Marcin Miłkowski

Obliczeniowe teorie świadomości

Analiza i Egzystencja 11, 133-154

2010

„Analiza i Egzystencja" 11 (2010)
ISSN 1734-9923

ARTYKUŁY

MARCIN MIŁKOWSKI*

OBLICZENIOWE TEORIE ŚWIADOMOŚCI

Słowa kluczowe: komputacjonizm, świadomość, trudny problem, funkcjonalizm
Keywords: computationalism, consciousness, hard problem, functionalism

Można niekiedy odnieść wrażenie, że czasy obliczeniowych teorii umysłu już minęły i pora poszukać rozwiązań bardziej radykalnych. Współcześni filozofowie lubują się wręcz w kontrowersyjnych propozycjach ontologicznych. Jedni ożywiają panpsychizm, do niedawna uważany za stanowisko kompletnie martwe. Inni przywołują popularny na przełomie XIX i XX wieku monizm neutralny, do niedawna prawie całkowicie zapomniany. Inni znowu głoszą wszechobejmujący sceptycyzm co do możliwości naukowego wyjaśnienia świadomości.

Uważam, że taki pesymizm jest dalece przedwczesny. Nie mamy – to prawda – żadnej kompletnej teorii świadomości, a więc nie istnieje też żadna pełna obliczeniowa teoria świadomości. Istnieją wyłącznie prototeorie. Mimo to można już dziś się zastanawiać, jakie są szanse powodzenia

* Marcin Miłkowski jest adiunktem w Zakładzie Logiki i Kognitywistyki Instytutu Filozofii i Socjologii PAN. Zajmuje się filozofią kognitywistyki i filozofią umysłu. Autor ponad 40 artykułów naukowych opublikowanych w językach polskim, angielskim i niemieckim. W roku 2008, wraz z Robertem Poczobutem, zredagował pracę zbiorową *Analityczna metafizyka umysłu. Najnowsze kontrowersje*. Obecnie pracuje nad książką na temat natury wyjaśniania obliczeniowego w kognitywistyce.

projektu obliczeniowego wyjaśnienia świadomości. Co więcej, sukcesy metod obliczeniowych w kognitywistyce: w teoriach postrzegania, funkcjonowania pamięci czy wnioskowania, w analizach kodu genetycznego w biologii, a także w komputerowych symulacjach zjawisk przyrodniczych mogą przynajmniej podważać zdecydowany sceptycyzm co do możliwości zastosowania tych metod również w badaniach nad świadomością.

Niżej będę argumentował, że świadomość nie daje się w pełni wyjaśnić w sposób obliczeniowy, a więc obliczeniowe wyjaśnienie nie jest wyjaśnieniem wystarczającym do zrozumienia wszystkich funkcji świadomości, jednak jej informacyjna natura sprawia, że jest ono konieczne do wyjaśnienia jednej z jej funkcji. Nie zakładam, że wyjaśnienie funkcji jakiegoś układu jest jednoznaczne z wyjaśnieniem wszystkiego, co jest interesujące w tym układzie, gdyż założenie to jest podważane przez przeciwników komputacjonizmu. Chcę pokazać, że nawet przy uchyleniu tego założenia komputacjonizm jest dzisiaj w teoriach świadomości bezkonkurencyjny.

W pierwszej części tekstu przypomnę ogólne argumenty na rzecz obliczeniowego wyjaśniania świadomości, a także bardzo skrótowo przedstawię zarysy najpopularniejszych dziś prototeorii, do których można zaliczyć teorie globalnej przestrzeni roboczej, teorię modelu świata, teorie myśli drugiego rzędu oraz teorie informacyjno-integracyjne. Następnie będę wykazywać, że teorie konkurencyjne albo są mylnie uważane za nieobliczeniowe, albo są nazbyt ogólnikowe, aby mogły cokolwiek wyjaśniać. Wobec tego nasuwa się wniosek, że nadal trzeba stawiać na komputacjonizm w odniesieniu do świadomości, bo nie ma on żadnych poważnych konkurentów.

Pojęcie świadomości jest trudne do zdefiniowania. Jest tak z kilku względów: po pierwsze, predykat „świadomy" odnosi się do bardzo wielu zjawisk w języku potocznym, także w sposób metaforyczny (np. „Unia Europejska jest świadoma trudności polskiego przemysłu zapałczanego"); po drugie, nie istnieje teoria naukowa, która jednoznacznie wskazywałaby naturę tego zjawiska; a po trzecie – jako zjawisko naturalne, z trudem poddaje się wyczerpaniu w definicji. Jak zauważono w odniesieniu do nazw rodzajów naturalnych, można je traktować niemal jako nazwy własne: zwolennicy takiego ujęcia głoszą, że rodzaje naturalne nie mają definicyjnych własności, na podstawie których się je identyfikuje, lecz ich nazwy są jedynie etykietami nadanymi pewnego rodzaju przedmiotom w świecie (przyczynowa teoria nazw), pozbawionymi konotacji. To więc, że woda składa się z cząstek H_2O nie jest jej cechą definicyjną, czyli faktem językowym dotyczącym wyrazu

„woda", lecz faktem empirycznym, odkrywanym w badaniach nad tym przedmiotem, który zwaliśmy „wodą". Nie rozstrzygając, czy taka skrajnie Millowska teoria nazw rodzajów naturalnych jest słuszna, warto zauważyć, że w odniesieniu do świadomości wydaje się ona bardzo pociągająca – nie znamy natury zjawiska, jakim jest świadomość, więc nie jesteśmy jej też w stanie porządnie zdefiniować. Po prostu teoretyczne ujęcie natury świadomości nie jest możliwe, zanim nie powstanie teoria; definicja terminu teoretycznego nie może być właściwie oceniana, zanim nie powstanie choćby istotny zaczątek teorii.

Nie znaczy to, że jesteśmy skazani wyłącznie na definicje ostensywne, choć można wątpić, czy da się kiedykolwiek zamknąć świadomość w zwartej formule definicyjnej *per genus proximum et differentiam specificam*. Zwykle rozróżnia się kilka pojęć świadomości: pojęcie świadomości jako *cechy podmiotu* (ang. *creature consciousness*) i jako *cechy stanu umysłowego* (ang. *state consciousness)*. Można powiedzieć o mnie, że jestem świadomy, co piszę (czyli mam świadomość jako podmiot); ale także można powiedzieć, że moja myśl na temat pisanego właśnie zdania jest myślą świadomą. Nie znaczy to, że myśl sama jest podmiotem psychicznym, lecz że jest uświadamiana, czyli jest przedmiotem świadomości. Mimo że obie cechy można rozróżnić, to jest też oczywiste, że nie sposób wyjaśnić, w jaki sposób jakaś istota jest podmiotem świadomym, nie wyjaśniając jednocześnie, w jaki sposób uświadamia sobie ona swoje myśli (pomijam behawioryzm jako ewentualną próbę wyjaśniania jedynie świadomości podmiotowej z pominięciem mechanizmów uświadamiania przedmiotu).

