Reprezentacje umysłowe jako aproksymacje stanów mózgu

Włodzisław Duch
Uniwersytet Mikołaja Kopernika w Toruniu
Katedra Informatyki Stosowanej

Streszczenie. *Neuronauki dokonały znacznego postępu w rozumieniu wyższych czynności poznawczych, w tym procesów decyzyjnych. Brakuje jednak zarówno prostych modeli, które pozwolą wyobrazić sobie te procesy, jak i głębszej refleksji nad wpływem tych wyników na zrozumienie natury umysłu, rozproszenia obaw, że nie jesteśmy tylko automatami. Płodny punkt widzenia na kwestię reprezentacji mentalnych daje próba zrozumienia, w jaki sposób informacja reprezentowana jest przez mózgi, jak w przybliżony sposób opisać stany mózgu tak, by można je było zinterpretować jako reprezentacje mentalne odnoszące się do umysłu. Czym jest w nas to, co podejmuje decyzje? Jest to kluczowa kwestia dla zrozumienia wielu zagadnień klasycznej filozofii, od wolnej woli i problemu ciała i umysłu poczynając. Chociaż iluzja* homunkulusa *jest silna można się od niej uwolnić. Analiza procesów podejmowania decyzji wymaga zrozumienia w jaki sposób zakodowane są w mózgu najprostsze pojęcia. Jedynie przez aproksymację fizycznych stanów mózgu, zawierających znacznie więcej informacji niż stany mentalne, możemy dokonać istotnego postępu w rozumieniu i opisie reprezentacji mentalnych, przydatnych nie tylko filozofom, ale też mających zastosowania w analizie języka naturalnego, kategoryzacji pojęć w psychologii i architekturach kognitywnych w sztucznej inteligencji. Ucieleśnienie jak i enaktywizm są dla rozwoju reprezentacji pojęć bardzo istotne ale nie wystarczające. Po analizie współczesnej wiedzy o reprezentacji pojęć w mózgu zaprezentowano podejście oparte o język układów dynamicznych, które oferuje dość prosty opis, pozwalający zrozumieć zaskakujące, często nieracjonalne decyzje podejmowane przez ludzi. Dzięki symulacjom komputerowym można się spodziewać znacznego postępu w rozumieniu reprezentacji mentalnych i wyższych czynności poznawczych, w szczególności procesów podejmowania decyzji.*

Studia z Kognitywistyki i Filozofii Umysłu, **3**: 5–28, 2009

1. Wstęp

Proste organizmy działają w sposób reaktywny, w oparciu o genetycznie uwarunkowane mikroprogramy zachowań. Mucha zamknięta w pokoju lata tak długo, aż zużyje całą dostępną jej energię. Latanie z gwałtowną zmianą kierunku jest zapewne dobrą strategią unikania drapieżników (ptaków, nietoperzy), przy jednoczesnej eksploracji dużego obszaru w sposób częściowo chaotyczny, a częściowo ukierunkowany przez stężenia zapachów. Latanie w zamkniętym pomieszczeniu nie jest jednak dobrą strategią, ale mucha nie potrafi jej zmienić. Na przykładzie insektów widać najlepiej związek pomiędzy wyuczonym a instynktownym zachowaniem, oraz ograniczenia z tego wynikające (De Marco i Menzel 2008). Zdolności komunikacyjne i nawigacyjne pszczół i innych insektów są zaskakująco duże, pomimo stosunkowo prostego układu nerwowego. Bodźce zmysłowe wyzwalają sekwencje mikroprogramów sterujących zachowaniem w ewolucyjnie korzystny sposób, jednakże procesy rozwojowe mogą zmodyfikować te zachowania.

Niewiele wiadomo o szczegółach oddziaływania pomiędzy wrodzonymi a wyuczonymi formami zachowań. U insektów dominują wrodzone formy zachowań, u zwierząt posiadających bardziej złożone mózgi instynktowne zachowania są w coraz większym stopniu modyfikowane, a u ludzi zwykle są dość głęboko ukryte. Jednak nawet u mrówek obserwujemy duże zdolności adaptacyjne wynikające z zależnych od kontekstu i unikalnych dla każdej kolonii bodźców zapachowych, pozwalających na sprawne działanie uwzględniające czas i miejsce pojawienia się określonych bodźców. Oprócz zapachu owady wykorzystują wzrok, smak, dotyk i odczuwanie mechanicznych wibracji. Mrówki porywające larwy innych gatunków w celu pozyskania robotników wykorzystują zjawisko wdrukowania (imprintingu).

Rodney Brooks wprowadził artykułem „Słonie nie grają w szachy" (Brooks 1986) nowe tendencje rozwojowe w robotyce, odrzucając pomysły budowania symbolicznych modeli umysłu opartych na wewnętrznych reprezentacjach stanów świata. Inteligentne zachowania prostych organizmów nie wymagają od nich przechowywania wewnętrznych reprezentacji, chyba że za takie uznać zbiór zależnych od kontekstu reakcji. Brooks próbował pokazać, że można stworzyć inteligentnego robota formując jego mózg, modelowany za pomocą sieci neuronowych, w naturalny sposób przez oddziaływanie otoczenia, podobnie jak rozwija się mózg dziecka. Był to zasadniczy odwrót od wcześniejszych prób budowy sztucznej inteligencji na czysto logicznych podstawach, przy całkowitym ignorowaniu biologii. Głównym projektem, mającym udowodnić słuszność takiego podejścia, miała być budowa robota o nazwie Cog

(Brooks i Stein 1994). Próba ta nie zakończyła się jednak powstaniem umysłu na wzór ludzki, a jedynie zbiorem odruchów pozwalających na proste reakcje. Pomimo tego robotyka rozwojowa stała się obecnie bardzo ważną dziedziną, powiązaną w psychologią rozwojową, a nadzieje na rozwój bardziej złożonych form poznania i działania są nadal silne. Jest to ciekawy eksperyment pomagający wyznaczyć granice pomiędzy wrodzonymi a wyuczonymi umiejętnościami. Jakiejś formy reprezentacji nie da się jednak uniknąć.

W tym artykule spróbuję pokazać, że płodny punkt widzenia na kwestię reprezentacji mentalnych daje próba zrozumienia, jak informacja reprezentowana jest przez mózgi, jak w przybliżony sposób opisać stany mózgu tak, by można je było zinterpretować jako reprezentacje mentalne odnoszące się do umysłu. Zacznę jednak od paru uwag na temat natury umysłu i iluzji *homunkulusa*.

2. *Homunkulus*

Większość badań psychologicznych i rozważań filozoficznych krytycznie podchodziła do neurobiologicznych podstaw procesów podejmowania decyzji i myślenia. Jedną z przyczyn tej niechęci mogła być silna, chociaż rzadko w pełni uświadamiana, wiara w *homunkulusa*, nadrzędne „ja", które pociąga za sznurki i wydaje decyzje realizowane przez mózg. W historii ludzkiej myśli jedynie tradycje, wywodzące się z kontemplacyjnych szkół religijno-filozoficznych Indii, rozwinięte na Dalekim Wschodzie (Chiny, Korea, Japonia), zdołały uwolnić się od iluzji *homunkulusa*. W jednym z najczęściej recytowanych w buddyjskich klasztorach całej Azji Południowo-Wschodniej i Dalekiego Wschodu tekstów, *Maha Pradżnia Paramita Hridaja Sutra* (w swobodnym tłumaczeniu oznacza to „Sutra Serca Wielkiej Mądrości Osiągającego Drugi Brzeg"), podkreślane jest to bardzo mocno (Austin 1988). W tym krótkim tekście Budda wyjaśnia swojemu uczniowi Sariputrze, czym jest wyzwolenie od cierpienia związane z odrzuceniem wszelkich złudzeń. W filozofii indyjskiej (Radhakrishnan 1958) ludzka osobowość uważana była za wytwór pięciu złożonych elementów, czyli *skandh* (dosłownie „skupisk", co wskazuje na przybliżony sposób opisu). Te elementy to forma, uczucia, postrzeganie, wola i świadomość. Mędrzec postrzega, że każda z nich istnieje tylko jako zmienna konfiguracja relacji elementów, pozbawiona rzeczywistej substancji – w tekście Sutry czytamy: „w głębi mądrości ujrzał pustkę wszystkich pięciu *skandh*". Wszystkie sfery zmysłowe i akty odczuwania mają iluzoryczną naturę. Formy przeżywanych wrażeń nie mają trwałej tożsamości, są chwilową konfiguracją pobudzeń mózgu, co w „Sutrze Serca" lapidarnie określa

się jako „forma jest tylko pustką, pustka jest tylko formą". Podobnie pozostałe elementy: „Uczucia, myśli, wola i świadomość sama są również takie". Sutra ta przetłumaczona została na język chiński około 172 roku. W znacznie późniejszym tekście *Hsin hsin ming* („Wersety Wiary w Umysł"), napisanym przez Seng-tsana, trzeciego patriarchę Zen (zmarłego w 606 roku), czytamy (Kapleau 1992, s. 173):

> Rzeczy się jawią za sprawą umysłu,
> za sprawą rzeczy umysł się pojawia.

