W. H. Freeman.

Glorennec, P. Y. (2000). Reinforcement Learning: an overview. W: *European symposium on intelligent techniques* (s. 17-35). Germany: Aechen.

Goodman, N. (1972). Seven strictures on similarity. W: *Problems and projects*. New York: The Bobbs-Merril Co.

Gray, W. D. (2007). Composition and control of integrated cognitive systems. W: W. D. Gray (Red.), *Integrated models of cognitive systems* (s. 3-12). Oxford University Press.

Green, D. M., i Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.

Greenwald, A. G., McGhee, D. E. i Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*, *74*, 1464-1480.

Griffiths, T. L., Kemp, C. i Tenenbaum, J. B. (w druku). Bayesian models of cognition. W: R. Sun (Red.), *Cambridge handbook of computational cognitive modeling.* Cambridge University Press.

Griffiths, T. L., i Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology*, *51*, 354-384.

Grobler, A. (2000). *Prawda a względność.* Kraków: Aureus.

Grobler, A. (2006). *Metodologia nauk*. Kraków: Aureus Znak.

Grünwald, P. D. (2005). Introducing the MDL principle. W: I. J. Myung i M. A. Pitt (Red.), *Advances in minimum description length: Theory and application* (s. 5-22). Cambridge: MIT Press.

Grünwald, P. D. (2007). *The minimum description length principle*. Cambridge: MIT Press.

Guttman, N., i Kalish, H. I. (1956). Discriminability and stimulus generalization. *Journal of Experimental Psychology*, *51*, 79-88.

Hallam, J. C. T., i Malcolm, C. A. (1994). Behavior-perception, action, and intelligence-the view from situated robotics. *Philosophical Transactions of the Royal Society A*, *349*, 29-42.

Harlow, L. L., Mulaik, S. A. i Steiger, J. H. (Red.). (1997). *What if there were no significance test?* Mahawah, NJ: Erlbaum.

Harnad, S. (1987). The induction and representation of categories. W: S. Harnad (Red.), *Categorical perception: The groundwork of cognition.* New York: Cambridge University Press.

Harnad, S. (1993). Problems, problems: The frame problem as a symptom of the symbol grounding problem. *PSYCOLOQUY*, *4*.

Harnad, S. (2002). Symbol grounding and the origin of language. W: M. Scheutz (Red.), *Computationalism: New directions* (s. 143-158). Cambridge: The MIT Press.

Heathcote, A., Brown, S. D. i Mewhort, D. J. K. (2000). The power law repealed: The case for an exponential law of practice. *Psychonomic Bulletin & Review*, *7*, 185-207.

Hempel, C. G. (1966). *The philosophy of natural science.* New York: Prentice-Hall.

Hempel, C. G., i Oppenheim, P. (1948). Studies in the logic of explanation. *Philosophy of science*, *15*, 135-175.

Hengst, B. (2003). *Safe state abstraction and discounting in hie-*

rarchical reinforcement learning (Rap. Tech.). Pod adresem `http://www.cse.unsw.edu.au/ bernhardh/`

Huber, P. J. (2004). *Robust statistics*. Wiley.

Hull, C. L. (1943). *Principles of behavior*. Appleton-Century-Crofts.

Hutter, M. (2004). *Universal artificial intelligence: Sequential decisions based on algorithmic probability*. Berlin: Springer.

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, *186*, 453-461.

Jensen, A. R. (2006). *Clocking the mind: Mental chronometry and individual differences*. Amsterdam: Elsevier.

Johnson-Laird, P. N., i Wason, P. C. (1970). A theoretical analysis into a reasoning task. *Cognitive Psychology*, *1*, 134-148.

Kaelbling, L. P., Littman, M. L. i Moore, A. W. (1996). Reinforcement learning: A survey. *Journal of Artificial Intelligence Research*, *4*, 237--285.

Kass, R. E., i Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773-795.

Keller, H., Schneider, K. i Henderson, B. (Red.). (1994). *Curiosity and exploration*. Berlin: Springer-Verlag.