W encyklopedycznych ujęciach (por. Van Gulick 2009) wśród podstawowych wyznaczników świadomości podmiotowej wskazuje się: (1) odbieranie wrażeń zmysłowych; (2) bycie na jawie; (3) bycie świadomym samego siebie; (4) posiadanie stanów mających określone jakości; (5) bycie podmiotem stanów świadomych. Jak widać, wyznaczniki (4) i (5) odnoszą się do określonych stanów świadomości – przynajmniej niektórym z nich przypisuje się swoisty charakter jakościowy. Aby przybliżyć formę tej wyliczanki do definicji równoważnościowej, można powiedzieć, że podmiot jest świadomy zawsze i tylko wtedy, gdy odbiera wrażenia zmysłowe, czyli posiada odpowiednie narządy sensoryczne; nie znajduje się w fazie snu, posiada stany świadome, a także przynajmniej może bywać świadomy tego, że jest świadomy. Zwolennicy teorii świadomości drugiego rzędu ostatni

warunek sformułują z użyciem dużego kwantyfikatora i usuną operator możliwości.

Do najczęściej wskazywanych cech stanów świadomych zalicza się to, że: (1) są one uświadamiane bądź dostępne podmiotowi; (2) jedne z nich są stanami intencjonalnymi (stanami dotyczącymi czegoś; świadomymi tranzytywnie), a inne nie (zwykle jakościowe lub całościowe; świadomymi intranzytywnie); (3) mają pewne cechy jakościowe czy też fenomenalne; (4) tworzą strumień świadomości, na którego temat podmiot może utworzyć narrację.

Żadna z istniejących teorii nie wyjaśnia, w jaki sposób powstają wymienione istotne cechy czy też wyznaczniki świadomości. Niemniej jednak ujęcia obliczeniowe potrafią ująć przynajmniej sporą część z nich, choć być może nie wyjaśnią natury stanów jakościowych.

1. Motywacje i natura podejścia obliczeniowego

Po czym można poznać, że świadomość jest wyjaśniona? Pod jakim względem i co należy wyjaśnić? Wiemy, jakie cechy świadomości się wskazuje, ale nie wiemy, jak analizować samą operację wyjaśniania. Dopóki nie ustali się ram, w których to wyjaśnienie ma się lokować, dyskusja może być trudna. Zakładam, że świadomość jest zjawiskiem (również) biologicznym. W związku z tym wyjaśnianie świadomości można przeprowadzić zgodnie z analizą Tinbergena (1963) pod czterema względami:

- funkcji świadomości,
- jej pochodzenia ewolucyjnego (w sensie filogenezy),
- jej mechanizmów,
- a także jej rozwoju (ontogenezy).

Pochodzenie ewolucyjne ani rozwój świadomości u poszczególnych osobników nie jest wyjaśnialne w sposób obliczeniowy, o ile wszystko nie ma natury obliczeniowej[1]. Nawet jeśli procesy ewolucyjne cechują się

[1] Zwolennicy tzw. pankomputacjonizmu każde wyjaśnienie uznawać będą za obliczeniowe, gdyż każdy proces uznają za obliczeniowy. To jednak zwycięstwo natury czysto werbalnej, gdyż również i pankomputacjonista odróżni komputer podłączony do sieci elektrycznej i wykonujący obliczenia od uszkodzonego i niesprawnego. Otóż urządzenia, które normalnie zwiemy „komputerami", w pankomputacjonizmie realizują obliczenia na co najmniej dwóch poziomach: poziomu algorytmu „fizycznego" (czyli praw fizyki)

dużą regularnością, to wydaje się, że jako takie nie stanowią bynajmniej odrębnych mechanizmów względnie odizolowanych od otoczenia, których struktura i interakcje wewnętrzne podlegałyby typowemu wyjaśnianiu mechanistycznemu na poziomie obliczeniowym (por. Miłkowski 2009b). Założę więc ostrożnie, że oponent komputacjonizmu nie musi się mylić, sądząc, że eksplanacja obliczeniowa nie chwyta wszystkich istotnych aspektów świadomości, które trzeba byłoby wyjaśnić, traktując świadomość tak jak każde inne zjawisko ewolucyjne. Jeśli się myli – tym lepiej dla komputacjonizmu, lecz dla obliczeniowych teorii świadomości będzie to kwestia obojętna (gdyż one operują na innym poziomie złożoności niż poziom całego procesu ewolucyjnego).

Są jednak aspekty, które z pewnością można wyjaśnić obliczeniowo. Jedną z funkcji świadomości – co uznaję za mało kontrowersyjne – jest przetwarzanie informacji o środowisku, np. planowanie działań i spostrzeganie (czyli odbieranie wrażeń zmysłowych na jawie), a także przetwarzanie informacji o stanie ciała[2]. Jest to przetwarzanie w pełnym sensie tego słowa: świadomość nie tylko zawiera nośniki informacji (znaki czy reprezentacje), jak np. książka czy drewno ze słojami, ale w ramach organizmu informacje te są wykorzystywane do działania, a więc mają rolę eksplanacyjną w zachowaniu organizmu (por. Miłkowski 2008). Ponieważ w biologii mechanizmy tłumaczą istnienie funkcji, obliczeniowo należy wyjaśnić istnienie informacyjnych zdolności świadomości. Zakłada się więc, że istnieją mechanizmy, które służą do przetwarzania informacji i tłumaczy się ich strukturę oraz zachodzące w nich procesy.

Mówiąc krótko, wyjaśnienie obliczeniowe dotyczy świadomego przetwarzania informacji: tłumaczy się pojawienie tej funkcji świadomości, pokazując, jakie są jej mechanizmy. Nie wyjaśnia się natomiast, jakie ma ona własności adaptacyjne ani jak wpisuje się w osobniczy kalendarz rozwojowy. Jest to bowiem niemożliwe w kategoriach obliczeniowych.

---

i poziomu algorytmu szczegółowego (czyli rzeczywiście zaimplementowanych przez człowieka procesów obliczeniowych w komputerze). Por. Miłkowski 2007. Kiedy piszę niżej o obliczeniach (bez kwalifikatora), mam na myśli obliczenia w sensie wąskim, czyli obliczenia poziomu wyższego niż poziom algorytmu fizycznego.

[2] Pod tym względem świadomość *nie* różni się od nieświadomych procesów przetwarzania informacji. Niektóre teorie jednak wskazują, że sposób przetwarzania informacji może doprowadzić do ich uświadomienia (por. niżej teoria pamięci roboczej Baarsa).

Założeniem sensowności zamiaru wyjaśniania świadomości jest to, że nie jest ona epifenomenem. Być może niektóre aspekty świadomości są epifenomenalne (niektórzy uznają, że takie są jakości przeżyć świadomych), lecz trudniej uznać, że świadome przetwarzanie informacji ma taki charakter. Oczywiście, konsekwentny epifenomenalista może twierdzić, że to nieświadome przetwarzanie informacji ma rolę przyczynową w zachowaniu, a nie uświadomienie sobie tych informacji. Weźmy jednak prosty przykład. Zwykle sądzi się, że to zaburzenia świadomości, np. wywołane spożyciem alkoholu, upośledzają umiejętność kierowania pojazdami. Epifenomenalista będzie twierdził, że upośledzają przetwarzanie informacji (co jest procesem wywołanym na drodze chemicznej), a świadome wrażenia towarzyszące upojeniu alkoholowemu jedynie towarzyszą temu upośledzeniu, nie zaś je powodują. Jednak jednocześnie w języku epifenomenalisty odtworzone zostaną wszystkie różnice pojęciowe, które normalnie wiążemy ze świadomością: odróżni się stan jawy od snu, ślepotę od widzenia (możliwości odbierania wrażeń wzrokowych), daltonizm od normalnego widzenia barw (a więc pewne zdolności odbioru jakości wrażeń), śpiączkę od śmierci... Skoro tak, to istnieje dokładny przekład z naszego języka na epifenomenalistyczny. Świadomość nie występuje w nim po prostu pod własną nazwą, tylko pod nazwą złożoną. Strukturalnie jednak oba języki są równoważne pod względem mocy wyrazu. A to z kolei sugeruje, że epifenomenalizm jest stanowiskiem czysto werbalnym.