To umysł pozwala nam odróżnić od siebie rzeczy w świecie, a świat pozwala mu się utworzyć — przyczyna nie da się odróżnić od skutku, gdyż nie ma tu prostej liniowej przyczynowości. W pieśni *Zazen Wasan* Hakuina czytamy (Kapleau 1992, str. 169): „Wtedy brama do jedności przyczyny i skutku zostanie otwarta na oścież". Dzięki głębokiej introspekcji starożytni myśliciele uwolnili się od idei *homunkulusa*, owego „ja" sterującego ciałem i podejmującego decyzje. Odrzucenie *homunkulusa* i uznanie, że człowiek jest jednością psychofizyczną, jest jednak rzeczą bardzo trudną i niektóre tradycje buddyjskie zrobiły tu krok wstecz, głosząc mgliste idee zachowania „strumienia świadomości" (Duch 2006). Powstaje bowiem głęboko zakorzeniona obawa, że jeśli jestem tylko swoim mózgiem i ciałem, to „mnie" tak naprawdę nie ma. Ta obawa jest (nie zawsze uświadomioną) motywacją do poszukiwania alternatywnych rozwiązań problemu podejmowania decyzji. Eccles i Popper w książce *Mózg i jaźń* (1999) głoszą potrzebę dualistycznego rozumienia człowieka twierdząc, że bez niej nie można uzasadnić ludzkiej godności. Echa takiego myślenia widać w wierszu „Zdania" Stefana Mellera (2008): „Trzecie najgłupsze zdanie / Jakie znam to: / — jestem niewierzący — / Bo znaczy tyle co / Wiem, że mnie nie ma". Czym jest w nas to, co podejmuje decyzje? Jest to kluczowa kwestia dla zrozumienia wielu zagadnień klasycznej filozofii, od wolnej woli i problemu ciała i umysłu poczynając, ma też ważne implikacje teologiczne, leży u podstaw dualistycznych wizji człowieka. Chociaż filozofowie, fizycy i neurobiolodzy włożyli dużo wysiłku w opracowanie modeli dualistycznych, żaden z nich nie pogłębia rozumienia procesów umysłowych i nie widać w tym kierunku żadnego postępu.

W ostatnich latach dzięki neuronaukom dokonał się znaczny postęp w rozumieniu wyższych czynności poznawczych, w tym procesów decyzyjnych. Brakuje jednak zarówno prostych modeli, które pozwolą wyobrazić sobie te procesy, jak i głębszej refleksji nad wpływem tych wyników na zrozumienie natury umysłu, rozproszenia obaw, że nie jesteśmy tylko automatami. Badania nad mózgiem nie są zagrożeniem dla godności ludzkiej, a stwierdzenie „mnie nie ma" jest po prostu fałszywe.

Sprowadza się to do ustalenia, czy „ja" i jego decyzje to jeden z wielu zachodzących w mózgu procesów, czy też jest to autonomiczny, nadrzędny, niematerialny czynnik kontrolujący mózg i ciało. Z naukowego punktu widzenia nie ma wątpliwości, że „ja" jest jednym z wielu procesów zachodzących w mózgu. Neurobiologia jaźni rozwija się bardzo szybko (Northoff i Panksepp 2008). Komputerowe modele chorób psychicznych i syndromów neuropsychologicznych (Duch 2007) pozwalają przynajmniej w jakościowy sposób zrozumieć przyczyny patologii i normalnego funkcjonowania mózgu, Nie ma wątpliwości, że dokładniejsze modele procesów nagrody i procesów decyzyjnych pozwolą na odtworzenie rezultatów eksperymentów w tej dziedzinie. Rozumienie autonomiczności jaźni jako jednego z procesów realizowanych przez mózg pozostawia nadal wiele do życzenia. Dlatego pozwolę sobie na kilka uwag na ten temat, chociaż głównym celem tego artykułu jest przedstawienie prostego modelu pozwalającego na zrozumienie relacji pomiędzy neurodynamiką opisującą zachodzące w mózgu procesy, a symboliczną interpretacją procesów decyzyjnych.

Procesy zachodzące w tkankach neuronowych mózgu są warunkiem koniecznym istnienia umysłu i powstania jaźni, podobnie jak warunkiem powstania jakiejkolwiek struktury biologicznej jest jej realizacja w oparciu o związki chemiczne zbudowane z atomów węgla i innych pierwiastków. Jednakże struktura organizmów biologicznych, przyczyna istnienia licznych organów spełniających specyficzne funkcje, jest wynikiem milionów lat rozwoju, adaptacji ewolucyjnych umożliwiających sprawne działanie pozwalające na przetrwanie gatunku w zmiennych, niekorzystnych warunkach. Nie można jej zrozumieć badając samą budowę organizmu, strukturę genomów i białek. Koncepcja *autopoesis* nie oddaje tu istoty rzeczy, gdyż dotyczy tylko systemowej organizacji reprodukujących się układów, podczas gdy nacisk trzeba położyć na ewolucyjny sens emergentnych własności. Analogicznie, badanie procesów neuronowych nie pozwoli w pełni zrozumieć indywidualnego umysłu, którego istotą jest specyficzna struktura relacyjna przyjmowanych przez mózgi stanów. Sensu tych relacji nie da się zrozumieć w oparciu o procesy neurofizjologiczne, bo procesy, które odpowiadają za stany mentalne, mają rację bytu tylko ze względu na istnienie umysłu wynikającego z niepowtarzalnej historii jednostki.

Obserwacja przejścia pomiędzy stanami Ψ_α i Ψ_β mózgu jest tylko zewnętrznym opisem zmian w nim zachodzących; by zrozumieć perspektywę wewnętrzną trzeba odwołać się do historii organizmu, skojarzyć w intencjonalny sposób stany Ψ_α i Ψ_β z kontekstem środowiskowym, w którym podobne stany występowały wcześniej. Np. stan Ψ_1 może wiązać się z wspomnieniem dawnej melodii, wywołując szereg skojarzeń

$\Psi_2, \Psi_3, \ldots, \Psi_\nu$, a więc stanów mentalnych realizowanych przez kolejne pobudzenia mózgu. Zależą one od kultury, w której wychowała się dana osoba, co można zweryfikować za pomocą metod neuroobrazowania (Northoff i Panksepp 2008). Dotyczy to również podstawowych mechanizmów poznawczych: Europejczycy zwracają większą uwagę na obiekty pierwszoplanowe, zapamiętując dotyczące ich szczegóły, Japończycy bardziej na relacje pomiędzy obiektami i całym środowiskiem. Różnice te widoczne są nawet w prostych eksperymentach, w których Europejczycy zapamiętują lepiej położenie pręta w ramce, nie zwracając uwagi na położenie ramki, które wpływa na percepcję Japończyków. Znajduje to odbicie w różnicach aktywności pomiędzy obszarami mózgu odpowiedzialnymi za przetwarzanie informacji o widzianych obiektach (boczna kora potyliczna) oraz o tle, w jakim się pojawiają (zakręt przyhipokampowy). Wspomnienia Chińczyków związane są bardziej z sytuacją społeczną niż ich indywidualną rolą, jak to się dzieje w przypadku ludzi Zachodu.

Poznawcze neuronauki społeczne (ang. *social cognitive neurosciences*) odkrywają wiele takich zależności na różnych poziomach. Szczególnie interesujące są badania wpływu kultury na pojęcia związane z „ja", stanowiące podstawę do autorefleksji i samoświadomości. Ocena, czy dana cecha, wyświetlana na monitorze, pasuje do badanej osoby prowadzi u niej do zwiększonej aktywności brzuszno-przyśrodkowej części kory przedczołowej (VMPFC) oraz przylegającej do niej przykolankowej przedniej części zakrętu obręczy (pACC). Pojęcia odnoszące się w różny sposób do „ja": pamięci autobiograficznej, reakcji emocjonalnych, relacji społecznych, rozpoznawania twarzy, odczuć, że jest się sprawcą działania i innych aspektów „siebie", prowadzą do aktywacji różnych części przyśrodkowych obszarów kory (Northoff i in. 2006). Silniejsza aktywacja VMPFC u ludzi z Zachodu nie obejmuje bliskiej rodziny, podczas gdy u Chińczyków odnosi się zarówno do siebie jak i własnej matki. Nawet przekonania religijne znajdują swoje odzwierciedlenie w różnym rozkładzie pobudzeń pomiędzy brzuszną (VM) i grzbietowo-przyśrodkową (DM) częścią kory przedczołowej (PFC). Aktywność DMPFC wiąże się z teorią umysłu (Frith i Frith 2005), czyli zdolnością do przypisywania stanów mentalnych innym ludziom i widzenia świata z ich perspektywy. Przekonania religijne wydają się więc wpływać na zmniejszenie egocentrycznej perspektywy na rzecz oceny siebie z perspektywy zewnętrznej.