Klauer, K. C., Voss, A., Schmitz, F. i Teige-Mocigemba, S. (2007). Process components of the implicit association test: A diffusion model analysis. *Journal of Personality and Social Psychology*, *93*, 353-368.

Korb, K. B., i Nicholson, A. E. (2003). *Bayesian artificial intelligence*. Chapman & Hall.

Krantz, D. H. (1999). The null hypothesis significance testing in psychology. *Journal of the American Statistical Association*, *44*, 1372-1381.

Krzyżewski, K. (1989). Przyczynek do kwestii wzajemnych relacji aspektów struktury i funkcji w neobehawioryzmie. *Prace Psychologiczne*, *6*, 7-17.

Kukla, A. (1989). Nonempirical issues in psychology. *American Psychologist*, *44*, 785-794.

Kukla, A. (2001). *Methods of theoretical psychology*. Cambridge: The MIT Press.

Lakatos, I. (1978). *The methodology of scientific research programmes*. Cambridge: Cambridge University Press.

Langley, P., Laird, J. E. i Rogers, S. (2009). Cognitive architectures: Research issues and challenges. *Cognitive Systems Research*, *10*, 141-160.

Larson, G. E., i Alderton, D. L. (1990). Reaction time variability and intelligence: A „worst performance" analysis of individual differences. *Intelligence*, *14*, 309-325.

Lashley, K. S., i Wade, M. (1946). The Pavlovian theory of generalization. *Psychological Review*, *53*, 72-87.

Lee, M. D., i Wagenmakers, E.-J. (2003). Bayesian statistical inference in psychology: Comment on Trafimow (2003). *Psychological Review*, *112*, 662-668.

Lewis, D. (1980). Mad Pain and Martian Pain. W: *Readings in philosophy of psychology* (s. 216-222). Cambridge, MA: Harvard University Press.
Li, M., i Vitányi, P. M. (1997). *An introduction to Kolmogorov complexity and its applications* (2 Wyd.). New York: Springer.
Lober, K., i Shanks, D. (2000). Is causal induction based on causal power?: Critique of Cheng (1997). *Psychological Review*, *107*, 195-212.
Loftus, G. R. (2002). Analysis, interpretation, and visual presentation of experimental data. W: J. T. Wixted i H. Pashler (Red.), *Stevens' handbook of experimental psychology* (3 Wyd., Tom. 4, s. 339-390). New York: Wiley Press.
Logan, G. D. (2002). Parallel and serial processing. W: J. T. Wixted i H. Pashler (Red.), *Stevens' handbook of experimental psychology* (3 Wyd., Tom. 4, s. 271-300). New York: Wiley Press.
Lovejoy, W. (1991). A survey of algorithmic methods for partially observed Markov decision processes. *Annals of Operations Research*, *18*, 47-65.
Luce, R. D. (1963). Detection and recognition. W: R. D. Luce, R. R. Bush i E. Galanter (Red.), *Handbook of mathematical psychology* (s. 103-189). New York: Wiley.
Luce, R. D. (1986). *Response times: Their role in inferring elementary mental organization*. Oxford University Press.
Luce, R. D. (1995). Four tensions concerning mathematical modeling in psychology. *Annual Reviews of Psychology*, *46*, 1-26.
MacLeod, C. M. (1991). Half a century of research on the stroop effect: an integrative review. *Psychological Bulletin*, *109*, 163-203.
Macmillan, N. A., i Creelman, C. D. (2002). *Detection theory: A user's guide* (2 Wyd.). Mahawah, NJ: Erlbaum.
Marr, D. (1982). *Vision*. San Francisco: W. H. Freeman.
McClelland, J. L. (1979). On the time relations of mental processes: An examination of systems of processes in cascade. *Psychological Review*, *86*, 287-330.
McCulloch, W. S., i Pitts, W. H. (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, *5*, 115-133.
Meehl, P. E. (1967). Theory testing in psychology and physics: A methodological paradox. *Philosophy of Science*, *34*, 103-105.
Miller, J. O. (1982). Discrete versus continuous stage models of human information processing: In search of partial output. *Journal of Experimental Psychology: Human Perception and Performance*, *8*, 273-296.
Miller, J. O. (1988). Discrete and continuous models of human information processing: Theoretical distinctions and empirical results. *Acta Psychologica*, *67*, 191-257.
Miller, J. O., Ham, F. van der i Sanders, A. F. (1995). Overlapping stage models and reaction time additivity: Effects of the activation equation. *Acta Psychologica*, *90*, 11-28.
Montague, P. R., Dayan, P. i Sejnowski, T. J. (1996). A framework for mesencephalic