1.1. Teoria informacji: przetwarzanie a przesyłanie

Pojęcie „informacji" należy do najmodniejszych w nauce; pojawia się w bardzo wielu znaczeniach i kontekstach (por. Poczobut 2005). Powstaje więc pytanie, w której teorii informacji można wyjaśniać działanie mechanizmu przetwarzania informacji umysłowych. Zgodnie z moją propozycją, najlepiej nadają się do tego narzędzia informatyczne stosowane w obliczeniowym paradygmacie w kognitywistyce. Tylko bowiem w języku informatyki w wystarczająco szczegółowy sposób można opisać nie tylko struktury informacyjne (struktury danych, reprezentacje), ale i ich przetwarzanie.

Struktury informacyjne można opisywać oczywiście bez powoływania się na pojęcia informatyczne. Jak wiadomo, klasyczna teoria informacji Shannona jest matematyczną teorią służącą do opisywania informacji przesyłanych w kanale zawierającym szum. Znakomicie pozwala wyliczyć

prawdopodobieństwo uzyskania poprawnego sygnału przy znanym szumie, ale sama nie opisuje szczegółowo mechanizmów przetwarzania informacji. Wynika to stąd, że teoria ta nie posiada odpowiednich środków technicznych, które umożliwiłyby np. definiowanie różnych reprezentacyjnych struktur danych; można w niej jedynie zdefiniować kod. Samo pojęcie kodu nie pozwala jednak na opisanie jego regularnych przekształceń. Co więcej, systemy poznawcze nie tyle przesyłają informacje, a więc nie tyle są zainteresowane jej dokładnym odtworzeniem, ile ją selektywnie wykorzystują do generowania nowych informacji – przede wszystkim w działaniu itd. Procesów selekcji nie da się opisać środkami teorii Shannona, gdyż jest ona – jak głosi tytuł publikacji Shannona i Weavera (1948) – matematyczną teorią komunikacji.

Istnieje bardzo wiele konkurencyjnych teorii informacji semantycznej, które mają umożliwić opisanie nie tylko ilościowych zależności w kodzie i prawdopodobieństw kolejnych elementów sygnału, lecz również jej treściowy wymiar. Same te teorie zwykle jednak służą do dosyć ogólnikowego opisu znaczenia (rozumianego najczęściej jako konotacja, niekiedy też jako denotacja) informacji. Na przykład przyczynowa teoria odniesienia jest zbyt ogólnikowa, aby mogła dokładnie opisać zależności między reprezentacjami umysłowymi – stanowi tylko ramę dla dokładniejszych narzędzi. Inne teorie z kolei dosyć łatwo przekształcić z abstrakcyjnej teorii filozoficznej (np. semantyki sytuacyjnej) na narzędzia stosowane w informatyce (semantyka sytuacyjna może zainspirować opis systemów rozproszonych, por. Barwise i Seligman 1997, co z kolei zbliża się do zastosowań praktycznych). Wpisują się one wówczas w słownik informatyczny[3] i są eksplikowalne w terminach obliczeniowych.

Inną możliwością byłoby wykorzystanie pojęć cybernetycznych lub samej teorii sterowania. Ale i tu zwykle korzysta się z języka informatyki, choć z powodów – jak sądzę – ideologicznych niektórzy autorzy starają się zdystansować od klasycznego ujęcia obliczeniowego[4]. Pojęcia cybernetyczne

[3] *Notabene*, semantyka uprawiana wyłącznie środkami logiki formalnej może być uznawana za semantykę typu obliczeniowego, gdyż systemy logiczne są równoważne co do mocy wyrazu odpowiednim systemom obliczeniowym. Tak więc środki logiki formalnej zaliczam do narzędzi pojęciowych informatyki. Np. książka Barwise'a i Seligmana została wydana w serii poświęconej informatyce teoretycznej.

[4] Mam na myśli głównie zwolenników tzw. dynamicznego podejścia do systemów poznawczych.

same w sobie świetnie mogą opisywać procesy przetwarzania informacji i zachowanie systemów złożonych, jednak niektóre zależności dosyć trudno w nich wyrazić. Załóżmy, że chcemy wyjaśnić działanie systemu, który posługuje się regułą *modus ponens* we wnioskowaniu. Wyrażenie zależności między wejściem a wyjściem systemu za pomocą złożonego równania różniczkowego, czyli klasyczny opis cybernetyczny, wydaje się dosyć kłopotliwe i nie ujawnia prostej logicznej reguły w zachowaniu systemu. To samo dotyczy teorii sterowania. Dlatego też w takich przypadkach warto posługiwać się pojęciami z zakresu teorii informacji, logiki i informatyki. Należą one bowiem wszystkie do tej samej kategorii: są to abstrakcyjne pojęcia matematyczne i mają identyczny status. Jako takie nie są żadnymi teoriami empirycznymi, tylko matematycznymi środkami wyrazu stosowanymi do opisu rzeczywistości empirycznej.

Warto zauważyć, że opis cybernetyczny czy opis w kategoriach teorii sterowania nie wykluczają opisu obliczeniowego. Są to opisy mechanizmów innego poziomu: teoria sterowania zwykle zajmuje się oddziaływaniem całego systemu z otoczeniem, pomijając jego strukturę wewnętrzną (choć w równaniu stanu układu dynamicznego ze względów technicznych wygodniej jest postulować stany wewnętrzne), natomiast obliczeniowo opisuje się przede wszystkim endostrukturę systemów przetwarzających informacje. Wyjaśnianie obliczeniowe jest na tyle ogólne, że wyjaśnia złożone mechanizmy przetwarzania informacji, lecz interakcje ze środowiskiem o naturze sprzężenia zwrotnego łatwiej opisać w kategoriach cybernetycznych.