Techniki eksperymentalne używane w badaniach transkulturowych pokazują wpływ środowiska na mechanizmy poznawcze i reprezentacje mentalne determinujące postrzeganie siebie i innych. Ten wpływ może wykraczać poza modulację ustalonych wzorców działania mózgu, może

to być wpływ decydujący o powstaniu pewnych dynamicznych struktur w oparciu o neuronalny substrat. Pod względem ogólnej budowy zdrowe mózgi są do siebie podobne, ich struktura neuroanatomiczna powstała w wyniku długotrwałych procesów ewolucyjnych. Potrzeby organizmu i jego możliwości poznawcze, znajdujące odbicie w strukturze mózgu, stwarzają ramy dla powstania subiektywnego obrazu świata. Jednak struktura połączeń w mózgu, decydująca o możliwych do powstania stanach mentalnych, jest wynikiem indywidualnej historii. Zrozumienie relacji pomiędzy następującymi po sobie stanami mentalnymi możliwe jest tylko przy uwzględnieniu tej indywidualnej historii, włącznie z intencjonalnymi odniesieniami stanów umysłu do stanów świata. Dlatego algorytmiczne wyjaśnianie stanów umysłu nie jest możliwe bez symulacji indywidualnej historii, a więc symulacji stanów całego świata, który się na nią składa. Jeśli nawet uznamy umysł za realizację pewnego algorytmicznego procesu wykonywanego przez mózg, to autonomię decyzji zapewnia jego indywidualna historia, częściowo zapisana w śladach pamięci epizodycznej, semantycznej i proceduralnej, a częściowo wyryta w całej strukturze mózgu, określającej jego potencjalnie dostępne stany dynamiczne, określającej indywidualne cechy osobowości. W każdym konkretnym przypadku za decyzje podejmowane przez człowieka odpowiedzialna jest neurodynamika jego mózgu. Wpływają na nią wszystkie wymienione wyżej czynniki, jak też i stan biochemiczny mózgu, związany z ogólnym stanem organizmu, zdrowiem, pożywieniem, wysiłkiem, snem i wieloma innymi czynnikami. Większość podejmowanych decyzji interpretowana jest jako decyzje „ja", chociaż czasami „ja" zaskakiwane jest przez „swoje" działania, których potem żałuje lub uznaje za nieprzemyślane.

Do rozbieżności dochodzi w kilku sytuacjach. Jeśli mam zbyt mało czasu by skorygować plany działania podsuwane przez mózg zgodnie ze swoimi celami i wartościami mogę powiedzieć lub zrobić coś nieodpowiedniego. Jeśli czas pozwala na refleksję może się okazać, że nie znam sam siebie, zrobię coś, z czego będę niezadowolony, gdyż moje przewidywania własnych reakcji jest niezgodne z rzeczywistością, a więc „ja" ma błędny model przewidywań neurodynamiki mózgu. Może się też pojawić silny przymus wewnętrzny, który pomimo prawidłowej oceny szkodliwości decyzji nie pozwoli na jej zablokowanie, jak to się dzieje w przypadku silnych uzależnień. Można mieć całkiem fałszywe wyobrażenie o sobie i swoich intencjach. Szczególnie zaskakujące są przypadki zaburzeń postrzegania przestrzeni. Heterotopagnosia jest niezdolnością do wskazywania innych osób w przestrzeni pozapersonalnej; pacjenci odnoszą wszystko do siebie, niektórzy nie są zdolni do określenia granic swojego „ja". Poznanie „samego siebie", do którego wzywali starożyt-

ni mędrcy, nie jest łatwym zadaniem, a odwarunkowanie, pozbycie się złudzenia istnienia *homunkulusa*, jest rzeczą nadzwyczaj trudną (Austin 1998, 2006). Dokładna symulacja umysłu jest nieprawdopodobna, można jedynie myśleć o pewnych aproksymacjach, pozwalających zrozumieć procesy podejmowania decyzji.

Taki obraz człowieka jest równie odległy od starożytnych wyobrażeń co świat geocentryczny od współczesnej astronomii. Z jednej strony mamy materialny substrat, ale z drugiej strony niezliczone potencjalne stany relacyjne tego substratu, stany które mają odmienny status ontologiczny. Z perspektywy wewnętrznej mamy świat wirtualny, świat intencjonalnych odniesień do rzeczy, sytuacji czy ludzi, którzy mogą już nie istnieć. Nie jest to świat materialny, chociaż do jego istnienia konieczne są materialne stany mózgu. Człowiek jest nie tylko psychofizyczną całością, jest czymś więcej niż tylko swoim „ja", mózgiem czy organizmem, jest też zogniskowanym odbiciem wielu zdarzeń, które wpłynęły na uformowanie jego umysłu, od relacji rodzinnych, ogólnego wychowania i wykształcenia, po indywidualne i niepowtarzalne przeżycia, które go uformowały. Podmiotowość, autonomia i odpowiedzialność nie odnoszą się do „ja", jednego z wielu procesów realizowanych przez mózg, ale do całego człowieka.

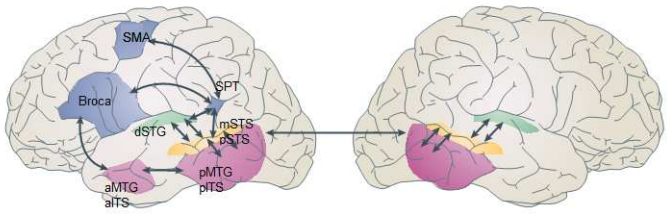
3. Wiedza koncepcyjna w mózgu

Analiza procesów podejmowania decyzji wymaga zrozumienia w jaki sposób zakodowane są w mózgu najprostsze pojęcia. Badania nad reprezentacją symboli są bardzo trudne: badania na zwierzętach niewiele mówią o języku, a bezinwazyjne metody neuroobrazowania nie są dostatecznie precyzyjne by ujawnić dokładne konfiguracje pobudzeń neuronów powstające w czasie używania jakiegoś pojęcia. Mamy jednak dość spójny model, oparty na synchronicznej aktywności grup neuronów (NCA, *neural cell assemblies*), opisujący wyniki badań neurolingwistycznych (Damasio i in. 1996;Pulvermuller 2003; Dehaene i in. 2005), zgodny z ogólnymi mechanizmami pamięci (Lin, Osan, Tsien 2006). Słowa, lub bardziej ogólnie symbole, mają swoją formę (nazwę, wygląd) i swoje znaczenie. Forma symboli reprezentowana jest w mózgu przez silnie ze sobą połączone lokalne podsieci mikroobwodów neuronalnych, które wiążą ze sobą akustyczne (lub wizualne) i artykulacyjne wzorce pobudzeń, pozwalając na ich identyfikację i wymowę. Znaczenie słów reprezentowane jest przez pobudzenia mózgu, skojarzone z ich formą fonologiczną i ortograficzną, łączące ze sobą percepcję i działanie, aktywujące korę zmysłową, ruchową, przedruchową i obszary podkorowe, tworząc rozszerzone reprezentacje pojęć (Pulvermuller 2003).

Rozszerzona część reprezentacji pojęcia jest rozkładem pobudzeń obszarów kory zmysłowej i ruchowej, co można utożsamiać z *qualiami* (interpretacja aktywności kory zmysłowej o różnej intensywności i rozkładzie) i predyspozycjami do działania. Z wewnętrznego punktu widzenia, w przestrzeni aktywacji systemu neuronów zakodowana jest w ten sposób nie tylko sieć pojęciowych relacji, ale również Hebbowskie korelacje pomiędzy postrzeganymi elementami na wielu poziomach, dające pewien model świata postrzeganego przez pryzmat *qualiów*.