dopamine systems based on predictive hebbian learning. *Journal of Neuroscience*, *16*, 1936-1947.

Murphy, K. P. (2000). *A survey of pomdp solution techniques* (Rap. Tech.). http://citeseer.nj.nec.com/murphy00survey.html.

Myung, I. J., Forster, M. i Browne, M. W. (200). A special issue on model selection. *Journal of Mathematical Psychology*, *44*, 1-2.

Myung, I. J., Navarro, D. J. i Pitt, M. A. (2006). Model selection by normalized maximum likelihood. *Journal of Mathematical Psychology*, *50*, 167-179.

Myung, I. J., i Pitt, M. A. (2002). Mathematical modeling. W: J. T. Wixted i H. Pashler (Red.), *Stevens' handbook of experimental psychology* (3 Wyd., Tom. 4, s. 429-459). New York: Wiley Press.

Nachtigall, P. E. (1986). *Vision, audition, and chemoreception in dolphins and other marine mammals* (R. J. Schusterman, J. A.Thomas i F. G. Wood, Red.). Mahwah, NJ: Erlbaum.

Navon, D. (1984). Resources - a theoretical soup stone? *Psychological Review*, *91*, 216-234.

Newell, A. (1973). You can't play 20 questions with nature and win: Projective comments on the papers of this simposium. W: W. G. Chase (Red.), *Visual information processing* (s. 283-308). New York: Academic Press.

Newell, A. (1990). *Unified theories of cognition*. Cambridge: Harvard University Press.

Newell, A., i Rosenbloom, P. S. (1981). Mechanisms of skill acquisition and the law of practice. W: *Cognitive skills and their acquisition* (s. 1-55). Hillsdale, NJ: Erlbaum.

Ng, A. Y., i Russell, S. (2000). Algorithms for inverse reinforcement learning. W: *Proceedings of 17th international conference on machine learning* (s. 663-670). Morgan Kaufmann.

Oaksford, M., i Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review*, *101*, 608-631.

O'Donohue, W., i Buchanan, J. A. (2001). The weaknesses of strong inference. *Behavior and Philosophy*, *29*, 1-20.

O'Reilly, R. C., Frank, M. J., Hazy, T. E. i Watz, B. (2007). Pvlv : The primary value and learned value Pavlovian learning algorithm. *Behavioral Neuroscience*, *121*(1), 31-49.

Ostaszewski, P. (1997). *Zachowanie się organizmów wobec odroczonych wzmocnień*. Warszawa: Wydawnictwo Instytutu Psychologii PAN.

Otterlo, M. V. (2004). Reinforcement learning for relational MDPs. W: *Proceedings of the machine learning conference of belgium and the netherlands*.

Pearl, J. (1998). *Probabilistic reasoning in intelligent systems*. San Francisco: Morgan Kaufmann.

Pearl, J. (2000). *Causality: Models, reasoning and inference*. Cambridge: Cambridge University Press.

Perry, R. B. (1918). Docility and purposiveness. *Psychological Review*, *25*, 1-20.