1.2. O jakie obliczanie chodzi?

„Obliczanie" rozumiem szeroko. Obejmuje ono zarówno komputery cyfrowe, jak i analogowe. Ponieważ w innym miejscu dokładniej definiowałem mocne pojęcie obliczania i podawałem odpowiednie kryteria jego stosowania (Miłkowski 2009a), tu ograniczę się do skrótowych uwag. Przez proces obliczeniowy rozumiem taki zintegrowany i względnie odizolowany od otoczenia, zachodzący w jednolitym mechanizmie proces, który ma stan początkowy i stan końcowy, przy czym w stanie początkowym mogą występować stany będące wartościami wejściowymi, a w końcowym – będące wartościami wyjściowymi. W komputerach cyfrowych wartości są symbolami nad skończonym alfabetem, w analogowych – po prostu wartościami ciągłymi z określonego przedziału. Za pomocą środków informatycznych, głównie

kodu w określonym języku programowania (lub analogicznego modelu obliczeń), opisuje się przejście między stanem początkowym a końcowym. Przejścia między stanami następują wyłącznie na drodze oddziaływania między składnikami procesu obliczeniowego (z wyjątkiem ewentualnych, sporadycznych interakcji z otoczeniem). Obliczenie zachodzi w systemie faktycznie, gdy spełnione są dosyć intuicyjne kryteria adekwatności opisu matematycznego, które można odnieść do różnego rodzaju teorii matematycznych (np. teorii informacji Shannona opisującej kanał informacyjny) – chodzi o prostotę, wartość eksplanacyjną itp.

Wśród kryteriów adekwatności wyjaśniania obliczeniowego istotne jest dokładne określenie przejścia stanów systemu (zgodne z zachodzącymi w systemie związkami przyczynowo-skutkowymi). Innymi słowy, podany algorytm musi dokładnie odpowiadać zmianom w systemie fizycznym, jeśli wyjaśnienie nie jest idealizujące. Im mniej dokładnie algorytm odpowiada zmianom w systemie, tym bardziej idealizacyjne jest wyjaśnienie w jego kategoriach.

W wielu ujęciach występowanie reprezentacji w systemie jest warunkiem koniecznym obliczania (Fodor 1975, s. 73; Newell 1980; Pylyshyn 1984). Niestety, pojęcie reprezentacji nie należy do najjaśniejszych, a więc i dyskusja nad rolą reprezentacji w obliczaniu jest niełatwa. Jeśli przez „reprezentację" rozumieć znak, który ma niepustą denotację i niepustą konotację, to jest oczywiste, że takie reprezentacje nie muszą występować w systemach obliczeniowych. W banalny sposób można zbudować program, który nie zawiera żadnych reprezentacji, np. zmienne w kodzie nie będą się do niczego odnosić lub kod nie będzie zawierał żadnych zmiennych, a jedynie instrukcje. Jeśli jednak dopuścimy możliwość, że reprezentacje będą czysto formalnymi symbolami, wówczas banalnie prawdziwe jest, że obliczanie cyfrowe pociąga występowanie symboli znad skończonego, dyskretnego alfabetu (jest to element definicji formalnych wszystkich modeli obliczeń; w wypadku obliczeń analogowych są to wartości ciągłe, ale w ściśle określonym przedziale). Simon i Newell mają więc rację, że wszystkie komputery są nie tyle maszynami obliczeniowymi, ile raczej maszynami operującymi na symbolach. Te wewnętrzne symbole maszyny nie muszą mieć żadnego odniesienia, aby obliczanie efektywnie zachodziło. Warto zauważyć, że to samo można wyrazić w kategoriach przetwarzania informacji. Otóż każdy system przetwarzający informacje z konieczności musi zawierać informacje, a znaczy to, że będzie zawierać znaki, przynajmniej czysto formalne. Istnie-

nie informacji pociąga więc za sobą istnienie reprezentacji. W moim ujęciu takie czysto formalne reprezentacje występują nie tylko we wszystkich systemach przetwarzających informacje, ale w ogóle we wszystkich systemach semiotycznych, czyli znakowych (np. słoje w drewnie są reprezentacjami, w tym wypadku nawet mającymi konotację i denotację). Ponieważ każdy skutek z konieczności jest też informacją o swojej przyczynie, to wszelkie związki przyczynowe można by postrzegać jako reprezentujące. Trzeba pamiętać, że to pojęcie reprezentacji jest bardzo słabe: z tego, co napisałem, nie wynika przecież, że kula bilardowa zawiera zdaniowy opis przyczyny toru swojego ruchu. Format wewnętrznych reprezentacji (czy też formalnych symboli) systemu zależy od struktury układu, a więc także mechanizmu obliczeniowego, implementującego określony model obliczeń.

Jeśli mamy świadectwa, że w systemie występują reprezentacje – i to reprezentacje w sensie mocniejszym niż tylko formalne symbole czy skutki przyczyn pozasystemowych – to może to wskazywać, że układ przetwarza informacje, a więc oblicza (operuje na symbolach, jak powiedzieliby Simon i Newell). Innymi słowy, występowanie faktycznie przetwarzanych reprezentacji w sensie mocniejszym jest warunkiem wystarczającym do uznania systemu za obliczeniowy. Samo występowanie reprezentacji w sensie słabszym – nie, gdyż muszą być jeszcze spełnione dodatkowe warunki, aby mechanizm można było uznać za *przetwarzający* informacje.

Wyjaśnianie obliczeniowe jest rodzajem wyjaśniania mechanistycznego (por. Piccinini 2007). Oznacza to, że takiemu wyjaśnianiu poddają się układy będące systemami względnie izolowanymi od otoczenia; w układach tych muszą występować części, których wzajemne oddziaływanie będzie tłumaczone w kategoriach obliczeniowych. Systemami obliczeniowymi w mocnym sensie są tylko takie układy, w których części powstają regularnie ze względu na typ systemu (np. ze względu na zamysł konstruktora lub kod DNA) – tylko bowiem w takich systemach części mogą być funkcjonalne w odpowiednim biologicznym sensie (por. Krohs 2009).

W proponowanym przeze mnie ujęciu obliczanie jest równoważne przetwarzaniu informacji; obejmuje więc także klasyczne pojęcie obliczania cyfrowego, modele symbolicznej sztucznej inteligencji, koneksjonizm itp.

1.3. Kłopoty komputacjonizmu

Ponieważ pojęcia obliczeniowe i informacyjne mają charakter matematyczny, mogą być stosowane podobnie jak fundamentalne pojęcia fizyczne – do obiektów fizycznych dowolnej skali. Może to rodzić poczucie, że są one same w sobie fundamentalnymi pojęciami fizycznymi. Jest to stanowisko tzw. pankomputacjonizmu, który głosi, że wszystkie procesy fizyczne są w istocie obliczeniowe. Zgodnie z moim ujęciem, pankomputacjonizm jest fałszywy – a to ze względu na liczne dodatkowe założenia i kryteria, które przyjmuję w celu wzmocnienia pojęcia obliczania (por. Miłkowski 2009a). Niemniej jednak uniwersalność pojęć obliczeniowych i informacyjnych niesie ze sobą swoisty problem – skoro w taki sposób można opisać bardzo dużą klasę obiektów (nawet przy przyjęciu moich obostrzeń), to pojawia się zarzut, że są to opisy puste, bo czysto formalne. Ale dokładnie to samo można by powiedzieć o aparacie matematycznym w fizyce współczesnej. Jeśli jednak nie negujemy realności własności opisywanych matematycznie, to ten sceptycyzm jest słabo motywowany (por. też Miłkowski w druku).

Zakładam, że pojęcie obliczania można na tyle sprecyzować, iż nie będzie poczucia zbytniej liberalności w jego stosowaniu. Ważniejsze są jednak zarzuty, które wysuwa się wobec samej idei obliczeniowego wyjaśniania świadomości.