Eksperymenty psycholingwistyczne dowodzą, że analogowy sygnał akustyczny, docierający do ucha, zamieniany jest w pierwotnej korze słuchowej na reprezentację fonologiczną, składającą się z dyskretnych elementów. Niewielki zbiór fonemów łączy się ze sobą w czasowo uporządkowany sposób tworząc utrzymujący się przez ułamki sekundy stan rezonansowy, reprezentujący formę słowa. Z badań nad potencjałami wywołanymi wynika, że już po 90 milisekundach następuje aktywacja rozszerzonych podsieci kory zmysłowej a po 200 ms widać aktywność kory ruchowej, która odnosi sens danego słowa do możliwości działania w świecie (Pulvermuller 2003). Rozpoznawanie mówionego słowa lub postrzeganego symbolu jest skomplikowanym procesem, w którym istotną rolę pełni torowanie (zmniejszanie progu pobudliwości grup neuronów) na poziomie morfologicznym, gdy rozpoznane fragmenty (fonemy, grafemy) aktywują wiele rzeczywistych słów jak i słowopodobnych kombinacji. Słowa polisemiczne mają jednakową reprezentację fonologiczną, ale kontekst prowadzi do automatycznej aktywacji odmiennych podsieci, tworząc odpowiednie rozszerzenia semantyczne. Okada i Hickok (2006) kontrastując słowa należące do obszarów o dużej gęstości fonologicznej (podobne ze względu na brzmienie do wielu słów) i słowa o małej gęstości (unikalne brzmienie) pokazali, że łączenie pobudzeń fonologicznych w reprezentacje leksykalne zachodzi w tylnej części bruzdy skroniowej górnej (STS). Kilku autorów, w tym Dehaene i współpracownicy (2005), twierdzili, że w lewej bruździe skroniowopotylicznej mieści się obszar wzrokowej formy słów (*Visual Word Form Area*, VWFA), który reaguje u ludzi potrafiących czytać i pisać tylko przy prezentacji słów w formie pisanej. Jest to nadal twierdzenie kontrowersyjne, krytykowane np. w pracy Price i Devlin (2003). Leżący w pobliżu boczny dolnoskroniowy obszar wielomodalny (*Lateral Inferotemporal Multimodal Area*, LIMA) reaguje zarówno na bodźce słuchowe jak i wzrokowe, i ma projekcje do obszarów fonologicznych jak i rozszerzonych (Gaillard i in. 2006). Jest dość prawdopodobne, że w strumieniu przetwarzającym informacje słuchowe istnieje obszar homologiczny do VWFA, położony w przedniej części lewej bruzdy skroniowej, chociaż dane na ten temat są nadal kontrowersyjne (Dehaene i Naccache

2001; Dehaene i in. 2005). W każdym razie przetwarzanie mowy zmierza (Hickok, Poeppel 2007) od sygnału akustycznego (grzbietowe części STS, bruzdy skroniowej górnej, w obu półkulach) przez dyskretne, elementarne segmenty (fonemy, środkowa i tylna część STS), do nieco bardziej złożonych struktur sylabicznych, a następnie morfemów (tylna część MTG i ITS, środkowego zakrętu skroniowego i dolnej bruzdy skroniowej) stanowiących elementarne jednostki leksykalne. Są one połączone z wieloma obszarami kory i ośrodkami podkorowymi. Wreszcie informacja o znaczeniu słowa analizowana jest w przedniej części lewego płata skroniowego (ATL, *Anterior Temporal Lobe*), który ma dostęp zarówno do leksykalnych jak i rozszerzonych reprezentacji. Lateralizacja mowy dotyczy głownie tego obszaru, oraz obszaru Broca związanego z artykulacją mowy.



RYS. 1. Organizacja przetwarzania informacji fonologiczno-leksykalnej w modelu, który zaproponował Hickok, Poeppel (2007). Objaśnienia skrótów w tekście.

W anomii wzrokowo-przedmiotowej pacjenci mają trudności z prawidłowym nazywaniem oglądanych przedmiotów, ale dotykając ich lub wskazując, do czego mają służyć, nie robią wielu błędów. Caramazza i jego koledzy (2006) zaproponowali do wyjaśnienia rezultatów eksperymentów z takimi pacjentami hipotezę unitarnej organizacji wiedzy koncepcyjnej (*Organized Unitary Content Hypothesis*, OUCH), opartą na preferencjach pomiędzy specyficznymi modalnościami i różnymi typami informacji semantycznej. Hipoteza uprzywilejowanego dostępu może również wyjaśnić przypadki anomii specyficznych kategorii semantycznych, w których np. trudności w nazywaniu dotyczą tylko kategorii zwierząt, ale nie roślin. Caramazza i Mahon (2006) proponują hierarchiczną organizację wiedzy, w której istnieje kilka domen o specyficznej organizacji (ludzie, zwierzęta, rośliny, narzędzia), różniących się rodzajem (również modalnością) informacji na temat obiektów danego typu. Organizacja wiedzy pojęciowej w mózgu powinna być różna w zależności od obiektu, z którym mamy do czynienia. Np. rozpoznawanie twarzy jest specyficzną funkcją, która ma zastosowanie tylko w stosunku do ludzi (lub ogólniej, osobników tego samego gatunku).

Tego rodzaju reprezentacja pojęć umożliwia myślenie symboliczne. Kontekst powoduje pobudzenie rozszerzonych podsieci, które definiują (poprzez relacje z innymi pojęciami) znaczenie słów. Wzajemne hamowanie konkurencyjnych procesów w mózgu zapewnia ograniczenie sensownych odpowiedzi i działań do tych, które są łatwo osiągalne w tak pobudzonej sieci. Podobieństwo fonologiczne i semantyczne pomiędzy słowami może prowadzić do podobnych aktywacji mózgu, ale w większości przypadków torowanie całkowicie ujednoznacznia interpretacje sensu słów — procesy neuronowe typu „zwycięzca bierze wszystko" (O'Reilly, Munakata 2000) powodują, że nic innego „nie przychodzi do głowy".

4. Stany mózgu i ich aproksymacje

Jeśli wyobrażam sobie jakiś konkretny obiekt, np. taki jak „klawiatura", której właśnie używam, w moim mózgu powstaje wyobrażenie składające się z pobudzeń uwzględniających w różnym stopniu wkład ze strony:

- układu wzrokowego: ogólny kształtu obiektu, kształty wyodrębnionych, licznych klawiszy; znaki na tych klawiszach;
- kory czuciowej, ruchowej i przedruchowej: dotyk plastiku i metalu, ruch palców używanych do naciskania klawiszy, układ dłoni na klawiaturze;
- kory skojarzeniowej: związek klawiatura i urządzeń, do których jest podłączona (komputera, programu, ekranu);
- kory słuchowej, obszar Broca: reprezentacji dźwięku, wymowy słów „klawiatura" i „keyboard".

Każde z tych pobudzeń, będące wynikiem rozchodzenia się aktywacji w sieci neuronów, można reprezentować przez zmienny w czasie rozkład prawdopodobieństwa aktywacji określonych grup neuronów. Gdybyśmy znali zbiór grup neuronów i mieli informację o ich aktywacji w różnych warunkach, to pobudzenia te można by opisać za pomocą odpowiednich funkcji bazowych $\psi_i(s, Kont)$, gdzie i jest numerem grupy neuronów (oraz odpowiedniej funkcji bazowej), s jest danym symbolem (pojęciem) a $Kont$ jest kontekstem, w którym ten symbol się pojawia. Całkowita aktywacja mózgu, pojawiająca się w wyniku prezentacji symbolu s w kontekście $Kont$ będzie wówczas aproksymowana za pomocą kombinacji liniowej:

$$\Psi(s, Kont) = \sum_i c_i \psi_i(s_i, Kont)$$

Funkcje bazowe opisują prymitywy percepcyjne, ruchowe, emocjonalne czy abstrakcyjne, coś, co można otrzymać za pomocą neurobrazowania po uśrednieniu aktywacji dla wielu osób by zobaczyć, jakie obszary mózgu (korowe i podkorowe) się aktywizują przy pojawieniu się danego pojęcia. Z formalnego punktu każda grupa neuronów, której aktywność reprezentuje funkcja bazowa, działa jak filtr, wydobywający pewne cechy z pierwotnych perceptów dotyczących pojęcia odpowiadającego symbolowi s. Percepty niekoniecznie muszą się odnosić do danych zmysłowych, mogą to być skojarzenia czysto wewnętrzne, wynikłe z pobudzeń kory związanych ze śladami pamięci i skojarzeń pomiędzy nimi, lub dalece przetworzonych informacji przez wcześniejsze obszary korowe i podkorowe, z którymi dane pole $\phi_i(s_i, Kont)$ jest połączone. Dlatego argumentem tej funkcji bazowej nie jest bezpośrednio s, ale s_i, zbiór zmiennych, które wpływają na stan danego pola. Każde pojęcie ma nieco inny sens w zależności od kontekstu, czyli wcześniejszych aktywacji mózgu (torowania), zmieniając się w różnych skalach czasowych. Widać stąd, że statyczne pojęcia reprezentacji mentalnych lub próby określenia sensu pojęć za pomocą ontologii czy tezaurusów, mogą być jedynie grubym przybliżeniem do rzeczywistego, dynamicznego charakteru stanów mózgu.