Piccinini, G. (2004). Functionalism, computationalism, and mental states. *Studies in History and Philosophy of Science*, *35*, 811-833.
Pisula, W. (2003). *Psychologia zachowań eksploracyjnych zwierząt*. Gdańsk: Gdańskie Wydawnictwo Psychologiczne.
Pitt, M. A., Myung, I. J. i Zhang, S. (2002). Toward a method of selecting among computational models of cognition. *Psychological Review*, *109*, 472-491.
Platt, J. R. (1964). Strong inference. *Science*, *146 (3642)*, 347-353.
Popper, K. (1977). *Logika odkrycia naukowego*. Warszawa: PWN.
Port, R. F., i Gelder, T. V. (Red.). (1995). *Mind as motion: explorations in the dynamics of cognition*. Cambridge: The MIT Press.
Precup, D., Sutton, R. S. i Singh, S. (1998). Theoretical results on reinforcement learning with temporally abstract options. W: *Proceedings of the 10th European Conference on Machine Learning*. Springer.
Press, J. (2003). *Subjective and objective bayesian statistics*. Hoboken, NJ: Wiley.
Psotka, J., i Mutter, S. A. (1988). *Intelligent tutoring systems: Lessons learned*. Lawrence Erlbaum Associates.
Putnam, H. (1960). Minds and machines. W: *Dimensions of mind* (s. 148-180). New York: New York University Press.
Putnam, H. (1973). Meaning and reference. *Journal of Philosophy*, *70*, 699-711.
Putnam, H. (1980). The nature of mental states. W: N. Block (Red.), *Readings in philosophy of psychology* (s. 223-231). Cambridge: Harvard University Press. (Oryginalna praca opublikowana w roku 1967)
Putnam, H. (1988). *Representation and reality*. Cabmridge: MIT Press.
Putnam, H. (1999). *The threefold cord: Mind, body, and world*. New York: Columbia University Press.
Pylyshyn, Z. W. (Red.). (1987). *The robot's dilemma: The frame problem in artificial intelligence*. Norwood, NJ: Ablex.
Pylyshyn, Z. W. (1989). Computing in cognitive science. W: M. Posner (Red.), *Foundations of cognitive science* (s. 49-92). Cambridge: MIT Press.
Rachlin, H. (2004). *The science of self-control*. New York: Harvard University Press.
Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*, 59-108.
Ratcliff, R. (1988a). Continuous versus discrete information processing: Modeling the accumulation of partial information. *Psychological Review*, *95*, 238-255.
Ratcliff, R. (1988b). A note on mimicking additive reaction time models. *Journal of Mathematical Psychology*, *32*(2), 192-204.
Ratcliff, R., i McKoon, G. (1997). A counter model for implicit priming in perceptual word identification. *Psychological Review*, *104*, 319-343.
Ratcliff, R., i Rouder, J. R. (1998). Modeling response times for two-choice decisions. *Psychological Science*, *9*, 347-356.
Ratcliff, R., Schmiedek, F. i McKoon, G. (2008). A diffusion model explanation of the worst performance rule for reaction time and IQ. *Intelligence*, *36*, 10-17.