Jednym z nich jest to, że nie wyjaśni się w ten sposób jakości przeżyć świadomych[5]. Pewne aspekty przeżyć można tłumaczyć obliczeniowo (tak czyni np. Jackendoff 1987), lecz przyjmijmy, że istotnie obliczeniowe wyjaśnienie istnienia *qualiów* jest niemożliwe. Nie byłoby to jednak szczególnie zaskakujące, skoro nie istnieje też kompletne obliczeniowe wyjaśnienie, dlaczego piksele na monitorze komputera mają określony kolor (oczywiście przy założeniu, że pankomputacjonizm jest fałszywy)[6].

---

[5] W literaturze określa się je mianem *qualiów*, lecz pojęcie to rozumie się zwykle węziej niż tylko „jakość przeżycia świadomego", a sama definicja *qualiów* jest przedmiotem kontrowersji. Mówi się też często o świadomości fenomenalnej, która ma się cechować swoistymi cechami modalności zmysłowych. Por. Ciecierski 2003.

[6] Przy założeniu, że pankomputacjonizm jest prawdziwy, argument przebiega identycznie, tyle że zamiast wyrażenia „fizyczny" należy użyć wyrażenia „algorytm poziomu fizycznego". Obliczeniowe procesy zachodzące w karcie graficznej komputera są procesami innego poziomu niż procesy zachodzące w monitorze ciekłokrystalicznym, zwłaszcza takim, który jest np. zadrapany lub mruga po włączeniu w cykliczny, lecz nie zaplanowany przez programistę komputera sposób, bo wysechł elektrolit w kondensatorze

Obliczeniowo można wyjaśnić procesy, które generują odpowiedni sygnał na wejściu monitora (cyfrowy lub analogowy), lecz sam proces wyświetlania może zajść bez użycia jakiegokolwiek komputera, podobnie jak żarówka do świecenia nie wymaga obróbki żadnych danych. Koloru pikseli nie sposób wyjaśnić nie odwołując się do nieobliczeniowego pojęcia światła i innych pojęć optycznych, a także do własności chemicznych materiału, z którego wykonano ekran (np. ciekłokrystaliczny). Fizyczna *bezpośrednia* przyczyna, która sprawia, że piksel świeci kolorem białym, nie jest wyjaśnialna *całkowicie* w sposób obliczeniowy. Być może analogiczne ograniczenie dotyczy modalności zmysłowych, których działanie nie da się wyjaśnić wyłącznie w kategoriach obliczeniowych.

Nie świadczy to jednak przeciwko teorii obliczeniowej świadomości, lecz pokazuje jej ograniczenia; nie jest to teoria wszechogarniająca, a jedynie opisująca mechanizmy przetwarzania informacji. Sam wynik przetwarzania informacji – określony układ wrażeń w danej modalności zmysłowej – nie musi być już objęty takim wyjaśnianiem, co nie wyklucza, rzecz jasna, że takie wyjaśnienie jednak się powiedzie. Możliwe bowiem, że generowanie jakości przeżyć świadomych jest wynikiem oddziaływania nie tyle surowych procesów fizycznych, ile pewnych mechanizmów obliczeniowych wyższego poziomu. To jednak dzisiaj kwestia wyłącznie spekulacji, od której do prototeorii jeszcze bardzo daleko.

2. Obliczeniowe i alternatywne teorie świadomości

Świadome przetwarzanie informacji jest wyjaśniane współcześnie przede wszystkim w kategoriach obliczeniowych. Do najbardziej znanych ujęć należą:

- teoria globalnej przestrzeni roboczej,
- teoria modelu świata,
- teoria myśli drugiego rzędu,
- integracyjna teoria informacji.

znajdującym się w zasilaczu. Poziom danego procesu obliczeniowego najłatwiej zresztą wyróżnić, kiedy spróbujemy określić, w którym procesie trzeba wprowadzić zmiany z powodu pojawiającego się błędu lub uszkodzenia.

Istnieje też wiele ujęć hybrydowych, łączących elementy kilku teorii, gdyż nie wykluczają się one wzajemnie (nieco większy, ale nadal zwięzły przegląd podaje Seth 2007). Pokrótce przedstawię założenia wskazanych koncepcji, a następnie zajmę się trzema przypadkami, które zwykle uznaje się za przykłady wyjaśnień nieobliczeniowych. Chodzi mi w tym wypadku o:

– kwantową teorię świadomości Hameroffa,
– dualistyczną koncepcję Chalmersa,
– radykalny eksternalizm.

2.1. Przegląd teorii obliczeniowych

Zgodnie z teorią globalnej przestrzeni roboczej świadome stają się te stany przetwarzania informacji w systemie poznawczym, które pojawiają się w swoistej, globalnej przestrzeni roboczej. Głównym orędownikiem tej teorii jest Bernard Baars (1988). Jest ona rozwijana zarówno w ujęciu neuronalnym, jak i w badaniach nad sztuczną inteligencją. Pewnym wariantem tej koncepcji jest też teoria wielokrotnych szkiców Daniela Dennetta (2007). Teoria ta ma implementacje komputerowe, służące do badania jej własności (Baars, Franklin 2009). Niektórzy starają się ją uzupełnić tak, aby wyjaśniać także jakości przeżyć (Duch 2001).

Teoria Baarsa tłumaczy m.in. ograniczoną pojemność świadomości (przez ograniczoną pojemność przestrzeni roboczej), jej sekwencyjną naturę, a także to, że zdarzenia świadome mogą generować nieświadome procesy mózgowe. Baars podkreśla jednak, że globalna przestrzeń robocza jest blisko powiązana ze świadomymi przeżyciami, lecz samo znajdowanie się w globalnej przestrzeni nie jest równoznaczne z byciem przeżyciem świadomym. Innymi słowy, teoria ta nie ujmuje jakościowego charakteru zdarzeń świadomych, lecz stara się pokazać architekturę procesów przetwarzania informacji, które prowadzą do powstania m.in. świadomych przeżyć. Teoria ta wyjaśnia jednak, w jaki sposób stany przetwarzania informacji stają się stanami świadomymi, a więc pozwala rozróżniać stany nieświadome i świadome. Jest więc teorią stanów świadomych, a nie świadomego podmiotu (i nie wyjaśnia np. różnicy między jawą a snem).

Nieco mniejszą popularnością cieszy się teoria myśli drugiego rzędu (*Higher-Order Thought* – HOT) Davida Rosenthala (2005). Ona również ma na celu wyjaśnienie procesu uświadamiania, a jednocześnie swoistej

struktury samoświadomości, która towarzyszy świadomości (przynajmniej u człowieka). Zdaniem Rosenthala, myśl staje się świadoma, gdy jest przedmiotem innej myśli. Same myśli drugiego rzędu rzadziej bywają uświadamiane, lecz są przyczyną uświadamiania innych. W pewnej mierze teoria ta przypomina więc tradycyjną koncepcję zmysłu wewnętrznego.