Pobudzenia kory zmysłowej interpretowane są w jako własności postrzeganych obiektów a ich pierwotnym odnośnikiem jest świat zewnętrzny. Na poziomie reprezentacji mentalnych powinniśmy operować cechami spostrzeżeń, które możemy zidentyfikować w treści naszych przeżyć. Oznacza to, że zamiast opisywać stan mózgu jako liczne pobudzenia grup neuronów powinniśmy wprowadzić nowe zmienne, które określają intensywność lub rodzaj jakiegoś postrzeżenia, np. barwy dźwięku, koloru, kształtu czy szybkości ruchu. Stan mózgu można przetransformować do tej nowej przestrzeni, którą można nazwać przestrzenią umysłu.

Z matematycznego punktu widzenia relacje między stanami mózgu $\Psi(s, Kont)$ a stanami umysłu $\Phi(s, Kont)$ sprowadzają się więc do transformacji pomiędzy przestrzenią aktywacji grup neuronów opisywaną przez bazę $\psi_i(s_i, Kont)$ a przestrzenią zdarzeń mentalnych, której wymiarami są zmienne skorelowane z własnościami dającymi się wyodrębnić w doświadczeniu wewnętrznym, opisywane przez bazę $\phi_i(s_i, Kont)$, której elementy można powiązać z bazą aktywacji $\psi_i(s_i, Kont)$. Dotyczy to nie tylko percepcji, np. postrzegania kolo-

ru, gdzie rozkład pobudzeń w obszarze V4 kory wzrokowej da się skorelować z postrzeganą barwą, lecz również wszystkich innych stanów mentalnych, łącznie ze stanami emocjonalnymi.

Jedną z możliwości opisu powstawania reprezentacji mentalnych jest aproksymacja za pomocą modeli neuronowych i koneksjonistycznych (O'Reilly i Munakata 2000). Wymaga to utworzenia sieci neuronowej, której węzły niczego bezpośrednio nie reprezentują (można je uznać za nieuświadamialne reprezentacje mikrocech), a dopiero konfiguracje ich pobudzeń można zinterpretować jako cechy, które odpowiadają percepttom odnoszącym się do analizowanych obiektów. Podejście koneksjonistyczne reprezentuje bezpośrednio te konfiguracje pobudzeń jako węzły w sieciach na wyższym stopniu abstrakcji niż sieci neuronowe, ukrywa więc mikrocechy. Rozkład pobudzeń wartości tych cech (pobudzeń węzłów sieci koneksjonistycznej) reprezentuje dany obiekt. W sieciach neuronowych mamy do czynienia z konfiguracjami pobudzeń oscylującymi wokół średnich wartości rozkładów. W takiej rozproszonej reprezentacji automatycznie pobudzają się skojarzone fragmenty reprezentacji dla odpowiednich modalności, jest ona więc skoncentrowana na jakiejś domenie wiedzy koncepcyjnej. Udało się nam (Dobosz i Duch 2009) dokonać wizualizacji trajektorii pokazujących, jak zmienia się z upływem aktywność wszystkich użytych w symulacji neuronów. Każdy punkt na rysunku poniżej (por. rys. 2) odpowiada określonej wartości średniej aktywności 140 neuronów w danej chwili czasu. Zagęszczenia trajektorii reprezentujące atraktory neurodynamiki pokazują, że system fluktuuje wokół danego pojęcia przez jakiś czas, a następnie przechodzi do pojęć skojarzonych, nie powraca jednak do tego samego miejsca, bo historia jego ewolucji zmienia sytuację. Nawet taki obraz jest wielkim uproszczeniem, gdyż zmiany ogólnego pobudzenia (efekty związane z emocjami, uwagą, zmęczeniem) może w znacznym stopniu zmienić ten krajobraz. Reprezentacja mentalna jest tu atraktorem neurodynamiki, chwilowym spowolnieniem zmian aktywacji prowadzącym do kolejnych reprezentacji w procesie, który nazywamy skojarzeniowym. W zależności od stanu mózgu i historii ewolucji tego stanu krajobraz atraktorów zmienia się drastycznie, umożliwiając całkiem odmienne skojarzenia.

Funkcje bazowe, które opisują cechy (w tym również *qualia*) odnoszące się do obiektów mogą być kombinacją pobudzeń wielu neuronów, które przez odległe projekcje przenoszą informacje o konfiguracjach pobudzeń kory wzrokowej czy czuciowej do innych obszarów mózgu. Pojęcia abstrakcyjne mają wyraźnie uboższe reprezentacje, gdyż nie pobudzają kory zmysłowej ani kory ruchowej.

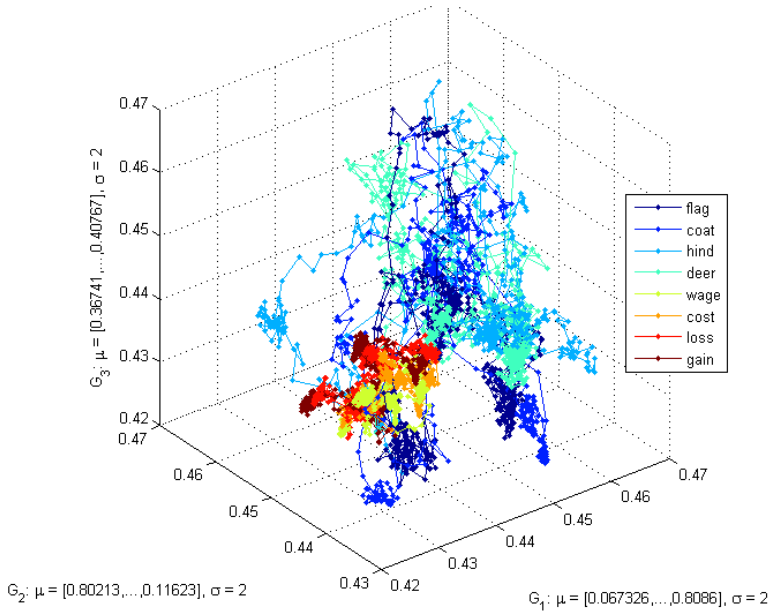
Prawdopodobieństwo pojawienia się skojarzenia pomiędzy dwoma stanami mentalnymi obliczyć można jako prawdopodobieństwo przejścia pomiędzy stanem początkowym $\Psi(s, Kont)$ a stanem końcowym

$\Psi(s', Kont')$ pobudzenia obszarów mózgu. Te prawdopodobieństwa powinny dać się przybliżyć jako iloczyny skalarne:

$$< \Psi(s, Kont)|\Psi(s', Kont') >$$

w odpowiedniej metryce. Kontekst jest tu bardzo skomplikowany, gdyż powinien uwzględnić nie tylko bieżącą sytuację, ale również skupianie uwagi, zmieniające pobudzenia i powodujące, że mamy różne skojarzenia, zależnie od tego, czy zwracamy uwagę na nazwę (fonologię), czy formę wizualną, funkcję, czy też wywołane przez dany obiekt emocje. Podobieństwa pomiędzy stanami umysłu są podstawą do tworzenia kategorii naturalnych. Skupienie różnych stanów mentalnych, silnie ze sobą skojarzonych (a więc o dużych prawdopodobieństwach przejść pomiędzy nimi), tworzy taką rozmytą kategorię i wszystko co jest z nią skojarzone – pod względem wyglądu, albo funkcji (czyli możliwości działania, ruchu) – jest kategoryzowane w podobny sposób. Nie ma zbioru definiujących cech dla pojęcia „krzesło", ale różnorakie skojarzenia pozwalają nam rozpoznać dany obiekt jako krzesło. Pobudzenia neuronów odpowiedzialnych za semantykę kategorii naturalnych nie tworzą zbiorów wypukłych lecz mogą się składać z wielu rozłącznych reprezentacji (atraktorów), które należą do tej samej kategorii (mają wspólne reprezentacje fonologiczne).

Rozwinięcia na różne funkcje bazowe stwarzają możliwości różnego opisu stanów mózgu. Może to być zapis wektorowy uśrednionych wartości współczynników rozkładu na funkcje bazowe $\phi_i(s_i, Kont)$ dla danego pojęcia s. Stan mentalny $\Phi(s)$ reprezentujący to pojęcie jest wówczas reprezentowany przez wektor, określający jakie cechy przypisać można danemu pojęciu. Taka reprezentacja stosowana jest często w analizie języka naturalnego (Manning i Schütze 1999). Brakuje w tej reprezentacji części związanej z fonetyką słowa, co utrudnia symulacje kreatywności na poziomie tworzenia słów i rozumienie skojarzeń na poziomie fonologicznym. W typowych zastosowaniach nie jest to konieczne i można takie reprezentacje stosunkowo łatwo rozszerzyć dodając część związaną z fonologią. Współczynniki wektorowe można interpretować jako prawdopodobieństwa rozkładu pobudzeń różnych obszarów mózgu, co nadaje pojęciu pewną strukturę, można np. skupić się na podobszarach ruchowych, w którym jest informacja o tym, co z danym przedmiotem możemy zrobić, jakie prymitywy ruchowe (reprezentowane za pomocą funkcji bazowych) składają się na takie działanie. Podobieństwa pomiędzy pojęciami są tu więc wynikiem tylko strukturalnych relacji a pojęcie jest opisane przez mikrocechy, które odnoszą się do jego



rys. 2.

percepcji, ogólnej wiedzy o nim, jak i możliwych sposobów interakcji z danym obiektem. W praktyce w reprezentacjach wektorowych brakuje odnośników do percepcji kształtu i innych własności przestrzennych.