Ratcliff, R., Thapar, A., Gomez, P. i McKoon, G. (2004). A diffusion model analysis of the effects of aging in the lexical-decision task. *Psychology and Aging*, *19*, 278-289.
Ratcliff, R., Thapar, A. i McKoon, G. (2001). The effects of aging on reaction time in a signal detection task. *Psychology and Aging*, *16*, 323-341.
Ratcliff, R., Thapar, A. i McKoon, G. (2003). A diffusion model analysis of the effects of aging on brightness discrimination. *Perception & Psychophysics*, *65*, 523-535.
Ratcliff, R., Thapar, A. i McKoon, G. (2006). Aging and individual differences in rapid two-choice decisions. *Psychonomic Bulletin & Review*, *13*, 626-635.
Read, T. R. C., i Cressie, N. A. C. (1988). *Goodness of fit statistics for discrete multivariate data*. New York: Springer.
Redish, A. D., Jensen, S., Johnson, A. i Kurt-Nehlson, Z. (2007). Reconciling reinforcement learning models with behavioral extinction and renewal: Implications for addiction, relapse, and problem gambling. *Psychological Review*, *114*(3), 784-805.
Rescorla, R. A. (1980). *Pavlovian second-order conditioning*. Erlbaum.
Rescorla, R. A., i Wagner, A. R. (1972). A theory of pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. W: A. H. Black i W. F. Prokasy (Red.), *Classical conditioning II* (s. 64-99). Appleton-Century-Crofts.
Richardson-Klavenh, A., i Bjork, R. A. (1988). Measures of memory. *Annual Review of Psychology*, *39*, 475-543.
Rissanen, J. J. (1986). Stochastic complexity and modelling. *The Annals of Statistics*, *14*, 1080-1100.
Rissanen, J. J. (2003). *Lectures on statistical modeling theory*. www.mdl-research.org.
Robbins, H., i Monro, S. (1951). A stochastic approximation method. *Annals of Mathematical Statistics*, *22*, 400-407.
Robert, C. P. (2001). *The bayesian choice* (2 Wyd.). Springer.
Roberts, S., i Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review*, *107*, 358-367.
Rubin, D. C., Hinton, S. i Wenzel, A. (1999). The precise time course of retention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*, 1161-1176.
Schiffrin, R. M., Lee, M. D., Kim, W. i Wagenmakers, E.-J. (2008). A survey of model evaluation approaches with a tutorial on hierarchical bayesian methods. *Cognitive Science*, *32*, 1248-1284.
Schiffrin, R. M., i Schneider, W. (1977). Controlled and automatic information processing: II. Perceptual learning, automatic attending, and a general theory. *Psychological Review*, *84*, 127-190.
Schneider, W., i Schiffrin, R. M. (1977). Controlled and automatic information processing: I. Detection, search and attention. *Psychological Review*, *84*, 1-66.
Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*, 461-464.

Schweickert, R., i Giorgini, M. (1999). Response time distributions: Some simple effects of factors selectively influencing mental processes. *Psychonomic Bulletin and Review*, *6*, 269-288.

Schweickert, R., i Townsend, J. T. (1989). A trichotomy: Interactions of factors prolonging sequential and concurrent mental processes in stochastic discrete mental (PERT) networks. *Journal of Mathematical Psychology*, *33*, 328-347.

Searle, J. (1980). Minds, brains and programs. *Behavioral and Brain Sciences*, *3*, 417-457.

Segundo, J. P., Galeano, C., Sommer-Smith, J. A. i Roig, J. A. (1961). Behavioral and EEG effects of tones „reinforced" by cessation of painful stimuli. W: J. F. Dalafresnaye (Red.), *Brain mechanisms and learning*. Blackwells Scientific Publishing.

Shepard, R. N. (1987). Towards a universal law of generalization for psychological science. *Science*, *237*, 1317-1323.

Shoham, Y., Powers, R. i Grenager, T. (2003). *Multi-agent reinforcement learning: a critical survey* (Rap. Tech.).

Simon, H. A. (1972). Theories of bounded rationality. W: *Decision and organization*. Amsterdam: North-Holland.

Simon, H. A. (1975). A behavioral model of rational choice. W: *Models of man, social and rational: Mathematical essays on rational human behavior in a social setting*. New York: Wiley.

Simon, H. A. (1991). Cognitive architectures and rational analysis. W: K. V. Lehn (Red.), *Architectures for intelligence: The 22nd Carnagie Mellon symposium on cognition* (s. 25-40). Hillsdale, NJ: Erlbaum. (Oryginalna praca opublikowana w roku 1988)

Skarda, C. A., i Freeman, W. J. (1987). How brains make chaos in order to make sense of the world. *Behavioral and Brain Sciences*, *10*, 161-195.

Sloman, A. (2002). The irrelevance of Turing machines to artificial intelligence. W: M. Scheutz (Red.), *Computationalism: New directions* (s. 87-128). Cambridge: The MIT Press.

Smith, B. C. (2002). The foundations of computing. W: M. Scheutz (Red.), *Computationalism: New directions* (s. 23-58). Cambridge: The MIT Press.

Sorkin, R. D., Hays, C. J. i West, R. (2001). Signal-detection analysis of group decision making. *Psychological Review*, *108*, 183-203.