Teoria Rosenthala ma dosyć oczywistą interpretację obliczeniową: otóż świadome są te reprezentacje, które są przedmiotem innych reprezentacji. Istnieje też wiele jej wariantów, m.in. Robert van Gulick (2004) łączy koncepcję HOT z koncepcją globalnej przestrzeni roboczej (tzw. koncepcja HOGS). Zaprogramowanie systemu, który posiada reprezentacje drugiego rzędu, jest banalnie proste – wystarczy, że system generuje kolejną metareprezentację na podstawie istniejących już w nim reprezentacji; np. system rozpoznający twarze i generujący na ekranie dodatkowe informacje dla człowieka z pewnością miałby w tym sensie uświadomione informacje na ekranie. Ze względu na zbyt daleko idącą prostotę – a jednocześnie kontrintuicyjność – prosta teoria HOT nie cieszy się zbyt dużym powodzeniem wśród informatyków kognitywistycznych.

Na przecięciu informatyki, modelowania statystycznego i neurologii leży natomiast integracyjna teoria informacji. Jej twórcą jest współpracownik Geralda Edelmana, Gulio Tononi (2004). W tym ujęciu świadomość cechuje się dwoma podstawowymi cechami: (1) ma naturę informacyjną; (2) jest silnie zintegrowana. Tononi opracował miarę złożoności integracyjnej dla sieci przesyłających informacje, która ma służyć do eksperymentalnego pomiaru stopnia świadomości. Miara ta nie jest na razie doskonała (zmienia się dosyć diametralnie w czasie dla tej samej sieci, a więc złożoność silnie zależy od stanu początkowego, w którym rozpoczyna się pomiar), a także bardzo trudna do efektywnego zastosowania; zainspirowała jednak szeroko zakrojone badania nad znalezieniem innych, łatwiejszych w użyciu ujęć. Do integracyjnych teorii – a więc opartych na pewnej mierze złożoności – należą też badania Christofa Kocha (2008). Świadomość ma być efektem integracji informacji w sieci nerwowej; neuronalne korelaty świadomości mają się więc cechować swoistym stopniem zintegrowania (synchronizacji). Mówiąc bardziej filozoficznym językiem, złożoność w teoriach integracyjnych jest inną nazwą emergencji stanów świadomych na stanach nieświadomych.

Warto zauważyć, że rola integracji i złożoności jest również bardzo istotna dla teorii globalnej przestrzeni roboczej (jednym z kryteriów wy-

różniania tej przestrzeni ma być gęstość oddziaływań przyczynowych). Dotychczasowe propozycje teorii obliczeniowych świadomości mają charakter wstępny, lecz zauważalna jest pewna konwergencja różnych ujęć informatycznych; samo przejście na poziom ilościowy świadczy o dojrzewaniu teorii (por. Seth, Dienes 2008). Jak jednak zauważył Ray Jackendoff (1987, s. 17), sama złożoność może być co najwyżej wyznacznikiem powstawania stanów świadomych, nie zaś pełnym wyjaśnieniem, dlaczego one zaistniały.

Jackendoff podkreśla, że nie wystarczy ujęcie struktury procesów prowadzących do powstania świadomości; konieczne jest także zbadanie struktury nośników informacji, czyli formatu reprezentacji świadomych. Zdaniem Jackendoffa, tylko informacje pośredniego szczebla mogą być treścią świadomości. Czyste dane percepcyjne, docierające z organów zmysłowych, nie są nam dostępne w świadomości (najniższy poziom reprezentacji); podobnie nie jesteśmy w stanie operować abstrakcyjnymi myślami w oderwaniu od modalności zmysłowych. Na przykład wyrazami operujemy zawsze w reprezentacji fonologicznej (lub graficznej, jeśli są to wyrazy języka tylko pisanego); nie są nam one dostępne w postaci czysto pojęciowej. Koncepcja Jackendoffa, choć powstawała nieco wcześniej od popularnych obecnie, jest bliska ujęciu Baarsa: bycie świadomym polega na znajdowaniu się w pamięci krótkoterminowej (STM) charakterystycznej dla danej modalności zmysłowej; reprezentacje z STM, na które skierowano uwagę, znajdują się w samym centrum świadomości.

Nieco inny charakter od poprzednich koncepcji, wyjaśniających głównie proces uświadamiania stanów świadomości – czy to pojedynczo, czy kolektywnie – ma teoria świadomości jako modelu świata (Johnson-Laird 1983; Metzinger 2003). Koncepcja ta nawiązuje nie tylko do faktu, że istnieje samoświadomość, co podkreślają teorie typu HOT, i że świadome informacje są globalnie dostępne oraz wykazują się daleko idącą integracją, ale także do tego, że integracja ta służy do budowania modelu wykorzystywanego w działaniu w świecie. Jednym z tych modeli jest model jaźni, który wyjaśnia integrowanie się uświadamianych informacji i ich znaczenie dla działania. Prekursorem świadomego modelu jaźni jest model nieświadomy, który zaimplementowano np. dla niektórych robotów (Bongard, Zykov, Lipson 2006). Natomiast podmiotowość związana ze świadomością wiąże się, zdaniem Metzingera (2003), z istnieniem fenomenalnego modelu jaźni, który jest wewnętrzną i dynamiczną reprezentacją organizmu, nierozpoznawaną jako reprezentacja (jest przezroczysta). Mówiąc w największym

uproszczeniu, jedną z kluczowych funkcji ludzkiej świadomości ma być ciągłe współreprezentowanie samej relacji reprezentowania świata w postaci dynamicznego modelu.

Podobne, symulacyjne ujęcie świadomości prezentuje Revonsuo (2005), który podkreśla, że świadomość jest rodzajem wirtualnej rzeczywistości. Ta wirtualna rzeczywistość pojawia się też we śnie, a jawa różni się tym, że następuje większa interakcja ze środowiskiem.

2.2. Teorie alternatywne

Do najbardziej popularnych koncepcji alternatywnych należą koncepcje oparte na mechanice kwantowej. Spekulacje związane z fizyką kwantową i świadomością mają różnorodny charakter; niektóre są związane z dualistycznym ujęciem kolapsu w mechanice kwantowej. Najbardziej jednak znaną teorią jest tzw. zorkiestrowana teoria świadomości (Orch OR) Stuarta Hameroffa (1998). Jak wiadomo, Hameroff wraz z Rogerem Penrose'em twierdzą, że świadomość ma zdolności przekraczające możliwości zwykłych komputerów (Penrose popiera to argumentem dotyczącym twierdzenia Gödla; por. Penrose 1995; Krajewski 2003). Jakie to są jednak możliwości? Intuicyjne rozpoznawanie prawd matematycznych jest jednak zdolnością z gatunku obliczeniowych. Hipotetyczne kwantowe efekty w makroskali mają bowiem zapewniać istnienie neurokomputera, który jest jednocześnie komputerem kwantowym, lecz jest to cały czas komputer (por. Hameroff 2007). Efekty kwantowe mają polegać w tym wypadku na istnieniu całościowej synchronizacji, czyli integracji na dużą skalę wielu komórek nerwowych.

Jest więc jasne, że koncepcja Hameroffa – mimo że wychodzi od krytyki tradycyjnego komputacjonizmu i opiera się na niestandardowej wykładni twierdzenia Gödla – jest odmianą teorii obliczeniowej. Można ją uznać za typ teorii integracyjnej, bo efekty kwantowe mają wspierać integrację informacji.