Jeszcze częściej stosowanym przybliżeniem wektorowym jest próba oceny korelacji statystycznych pomiędzy pobudzeniami poszczególnych funkcji bazowych dla różnych pojęć. W praktyce ocenia się to na podstawie kookurencji różnych pojęć tekstach (Manning i Schütze 1999). Do tego przybliżenia można również dojść badając transformacje pomiędzy funkcjami bazowymi reprezentującymi pobudzenia grup neuronów. Tego typu rozważań w analizie języka naturalnego dotychczas nie robiono, wyprowadzając z rozważań statystycznych reprezentacje wektorowe, które mają za zadanie uchwycić kontekst na podstawie kookurencji wyrazów. W efekcie takie reprezentacje dają bardzo prymitywne przybliżenie do użytecznych reprezentacji mentalnych, pomijające istotne własności strukturalne obiektów, jak i powiązania pomiędzy percepcją i działaniem. Podobnie drastycznym uproszczeniem są sieci semantyczne (Sowa 1991).

Wydaje się, że systematyczna analiza przybliżenia do dynamicznych procesów rozchodzenia się aktywności neuronalnej w mózgu i transfor-

macja tych stanów do przestrzeni, w której można zdefiniować stany mentalne jest dobrą drogą do lepszego opisu reprezentacji mentalnych i reprezentacji pojęć. Można w tym celu zastosować abstrakcyjną teorię reprezentacji, rozwiniętą dla potrzeb mechaniki kwantowej. Wektor stanu (funkcja falowa) może w niej zostać poddana dowolnej transformacji obrotów, ale nie zmienia to relacji pomiędzy różnymi wektorami stanu.

5. Zastosowania

Przedstawione powyżej rozważania oparte są na przekonaniu, że jedynie przez aproksymację fizycznych stanów mózgu, zawierających znacznie więcej informacji niż stany mentalne, możemy dokonać istotnego postępu w rozumieniu i opisie reprezentacji mentalnych, przydatnych nie tylko filozofom, ale też mających zastosowania w analizie języka naturalnego (Duch, Matykiewicz, Pestian 2008), kategoryzacji pojęć w psychologii (Duch 1997) i architekturach kognitywnych w sztucznej inteligencji (Duch 2010). Informatyka neurokognitywna (Duch 2009) usiłuje tworzyć praktyczne algorytmy czerpiąc inspiracje ze zrozumienia procesów zachodzących w mózgu. Jest już szereg przykładów, pokazujących użyteczność takich metod. Ucieleśnienie jak i enaktywizm są dla rozwoju reprezentacji pojęć bardzo istotne (Barsalou 2008), podkreślają konieczność wielomodalnych reprezentacji, ale to jeszcze nie wystarcza. Proces tworzenia się abstrakcyjnych reprezentacji w oparciu o reprezentacje percepcyjno-ruchowe stwarza niełatwe problemy (Mahon i Caramazza 2008), ale z punktu widzenia transformacji pomiędzy stanami mózgu i umysłu nie jest trudny do wyobrażenia.

Transformacje pomiędzy stanami mózgu a stanami umysłu stosuje się w różnego rodzaju detektorach kłamstwa wykorzystujących pomiary elektrofizjologiczne (głównie EEG). Podejście oparte na transformacjach jest też podstawą do tworzenia interfejsów mózg-maszyna (*Brain-Computer Interfaces*, BCI), do sterowania urządzeniami za pomocą myśli, a raczej intencji. Pojawiają się pierwsze gry komputerowe, wykorzystujące taki interfejsy, są też liczne zastosowania medyczne (Coyle i in. 2003), w tym urządzenia typu *neurofeedback*, przydatne w terapii licznych problemów psychicznych. W przypadku interfejsów mózg-komputer intencje ruchu, które należą do świata umysłu, zmieniają stany mózgu w sposób możliwy do odczytania za pomocą elektrod aparatury elektroencefalograficznej. Chociaż nie jest to precyzyjna informacja o stanie mózgu, aktywność każdej elektrody EEG reprezentuje uśrednienie aktywności wielu milionów neuronów, wystarcza to jednak do tego, aby dokonać transformacji informacji o aktywności mózgu na

informację pozwalającą rozróżnić kilka stanów mentalnych. W eksperymentach wykorzystujących fMRI (Hayens i in. 2007) udało się odróżnić stany umysłu związane z intencjami dodawania bądź odejmowania dwóch liczb od siebie. W pracy Mitchella i wsp. (2008) stworzono sieć neuronową przewidującą rozkład pobudzeń mózgu mierzonych w eksperymentach z fMRI przy prezentacji różnych rzeczowników. Model ten, nauczony przy wykorzystaniu reprezentacji wektorowej słów zawartych w dużym korpusie językowym, oraz na kilkudziesięciu wynikach rzeczywistych pomiarów fMRI, potrafi odróżnić od siebie stany mózgu spodziewane w eksperymentach z innymi rzeczownikami. W tym przypadku jako funkcje bazowe używa się aktywności pojedynczych wokseli w fMRI.

Odczytywanie stanów mentalnych na podstawie aktywności mózgu jest dużym problemem technicznym, ale już dzisiaj dyskutowane są problemy etyczne związane z możliwością podglądania prywatnych stanów mentalnych (Haynes i Rees 2005, 2006). Postęp w tym kierunku nie będzie jednak łatwy. Z jednej strony nie mamy dobrych technik eksperymentalnych by podejrzeć niewielkie grupy neuronów i stworzyć odpowiednią bazę funkcji do opisu stanów mózgu. Żadna z obecnie rozwijanych technik nie ma odpowiedniej rozdzielczości czasowej i przestrzennej by na to pozwolić. Analiza sygnałów EEG jest bardzo trudna i nie umiemy jeszcze znaleźć w nich precyzyjnej informacji o cechach stanów mentalnych, nie wiemy nawet, czy taka informacja jest w EEG. Z drugiej strony nie mamy też dobrej neurofenomenologii pozwalającej na precyzyjny opis stanów mentalnych i powiązanie ich ze stanami mózgu. Próba opisu doświadczenia wewnętrznego przedstawiona przez Hurlburta i Schwitzgebela (2007) nie jest wystarczająca by określić, jakiego rodzaju transformacji powinniśmy szukać dla opisu stanów mentalnych. Jedynie w prostych sytuacjach eksperymentalnych możemy ustalić, jakie zmienne odnoszące się do stanów mentalnych potrzebne są do dokonania wymaganych przez eksperyment rozróżnień.

Prezentowane tu podejście oferuje dość prosty język, pozwalający zrozumieć zaskakujące, często nieracjonalne decyzje podejmowane przez ludzi. Psychologia oferuje pewne racjonalizacje, ale są one arbitrarne i nie mają mocy wyjaśniającej (Duch 1997). Logiczne argumenty okazują się często błędne tam, gdzie dobrze się sprawdzają argumenty oparte na intuicji (Gigerenzer 2009). Wybór intuicyjny nie buduje konstrukcji logicznej w oparciu o obecność lub rbak pewnych cech, ale opiera się na doświadczeniu, podobieństwie do wcześniej spotkanych sytuacji, przyciągającym trajektorię aktywności neuronów do jakiegoś atraktora. Proces ten może nie dać się opisać za pomocą reguł logicznych. Eksploracja labiryntu, w którego lewej odnodze jest kęs jedzenia w 80% przypadków a w prawej tylko w 20% pokazuje, że szczur w 80% razy

idzie w lewo a w 20% idzie w prawo, chociaż gdyby zawsze przechodził do lewej odnogi znalazłby jedzenie częściej. W modelach pamięci można zobaczyć, że obszary atrakcji, które się tworzą dla decyzji „iść w lewo czy w prawo" są odbiciem statystyki częstości znajdowania pożywienia. Na poziomie interpretacji psychologicznej możemy powiedzieć, że szczur, który ma zwykle dużą konkurencję, nie zapominając o innych możliwościach czasami unika tłumów i zdobywa dodatkowe pożywienie. Na poziomie neurodynamiki widzimy trajektorie aktywności, które w 80% przypadków kończą się w atraktorze „iść w lewo", a w pozostałych „iść w prawo".