Sternberg, S. (1966). High-speed scanning in human memory. *Science*, *153*, 652-654.

Sternberg, S. (1969a). The discovery of processing stages: Extensions of Donders' method. *Acta Psychologica*, *30*, 276-315.

Sternberg, S. (1969b). Memory-scanning: Mental processes revealed by reaction-time experiments. *American Scientist*, *57*(4), 421-457.

Sternberg, S. (2001). Separate modifiability, mental modules, and the use of pure and composite measures to reveal them. *Acta Psychologica*, *106*, 147-246.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, *18*, 643-662.
Sutton, R. S., i Barto, A. G. (1981). Toward a modern theory of adaptive networks: Expectation and prediction. *Psychological Review*, *88*, 135-171.
Sutton, R. S., i Barto, A. G. (1990). Time-derivative models of Pavlovian reinforcement. W: M. Gabriel i J. Moore (Red.), *Learning and computational neuroscience: Foundations of adaptive networks* (s. 497-537). Cambridge: MIT Press.
Sutton, R. S., i Barto, A. G. (1998). *Reinforcement learning: An introduction*. Cambridge: MIT Press.
Swets, J. A. (1996). *Signal detection theory and ROC analysis in psychology and diagnostics*. Mahwah, NJ: Erlbaum.
Tenenbaum, J. B., i Griffiths, T. L. (2001). Generalization, similarity and bayesian inference. *Behavioral and Brain Sciences*, *24*, 629-641.
Tesauro, G. (1995). Temporal difference learning and TD-Gammon. *Communications of the ACM*, *38*, 58-68.
Thelen, E., i Smith, L. B. (1994). *A dynamic systems approach to the development of cognition and action*. Cambridge: The MIT Press.
Tolman, E. C. (1995). *Zachowanie celowe u ludzi i zwierząt*. Wydawnictwo Naukowe PWN.
Townsend, J. T. (1972). Some results concerning the identifiability of parallel and serial processes. *British Journal of Mathematical and Statistical Psychology*, *25*, 168-199.
Townsend, J. T. (1974). Issues and models concerning the processing of a finite number of inputs. W: B. H. Kantowitz (Red.), *Human information processing: Tutorials in performance and cognition* (s. 133-186). New York: Erlbaum.
Townsend, J. T. (1984). Uncovering mental processes with factorial experiments. *Journal of Mathematical Psychology*, *28*(4), 363-400.
Townsend, J. T. (1990a). Serial vs. parallel processing: Sometimes they look like Tweedledum and Tweedledee but they can (and should) be distinguished. *Psychological Science*, *1*, 46-54.
Townsend, J. T. (1990b). Truth and consequences of ordinal differences in statistical distributions : Toward a theory of hierarchical inference. *Psychological Bulletin*, *108*(3), 551-567.
Townsend, J. T., i Ashby, F. G. (1983). *Stochastic modeling of elementary psychological processes*. Cambridge: Cambridge University Press.
Townsend, J. T., i Evans, R. (1983). A systems approach to parallel-serial testability and visual feature processing. W: *Modern issues in perception* (s. 166-189). Berlin: VEB Deutscher Verlag der Wissenschaften.
Townsend, J. T., i Fific, M. (2004). Parallel versus serial processing and individual differences in high-speed search in human memory. *Perception & Psychophysics*, *66*(6), 953-962.
Townsend, J. T., Golden, R. i Wallsten, T. (2005). Quantitative training in psycholo-

gy is deteriorating: Traditional methodologists, mathematical psychologists, and psychology face a challenge. *Psychological Science Agenda*, *19*.

Townsend, J. T., i Nozawa, G. (1995). Spatio-temporal properties of elementary perception: An investigation of parallel, serial and coactive theories. *Journal of Mathematical Psychology*, *39*, 321-360.

Townsend, J. T., i Schweickert, R. (1989). Toward the trichotomy method of reaction times: Laying the foundation of stochastic mental networks. *Journal of Mathematical Psychology*, *33*, 309-327.