Wydawać by się mogło, że obliczeniowego charakteru nie mają koncepcje Davida Chalmersa (1996). Jego zdaniem, trudnego problemu świadomości – problemu, dlaczego w ogóle istnieją jakości przeżyć świadomych – nie sposób rozwiązać w sposób funkcjonalistyczny i obliczeniowy. Według niego istnienie takich jakości zależy od własności wewnętrznych przeżyć,

które to własności są skorelowane (nieznanymi dotychczas) fundamentalnymi prawami psychofizycznymi ze światem fizycznym.

Jednocześnie jednak, zdaniem Chalmersa, zwolennika mocnej sztucznej inteligencji, „łatwe problemy", też dotyczące przetwarzania informacji w sposób świadomy, można rozwiązywać w sposób obliczeniowy. Ponieważ, jak już wspomniałem, wyjaśnianie obliczeniowe nie może z konieczności wyjaśnić wszystkich aspektów świadomości, stanowisko Chalmersa trudno uznać za opozycyjne. Jest tylko bardzo swoistym stanowiskiem w ramach szeroko pojętego paradygmatu obliczeniowego.

Natomiast przedstawiciele radykalnego eksternalizmu (a także pokrewnych koncepcji, np. dynamicznego ujęcia poznania czy niektórych odłamów enaktywizmu) ujmują świadomość jako proces niereprezentacyjny, zachodzący między środowiskiem a mózgiem (Manzotti 2006)[7]. Manzotti ujmuje wszystkie procesy świadome jako rodzaj percepcji; i tak np. pamięć jest jego zdaniem opóźnioną percepcją (oczywistym problemem dla tego rodzaju koncepcji jest percepcja niewerydyczna i halucynacje, których przekonującego wyjaśnienia autor nie jest w stanie podać).

Manzotti, a wraz z nim wielu przedstawicieli robotyki inspirującej się enaktywizmem i cybernetyką, podkreśla, że świadomość należy postrzegać jako proces. Wydawać by się mogło, że Manzotti będzie więc odżegnywać się od wyjaśnienia świadomości w sposób obliczeniowy. Tymczasem jest wręcz przeciwnie; jest on gorącym orędownikiem badania świadomości w nurcie tzw. świadomości maszynowej (*machine consciousness*). Jego prace zmierzają do stworzenia sztucznej architektury świadomości, opartej na niesubstancjalnym pojmowaniu świadomości (Manzotti 2003). Innymi słowy, nawet bardzo odmienne ujęcie świadomości – postulujące zupełnie inną ontologię – pozostaje w ramach paradygmatu obliczeniowego. O ile mi wiadomo, w tym radykalnym nurcie na razie nie zaproponowano żadnych modeli biologicznych systemów poznawczych; tworzy się modele sztuczne.

---

[7] Oczywiście, w moim ujęciu takie postawienie sprawy stanowi sprzeczność pojęciową: świadomość jako informacyjna ma naturę reprezentacyjną z konieczności. Manzotti jednak uważa, że samo założenie, że reprezentacja jest odmienna od tego, co reprezentuje, stanowi świadectwo dualizmu ontologicznego typowego dla metafizyki XVII w. W ujęciu Manzottiego reprezentacje są tożsame z tym, co reprezentują (Manzotti 2003, s. 6); są one w istocie procesami o naturze prezentacyjnej (a nie reprezentacyjnej). W takim ujęciu niesłychanie trudno poradzić sobie z reprezentacjami niewerydycznymi.

3. Pozorne alternatywy

Mógłbym oczywiście listę pozornych wyjaśnień alternatywnych ciągnąć, gdyż badania nad świadomością są dziedziną bogatą i dopuszcza się w nich do głosu nawet dosyć egzotyczne pomysły. Istnieją oczywiście ujęcia silnie biologizujące, w których świadomość z konieczności występować musi tylko w organizmach biologicznych (a więc nie da się jej replikować w sposób obliczeniowy w systemach sztucznych). Jednak brak możliwości takiej replikacji nie oznacza jeszcze, że sam obliczeniowy sposób wyjaśniania świadomości będzie niepoprawny. Być może po prostu dodatkowe własności mechanizmów świadomości, o naturze innej od samego przetwarzania informacji, będą rozstrzygać o niemożności wytworzenia jej w systemach sztucznych. Pozostajemy jednak tutaj w sferze spekulacji.

Wyraźne jest jednak jedno: obecnie nie istnieją żadne dokładne wyjaśnienia mechanizmów świadomego przetwarzania informacji, które nie miałyby charakteru informacyjnego. Niekiedy są to wyjaśnienia dosyć nietypowe (jak u Hameroffa) lub mają postać pewnych postulatów do tworzenia systemów sztucznych (jak u Manzottiego), lecz pozostają w szerokich ramach komputacjonizmu.

Twierdzę więc, że wyjaśnianie obliczeniowe jest dziś najlepszym narzędziem eksplanacji funkcjonowania złożonych systemów przetwarzania informacji. Nie jest to jednak nigdy jedyne wyjaśnienie tych systemów: mechanizmy obliczeniowe z konieczności są tylko jednym z wielu innych mechanizmów w takich układach. Są one bowiem realizowane przez mechanizmy niższego poziomu, które wyjaśnia się w kategoriach czysto chemicznych, biologicznych lub elektronicznych[8]. W języku czysto informatycznym nie wyjaśnia się *całkowicie* specyfiki mechanizmów niższego poziomu: skąd biorą się kolory na monitorach, skąd bierze się dźwięk w głośnikach i dlaczego toner przylega do papieru w drukarce laserowej, a także dla-

[8] Nawet jeśli prawdziwe są koncepcje pankomputacjonistyczne, to prawdą pozostaje, że mechanizm obliczeniowy wyższego poziomu jest wyjaśniany *inaczej* niż mechanizm niższego poziomu. Prawdopodobieństwo, że realizują te same algorytmy, jest zresztą nikłe, a z pewnością operują one na innych danych. Z punktu widzenia twórcy obliczeniowego modelu świadomości nie ma więc znaczenia, czy jest on implementowany na komputerze, który jest implementowany na kolejnym komputerze, czy też realizowany środkami nieobliczeniowymi. Poziom realizacji niskiego poziomu jest dla tej teorii i tak niedostępny i się ona do niego nie odnosi, a więc także go nie wyjaśnia.

czego niektóre wyrażenia językowe mają odniesienie w rzeczywistości pozajęzykowej. Oznacza to, że każda obliczeniowa teoria ma cząstkowy charakter. Dotyczy to też obliczeniowych teorii świadomości. Możliwe, że nie wyjaśnią wielu jej aspektów, lecz nic nie wyjaśnia lepiej, w jaki sposób w świadomości integrowane są informacje z otoczenia organizmu. W teoriach tych stawia się ścisłe hipotezy, które mają tłumaczyć, jak zachodzi uświadamianie informacji nieświadomych, a niekiedy też i szerszą funkcję uświadamianych informacji, które są unifikowane w postaci dynamicznych modeli. Wyjaśniają więc powstawanie stanów świadomych, ich wzajemne oddziaływania oraz wskazują, jakie architektury mogłyby implementować mechanizmy, w których pojawiałyby się te stany.