Decyzje ludzi oparte są często na takim samym mechanizmie, a racjonalne wyjaśnienia, jakie do nich dodajemy są konfabulacjami. Reklamy często powtarzane pozostawiają głębokie ślady w pamięci, wpływając na decyzje zakupów. Początkowa sugestia, chociaż pozornie całkiem nie związana z decyzją, którą trzeba podjąć, ma na nią silny wpływ. Dan Ariely i George Loewenstein pokazali (Ariely 2008), że zapisanie dwóch ostatnich cyfr numerów ubezpiecznia (*social security*) przez studentów miało duży wpływ na deklarowane sumy podczas aukcji przedmiotów o cenach do 100$. Osoby, które zapisywały większe liczby skłonne są zapłacić więcej. Nawet taka nieświadoma i całkiem nie związana z późniejszym działaniem sugestia wpływa na podejmowane decyzję. Sugestie związane z zadaniami wyboru wpływają na nią jeszcze silniej. Dodanie okrojonej oferty tego samego produktu znacząco wpływa na podwyższenie oceny produktu, który chcemy wypromować. Jeśli rozważamy trzy możliwości, A, B i B', przy czym B' jest podobne do B ale gorsze pod jakimś istotnym względem, atraktory dla B i B' mają większą siłę przyciągania i pozostaje dość oczywisty wybór pomiędzy B i B'. Np. jeśli mamy do wyboru subskrypcję w sieci za 50$ lub wersję drukowaną za 125$ to niewiele osób wybiera wersję drukowaną, ale jeśli oferowana jest dodatkowo wersja „druk i wersja sieciowa" za tę samą cenę co druk to proporcje się odwracają. Jak pokazał Dan Ariely (2008) dotyczy to dowolnych ofert, łącznie z partnerem życiowym: jeśli jest trzech potencjalnych partnerów, z których A uznawany był wcześniej za bardziej atrakcyjnego, ale B ma kolegę podobnego do siebie, ale nieco mniej atrakcyjnego, preferencje przesuwają się w stronę B.

Bardziej skomplikowane sytuacje związane są z konkluzjami wymagającymi logicznego myślenia. Najprostsze są bezpośrednie skojarzenia, $A \Rightarrow B \Rightarrow C$, tworząc uporządkowany ciąg atraktorów następujących po sobie. Rozważmy jednak takie 3 zdania:

- Wszyscy członkowie gabinetu to złodzieje.
- Żaden muzyk nie jest członkiem gabinetu.

- Co można powiedzieć o relacji miedzy muzykami i złodziejami?

Nieco prostsza wersja, gdyż bliska znanemu schematowi myślenia „nie każdy uczony jest mędrcem", to:

- Wszyscy akademicy to uczeni.
- Żaden mędrzec nie jest akademikiem.
- Co można powiedzieć o relacji między uczonymi i mędrcami?

Po paru tygodniach bezowocnych zmagań podałem studentom prawidłową odpowiedź na pierwszą z tych wersji, jednak na egzaminie dla drugiej wersji podali kilkanaście błędnych lub niejednoznacznych odpowiedzi (tylko dwie ostatnie są poprawne):

1. Nie ma relacji pomiędzy tymi trzema zbiorami.
2. Bycie mędrcem świadczy o tym, że nie jest się uczonym.
3. Niektórzy mędrcy to uczeni.
4. Wszyscy uczeni nie będący akademikami są mędrcami.
5. Wszyscy uczeni to mędrcy.
6. Żaden mędrzec nie jest uczonym.
7. Nie wszyscy mędrcy to uczeni.
8. Istnieje taki mędrzec, który może być uczonym.
9. Mogą istnieć mędrcy, którzy są uczonymi.
10. Nie trzeba być mędrcem aby być uczonym.
11. Mędrzec uczonemu nie równy.
12. Nie wszyscy uczeni to mędrcy.
13. Nie każdy uczony jest mędrcem.
14. Istnieją tacy uczeni, którzy nie są mędrcami.
15. Istnieje uczony, który nie jest mędrcem.

6. Dyskusja

Rozważania na temat reprezentacji mentalnych i ich związek ze stanami mózgu można rozszerzyć na zagadnienia dotyczące kreatywności (Duch i Pilichowski 2007; Duch 2007), roli prawej półkuli mózgu w doświadczeniach wglądu (Duch 2007a) jak i praktycznych algorytmów analizy tekstu, pozwalających na tworzenie rozszerzonych reprezentacji (Duch i in. 2008, Duch 2009). Zbudowanie modelu umysłu uwzględniające perspektywę wewnętrzną jest nadal wielkim wyzwaniem. Idee dotyczące geometrycznego opisu stanów mentalnych (Duch 1997, 2001, 2002, 2002a) można połączyć z opisanym tutaj podejściem transformacyjnym do relacji mózg-umysł. Droga do stworzenia dokładnego modelu

takich relacji jest nadal daleka. Nie znamy szczegółów procesów zachodzących w mózgu, są trudności związane z badaniami eksperymentalnymi, brak jest dobrych metod matematycznych do analizy sygnałów i procesów rozchodzenia się aktywacji w rzeczywistych sieciach neuronowych, procesów przetwarzania informacji w oparciu o takie sieci. Jednakże nawet proste mózgo-podobne przetwarzanie informacji daje rezultaty, które ma cechy jakościowe porównywalne do obserwowanych funkcji; złożoność mózgu nie jest więc głównym problemem stojącym przed budową sztucznych umysłów (Duch 2005; Duch, Oentaryo, Pasquier 2008)!

Metody komputerowe dopiero od niedawna zaczęto stosować do modelowania syndromów neuropsychologicznych i chorób psychicznych (Parks i in. 1998). Na razie objęto nimi jedynie część zagadnień, które można w ten sposób badać. Nie brakuje problemów fundamentalnych, do których nie bardzo wiadomo, jak podejść. Należą do nich urazy psychogenne, zaburzenia osobowości i inne problemy wymagające pełnego modelu umysłu. Jednakże za pomocą modeli komputerowych udało się już teraz osiągnąć więcej, niż można było oczekiwać zdając sobie sprawę ze stopnia komplikacji takiego modelowania. Niewielka liczba założeń i stosunkowo proste sieci neuronowe pozwalają na zrozumienie zjawisk zachodzących przy uszkodzeniach i rehabilitacji mózgu.

Często jest to rozumienie jakościowe, metaforyczne, posługujące się luźnymi analogiami. Należy jednak przyznać, że dokonany został duży postęp, pojawił się nowy styl rozumowania ze specyficznymi problemami i pytaniami, język opisu nieredukowalny do języka używanego dotychczas w psychiatrii czy w psychofarmakologii (Duch 2007b). Dzięki symulacjom komputerowym można się spodziewać podobnego postępu w rozumieniu reprezentacji mentalnych i wyższych czynności poznawczych, w szczególności procesów podejmowania decyzji.

Podziękowania: Krzysztof Dobosz opracował program do wizualizacji atraktorów i przygotował rysunek ze strony 19.

Literatura

Austin, J. H. (1988). *Zen and the Brain: Toward an Understanding of Meditation and Consciousness*. Cambridge, MA: MIT Press.

Austin, J. H. (2006). *Zen-Brain Reflections: Reviewing Recent Development in Meditation and States of Consciousness*. Cambridge, MA: MIT Press.

Ariely, D. (2008). *Predictably Irrational. The Hidden Forces That Shape Our Decisions*. Harper-Collins.

Barsalou, L. W. (2008). Grounded Cognition. *Annual Reviews of Psychology* 59, s. 617–645.

Bowden, E. M., Jung-Beeman, M., Fleck, J. & Kounios, J. (2005). New approaches to demystifying insight. *Trends in Cognitive Science 9*, 322–328.

Brooks, R. (1986). Elephants don't play chess. *Robotics and Autonomous Systems* 6, s. 3–15.

Brooks, R, Stein, L. A. (1994). Building Brains for Bodies. *Autonomous Robotics* 1, s.7–25.

Caramazza, A., & Mahon, B. Z. (2006). The organization of conceptual knowledge in the brain: the future's past and some future directions. *Cognitive Neuropsychology* 23, 13–38.

Coyle, S, Ward, T., Markham, C. (2003). Brain-computer interfaces: A review. *Interdisciplinary Science Reviews* 28(2), s. 112–118.

Damasio, H., Grabowski, T. J., Tranel, D., Frnak, R. J, Hichiwa, R. D, Damasio, A. R. (1996). *A neural basis for lexical retrieval*. Nature 380, s. 499–505.