Townsend, J. T., i Thomas, R. D. (1994). Stochastic dependencies in parallel and serial models: Effects on systems factorial interactions. *Journal of Mathematical Psychology*, *38*, 1-34.

Townsend, J. T., i Wenger, M. J. (2004). A theory of interactive parallel processing: New capacity measures and predictions for a response time inequality series. *Psychological Review*, *111*, Psychological Review.

Townsend, P. (1975). The mind-body equation revisited. W: C.-Y. Cheng (Red.), *Psychological problems in philosophy*. Honolulu: University of Hawaii Press.

Trafimow, D. (2003). Hypothesis testing and theory evaluation at the boundaries: surprising insights from Bayes's theorem. *Psychological Review*, *110*, 526-535.

Treisman, A., i Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, *12*, 97-136.

Treisman, A., i Gormican, S. (1988). Feature analysis in early vision: Evidence from search assymetries. *Psychological Review*, *95*(1), 15-48.

Tversky, A. (1977). Features of similarity. *Psychological Review*, *84*, 327-352.

Voss, A., Rothermund, K. i Voss, J. (2004). Interpreting the parameters of the diffusion model: An empirical validation. *Memory & Cognition*, *32*, 1206-1220.

Wagenmakers, E.-J. (2009). Methodological and empirical developments for the Ratcliff diffusion model of response times and accuracy. *European Journal of Cognitive Psychology*, *21*, 641-671.

Wagenmakers, E.-J., i Brown, S. D. (2007). On the linear relation between the mean and the standard deviation of a response time distribution. *Psychological Review*, *114*, 830-841.

Wagenmakers, E.-J., i Waldrop, L. (2006). Special issue on model selection: Theoretical developments and applications. *Journal of Mathematical Psychology*, *50*.

Wald, A. (1947). *Sequential analysis*. John Wiley and Sons.

Wallace, C. S. (2005). *Statistical and inductive inference by minimum message length*. Springer-Verlag.

Ward, L. M. (2002). *Dynamical cognitive science*. Cambridge: The MIT Press.

Wason, P. C. (1968). Reasoning about a rule. *Quarterly Journal of Experimental Psychology*, *20*, 273-281.

Weeda, W. D., Wagenmakers, E.-J. i Huizenga, H. M. (2007). Empirical support for a

diffusion model account of the worst performance rule. *Manuskrypt w przygotowaniu*.
Wilcox, R. R. (1998). How many discoveries have been lost by ignoring modern statistical methods? *American Psychologist*, *53*, 300-314.
Wilcox, R. R. (2009). *Basics statistics: Understanding conventional methods and modern insights*. New York: Oxford University Press.
Wilkinson, L. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*, 594-604.
Wittgenstein, L. (2001). *Philosophical investigations*. Blackwell Publishing. (Oryginalna praca opublikowana w roku 1953)
Wolfe, J. M. (1994). Guided search 2.0: A revised model of visual search. *Psychonomic Bulletin and Review*, *1*, 202-238.
Wolfe, J. M. (2007). Guided search 4.0: Current progress with a model of visual search. W: W. D. Gray (Red.), *Integrated models of cognitive systems* (s. 99-119). Oxford University Press.
Wolfe, J. M., Cave, R. K. i Franzel, S. L. (1989). Guided search: An alternative to the feature integration model for visual search. *Journal of Experimental Psychology: Human Perception & Performance*, *15*, 419-433.
Zandt, T. V. (2002). Analysis of response time distributions. W: J. T. Wixted i H. Pashler (Red.), *Stevens' handbook of experimental psychology* (3 Wyd., Tom. 4, s. 461-516). New York: Wiley Press.
Zandt, T. V., i Ratcliff, R. (1995). Statistical mimicking of reaction time data: Single-process models, parameter variability, and mixtures. *Psychonomic Bulletin and Review*, *2*(1), 20-54.
Zandt, T. V., i Townsend, J. T. (1993). Self-terminating versus exhaustive processes in rapid visual and memory search: An evaluative review. *Perception & Psychophysics*, *53*, 563-580.

Spis tabel

Spis rysunków