Literatura

Baars B. 1988, *A Cognitive Theory of Consciousness*, Cambridge, MA: Cambridge University Press.

Baars B., Franklin S. 2009, *Consciousness is Computational: the LIDA Model of Global Workspace Theory*, „International Journal for Machine Consciousness" 1, s. 23–32.

Barwise J., Seligman J. 1997, *Information Flow. The Logic of Distributed Systems*, Cambridge UP.

Bongard J., Zykov V., Lipson H. 2006, *Resilient Machines Through Continuous Self-modeling*, „Science" 314, s. 1118–1121.

Chalmers D. 1996, *The Conscious Mind*, Oxford: Oxford University Press.

Ciecierski T. 2003, *O objaśnianiu pojęcia świadomości fenomenalnej*, „Przegląd Filozoficzno-Literacki" 4 (6), s. 33–47.

Dennett D. 2007, *Słodkie sny*, tłum. M. Miłkowski, Warszawa: Wydawnictwo Prószyński i S-ka.

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

Fodor J. 1975, *Language of Thought*, Cambridge, Mass: Harvard University Press.

Hameroff S. 1988, *Quantum Computation in Brain Microtubules? The Penrose-Hameroff „Orch OR" Model of Consciousness*, „Philosophical Transactions Royal Society London (A)" 356, s. 1869–1896.

Hameroff S. 2007, *The Brain is Both Neurocomputer and Quantum Computer*, „Cognitive Science" 31, s. 1035–1045.

Jackendoff R. 1987, *Consciousness and the Computational Mind*, Cambridge, Mass. MIT Press.

Johnson-Laird P. 1983, *Mental Models: Towards a Cognitive Science of Language, Inference, an Consciousness*, Cambridge: Cambridge University Press.

Koch C. 2008, *Neurobiologia na tropie świadomości*, tłum. G. Hess, Warszawa: Wydawnictwa UW.

Krajewski S. 2003, *Twierdzenie Gödla i jego interpretacje filozoficzne: od mechanicyzmu do postmodernizmu*, Warszawa: IFiS PAN.

Krohs U. 2009, *Functions as Based on a Concept of General Design*, „Synthese 166" (1), s. 69–89.

Manzotti R. 2003, *A Process-based Architecture for an Artificial Conscious Being*, „Axiomathes" 00, s. 1–28.

Manzotti R. 2006, *Outline of a Process View of Conscious Perception*, „Journal of Consciousness Studies" 13, 6, s. 45–79.

Metzinger T. 2003, *Being no one. The Self-Model Theory of Subjectivity*, Cambridge MA: MIT Press.

Miłkowski M. 2007, *Is Computationalism Trivial?*, [w:] G. Dodig Crnkovic, S. Stuart (eds.), *Computation, Information, Cognition – The Nexus and the Liminal*, Cambridge Scholars Publishing, s. 236–246.

Miłkowski M. 2008, *Postulowanie utajonych funkcji umysłu: realizm kontra antyrealizm*, [w:] *Utajone funkcje umysłu*, red. Sz. Wróbel, Poznań–Kalisz: Wydział Pedagogiczno-Artystyczny UAM w Poznaniu, s. 15–37.

Miłkowski M. 2009a, *O tzw. metaforze komputerowej*, „Analiza i Egzystencja" 9, s. 163–185.

Miłkowski M. 2009b, *Is Ewolution Algorithmic?*, „Minds and Machines" 19, 4, s. 465–475.

Miłkowski M. (w druku), *Is Computation Interpretation-based?*, „Semiotica".

Newell A. 1980, *Physical Symbol Systems*, „Cognitive Science" 4, s. 135–183.

Penrose R. 1995, *Nowy umysł cesarza*, tłum. P. Amsterdamski, Warszawa: Wydawnictwo Naukowe PWN.

Piccinini G. 2007, *Computing Mechanisms*, Warszawa: „Philosophy of Science" 74.4, s. 501–526.

Poczobut R. 2005, *Od informacji fizycznej do informacji fenomenalnej*, [w:] *Informacja a rozumienie*, red. M. Heller, J. Mączka, Kraków 2005, s. 177–193.

Pylyshyn Z. 1984, *Computation and Cognition. Toward a Foundation for Cognitive Science*, Cambridge, Mass.: MIT Press.

Revonsuo A. 2005, *Inner Presence: Consciousness as a Biological Phenomenon*, Cambridge, Mass.: MIT Press.

Rosenthal D. 2005, *Consciousness and Mind*, Oxford: Clarendon Press.

Seth A. 2007, *Models of Consciousness*, „Scholarpedia" 2 (1), s. 1328.

Seth A., Dienes Z. 2008, *Measuring Consciousness: Relating Behavioural and Neurophysiological Measures*, „Trends in Cognitive Sciences" 12, s. 314–321.

Shannon C., Weaver W. 1948, *A Mathematical Theory of Communication*, „Bell System Technical Journal" 27, s. 379–423, 623–656.

Tinbergen N. 1963, *On Aims and Methods in Ethology*, „Zeitschrift für Tierpsychologie" 20, s. 410–433.

Tononi G. 2004, *An Information Integration Theory of Consciousness*, „BMC Neuroscience" 5, 42.

Van Gulick R. 2004, *Higher-Order Global States (HOGS): An Alternative Higher--Order Model of Consciousness*, [w:] R.J. Gennaro (red.), *Higher-Order Theories of Consciousness: An Anthology*, John Benjamins.

Van Gulick R. 2009, *Consciousness*, [w:] *The Stanford Encyclopedia of Philosophy*, red. E.N. Zalta, http://plato.stanford.edu/archives/spr2009/entries/consciousness/.

COMPUTATIONAL THEORIES OF CONSCIOUSNESS

Summary

In this paper, I review the motivations for having a computational theory of consciousness to see if they turn out to be no longer plausible in the light of recent criticisms. These criticisms focus on the alleged inability of computational theories to deal with qualia, or qualities of experience (or objects of experience in some accounts), and with so-called symbol grounding. Yet it seems that computationalism remains the best game in town when one wants to explain and predict the dynamics of information processing of cognitive systems. Conscious information processing does

not seem to be explainable better within any other framework; computationalism regarding consciousness can only be discarded by supposing that consciousness is epiphenomenal in information processing.

I will argue that recent theories of consciousness that are to deal with the so-called hard problem of consciousness remain in their core computational if they do not subscribe to epiphenomenalism. For example, the quantum theory as proposed by Stuart Hameroff remains openly computational; the same goes for pan(proto)psychist speculation of David Chalmers. The qualitative character of information processing that Chalmers takes to explain the existence of subjective experience piggy-backs, so to say, on the very fact that there is information processing that is best explained in a computationalist framework. I also briefly show that other alternative accounts of consciousness (such as direct theories of consciousness) that were supposed to oppose computational and functionalist conceptions are not only compatible with them but require them to begin with.

In short, to discard credentials of computationalism in consciousness research one would have to show that it's possible to explain conscious information-processing mechanisms sufficiently in a non-computational way. And this has not been done by any of the critics of computational accounts. This all doesn't suggest, though, that computational explanation is sufficient for building a complete theory of consciousness; it seems however to be necessary.