Dehaene, S., Cohen, L., Sigman, M. & Vinckier, F. (2005). The neural code for written words: a proposal. *Trends in Cognitive Science 9*, s. 335–341.

Dehaene, S., & Naccache, L. (2001). Towards a cognitive neuroscience of consciousness: Basic evidence and a workspace framework. *Cognition 79*, 1–37.

De Marco, R. J., Menzel, R. (2008). Learning and memory in communication and navigation in insects. W: Menzel R, (red.) *Learning Theory and Behavior*, Vol. 1 of *Learning and Memory: A Comprehensive Reference* (4 vols). Oxford: Elsevier, s. 477–98.

Dobosz, K., Duch, W. (2009). Fuzzy Symbolic Dynamics for Neurodynamical Systems. *Neural Networks*, `http://dx.doi.org/10.1016/j.neunet.2009.12.005`

Duch, W. (1997). Platonic model of mind as an approximation to neurodynamics. W: S. Amari, N. Kasabov (red.) *Brain-like computing and intelligent information systems*. Springer, Singapore, rozdz. 20, s. 491–512

Duch, W. (2001). Neurokognitywna teoria świadomości. *Kognitywistyka i Media w Edukacji* 5(2), s. 47–67.

Duch, W. (2002). Geometryczny model umysłu. *Kognitywistyka i Media w Edukacji* 6, s. 199–230.

Duch, W. (2002a). Fizyka umysłu. *Postępy Fizyki* 53D, s. 92–103.

Duch, W. (2005). Brain-inspired conscious computing architecture. *Journal of Mind and Behavior* 26(1–2), s. 1–22.

Duch, W. (2006). Madhyamika, nauka i natura rzeczywistości. Uwagi na marginesie książki: Matthieu Ricard i Trinh Xuan Thuan, *Nieskończoność w Jednej Dłoni: Od Wielkiego Wybuchu do Oświecenia. Kognitywistyka i Media w Edukacji* 1–2, s. 293–316.

Duch, W. (2007). Intuition, Insight, Imagination and Creativity. *IEEE Computational Intelligence Magazine* 2(3), s. 40–52.

Duch, W. (2007a) Creativity and the Brain. In: Ai-Girl Tan (red.)*A Handbook of Creativity for Teachers*. Singapore: World Scientific Publishing, s. 507–530.

Duch, W. (2007b). Computational Models of Dementia and Neurological Problems. W: *Neuroinformatics*, C.J. Crasto (red.) *Methods in Molecular Biology* series (J. Walker, series ed.). Totowa, NJ: Humana Press, rozdz. 17, s. 307–336.

Duch, W. (2009). Neurocognitive Informatics Manifesto. W: *Series of Information and Management Sciences*, California Polytechnic State University, s. 264–282.

Duch, W. (2010). Architektury kognitywne. W: R. Tadeusiewicz (red.) *Neurocybernetyka teoretyczna*. Wyd. Uniwersytetu Warszawskiego.

Duch, W., Pilichowski, M. (2007). Experiments with computational creativity. *Neural Information Processing — Letters and Reviews 11*, s. 123–133.

Duch, W., Matykiewicz, P., Pestian, J. (2008). Neurolinguistic Approach to Natural Language Processing with Applications to Medical Text Analysis. *Neural Networks* 21(10), s. 1500–1510.

Durham, W.H. (1983). Testing the malaria hypothesis in West Africa. W: *Distribution and Evolution of Hemoglobin and Globin Loci*. New York: Elsevier Science Publishing Co., Inc.

Frith, C., Frith, U. (2005). Theory of mind. *Current Biology* 15, R644–R646.

Gaillard, R., Naccache, L., Pinel, P., Clémenceau, S., Volle, E., Hasboun, D., Dupont, S., Baulac, M., Dehaene, S., Adam, C., & Cohen, L. (2006). Direct intracranial, FMRI, and lesion evidence for the causal role of left inferotemporal cortex in reading. *Neuron 50*, s. 19–204.

Gigerenzer, G. (2009). Intuicja. Inteligencja nieświadomości. Warszawa: Prószyński i S-ka.

Gruszka, A., & Nęcka, E. (2002). Priming and acceptance of close and remote associations by creative and less creative people. *Creativity Research Journal* 14, s. 193–205.

Haynes, J.D. & Rees, G. (2005). Predicting the stream of consciousness from activity in early visual cortex. *Current Biology* 15, s. 1301–1307.

Haynes, J.-D. & Rees, G. (2006). Decoding mental states from brain activity in humans. *Nature Reviews Neuroscience* 7, s. 523–534.

Haynes, J.-D., Sakai, K., Rees, G., Gilbert, S., Frith, C. & Passingham, D. (2007). Reading hidden intentions in the human brain.*Current Biology* 17, s. 323–328.

Hickok, G. & Poeppel, D. (2007). The cortical organization of speech processing. *Nature Reviews Neuroscience* 8, s. 393–402.

Hurlburt, R.T., Schwitzgebel, E. (2007). *Describing Inner Experience? Proponent Meets Skeptic*. Cambridge, MA: MIT Press.

Itert, L. Duch, W. & Pestian, J. (2007). Influence of *a priori* Knowledge on Medical Document Categorization, *IEEE Symposium on Computational Intelligence in Data Mining*. IEEE Press, s. 163–170.

Jung-Beeman, M., Bowden, E. M., Haberman, J., Frymiare, J. L., Arambel-Liu, S., Greenblatt, R., Reber, P. J., &. Kounios, J. (2004). Neural activity when people solve verbal problems with insight. *PLoS Biology 2*, s. 500–510.

Lehmann, F. (red.), (1992). *Semantic Networks in Artificial Intelligence*. Oxford: Pergamon.

Lin, L., Osan, R., & Tsien, J. Z. (2006). Organizing principles of real-time memory encoding: neural clique assemblies and universal neural codes. *Trends in Neuroscience* 29(1), s. 48–57.

Kapleau, P. (1992). *Zen, Świt na Zachodzie*. Warszawa: ZBZ „Bodhidarma".

MacLean, P. (1990). *The Triune Brain in Evolution*. New York: Plenum Press.

Mahon, B.Z, Caramazza, A. (2008). A Critical Look at the Embodied Cognition Hypothesis and a New Proposal for Grounding Conceptual Content. *Journal of Physiology — Paris* 102, s. 59–70.

Manning, C.D. & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.

Matykiewicz, P., Duch, W., & Pestian, J. (2006). Nonambiguous Concept Mapping in Medical Domain, *Lecture Notes in Artificial Intelligence 4029*, s. 941–950.

Mednick, S.A. (1962). The associative basis of the creative process. *Psychological Review* 69, s. 220–232.

Meller, S. (2008). *Świat według Mellera. Życie i polityka: ku przyszłości*. Warszawa: Rosner i wspólnicy.

Mitchell, T. M., Shinkareva S. V., Carlson A., Chang, K. M., Malave, V. L., Mason, R. A., Just, M. A. (2008). Predicting human brain activity associated with the meanings of nouns. *Science* 30, 320(5880) s. 1191–95.

Northoff, G., Heinzel, A., de Greck, M., Bermpohl, F., Dobrowolny, H., Panksepp, J. (2006). Self-referential processing in our brain — a meta-analysis of imaging studies on the self. *Neuroimage* 15, 31(1), s. 440–457.

Northoff, G., Panksepp, J. (2008). The trans-species concept of self and the subcortical-cortical midline system. *Trends in Cognitive Science* 12(7), s. 259–64.

Okada, K., Hickok, G. (2006). Identification of lexical-phonological networks in the superior temporal sulcus using fMRI. *Neuroreport* 17, s. 1293–1296.

O'Reilly, R.C., Munakata, Y. (2000). Computational Explorations in Cognitive Neuroscience Understanding the Mind by Simulating the Brain. Cambridge, MA: MIT Press.

Parks, R. W., Levine, D. S., Long, D., (red.) (1998). *Fundamentals of Neural Network Modeling*. Cambridge, MA: MIT Press.

Popper, K., Eccles, J.C. (1999). *Mózg i jaźń*, Tom 1–3, Poznań: Wyd. Protext.

Price, C. J., Devlin, J. T. (2003). The myth of the visual word form area. *NeuroImage* 19, s. 473–481.

Pulvermuller, F. (2003). *The Neuroscience of Language. On Brain Circuits of Words and Serial Order*. Cambridge, UK: Cambridge University Press.

Radhakrishnan, S. (1958). *Filozofia indyjska*, tłum. Z. Wrzeszcz, t. 1–2, Warszawa: PAX.

Sowa, J. F. (red.), (1991). *Principles of Semantic Networks: Explorations in the Representation of Knowledge*. San Mateo, CA: Morgan Kaufmann Publishers.