Przegląd Filozoficzny – Nowa Seria
R. 22: 2013, Nr 2 (86), ISSN 1230–1493
DOI: 10.2478/pfns-2013-0025

Piotr Bołtuć

BICA jako szansa stworzenia świadomych maszyn

Słowa kluczowe: *teza inżynieryjna w świadomości maszyn, Biologicznie Zainspirowane Architektury Kognitywne (BICA), niereduktywna świadomość maszyn, D. Chalmers, N. Block*

W artykule tym oddzielam pojęcie świadomości funkcjonalnej i fenomenalnej. W części pierwszej ilustruję tezę, że maszyny spełniają już pewne kryteria świadomości funkcjonalnej, a należy przewidywać, że kryteria bardziej zaawansowane spełnią w skali kilku czy kilkudziesięciu lat. W części drugiej wyjaśniam pojęcie świadomości fenomenalnej, którego nie bronię, gdyż robię to w innych pracach. Następnie zaś przedstawiam argument, że w naturalistycznym systemie myślenia istnieje miejsce na nieredukowalną świadomość fenomenalną. Skoro taka świadomość byłaby wyjaśnialna na zasadach naturalistycznych, to należy przyjąć, iż może ona zostać skonstruowana także w maszynach, choć nie należy się tego spodziewać w krótkiej skali czasowej. Istnieją problemy epistemologiczne, które spowodowałyby, że tego rodzaju świadomość może być poznana tylko na zasadach miękkiej indukcji, np. poprzez *inference to the best explanation*. Wątek przewodni artykułu stanowi idea, że Biologicznie Zainspirowane Architektury Kognitywne (BICA), jakie obecnie są szeroko dyskutowane w nauce o sztucznej inteligencji (AI), nie powinny ograniczać się do metod programowania. Biologiczne architektury kognitywne opierają się na systemie nerwowym, który stanowi inny rodzaj *hardware* niż znane dzisiaj komputery. Stawiam tezę, iż świadomość pierwszoosobowa (fenomenalna) wymagałaby tworzenia biologicznie zainspirowanej architektury kognitywnej w zakresie *hardware*, a mianowicie stworzenia projektora jaźni w oparciu o specyfikacje poznane u ludzi i wyższych zwierząt.

1. Słów kilka o świadomości funkcjonalnej

Świadomość funkcjonalną definiuje się jako wykonywanie zadań kognitywnych na poziomie istot świadomych. Jednym z najstarszych funkcjonalnych kryteriów świadomości jest test Turinga. Alan Turing twierdził, że maszyna będzie świadoma, jeżeli korespondując z nią przez pięć minut na komunikatorze podobnym do Internetu, nie będziemy w stopniu statystycznie znaczącym umieli określić, czy rozmawiamy z człowiekiem, czy komputerem. Wiele interfejsów próbuje spełnić test Turinga w danych domenach, ale nikomu nie udało się jeszcze zbudować maszyny, która spełniłaby test Turinga w domenie otwartej. Wynika to z faktu, że staje się on testem na oszukiwanie, co jest o wiele trudniejsze niż zachowanie zgodne z prawdą. Jeżeli pytamy komputer o czynności czysto ludzkie, np. co jadł na śniadanie, to prawdziwa odpowiedź wydałaby jego status ontologiczny nie ze względu na sprawność intelektualną interakcji, a na jego treść (komputery nie jedzą śniadań). W nietrywialnych domenach zamkniętych programy komputerowe pokonują test Turinga.

Maszyny spełniają coraz wyższe kryteria świadomości funkcjonalnej. Wiodący dzisiaj program LIDA negocjuje w języku naturalnym przydział służby na lotniskowcach Stanów Zjednoczonych, co jest procesem złożonym, bo dotyczy setek pilotów i innych członków personelu (Franklin). LIDA bierze pod uwagę dużą liczbę zmiennych, takich jak wymogi sprawowanej służby, ważne powody osobiste (np. śluby, pogrzeby), stan zdrowia, wcześniejsze obciążenie i indywidualne preferencje. Wiadome jest, że LIDA wywiązuje się z tych zadań kognitywnych lepiej niż osoby ludzkie, bo żaden człowiek nie może uwzględniać równocześnie tak wielkiej liczby zmiennych. W mniejszej skali, programy oparte na podobnej zasadzie oceniają prace studentów z logiki. Mają one przewagę nad wykładowcami, bo pamiętają wszystkie prace złożone przez danego studenta i mogą np. wywnioskować, że w kolejnych egzaminach błędy wynikają z niedostatecznego opanowania *modus tollens*; radzą więc studentowi jego powtórzenie. Natomiast istnieje mała szansa, by jakiś człowiek oceniający dziesiątki egzaminów takie szczegóły zauważył. Także w wiele gier komputerowych komputery grają tak dobrze lub lepiej niż ludzie. Nie wspominam już o tym, że Deep Blue wiele lat temu pokonał w grze w szachy Kasparowa.

Programy komputerowe pozwalają na funkcjonowanie samolotów i samochodów bezzałogowych. W coraz większym stopniu są to nie tylko pojazdy obsługiwane przez osobę z odległego miejsca, jak się dzieje w przypadku tradycyjnych samolotów bezzałogowych, ale są to w ścisłym sensie pojazdy samosterujące. Potrzeba samosterowalności wynika bezpośrednio z faktu, że mają miejsce zaniki połączeń internetowych pozwalających na kierowanie takim pojazdem. Funkcjonalności rozwinięte np. dla uniknięcia przerw w dzia-

łaniu dronów na polu walki w kolejnych stadiach pozwalają na uzyskiwanie robotów transportowych lub walczących, które przekraczają zdolności nawet najlepszego ludzkiego pilota czy kontrolera.

W Japonii tworzy się „sztucznych kompanów" mających pomagać osobom starszym i dotrzymywać im towarzystwa. Roboty takie wykonują zadania pomocy domowej, ale także spełniają rolę *extended memory*, przechowując np. rodzinne zdjęcia i rozmawiając z osobą, którą się opiekują, tak jak bliski człowiek (Floridi).

Tyle przykładów z dziedzin potocznych. Wskazują one, że jeśli świadomość określimy jako warunek zaawansowanego funkcjonowania w świecie, przypominającego zachowania ludzi, to komputery i napędzane przez nie roboty takie kryteria w znacznej mierze spełniają. Przejdziemy obecnie do zagadnień teoretycznie trudniejszych.

Zgodnie z fizyczną interpretacją tezy Churcha-Turinga każdy proces w rzeczywistości fizycznej może zostać w pełni opisany przez (ciągłą lub nieciągłą) funkcję algorytmiczną. To oznacza, że każde zachowanie także organizmów żywych może zostać w pełni opisane za pomocą algorytmów. Z drugiej zaś strony, wszystkie procesy, jakim odpowiadają funkcje algorytmiczne, mogą zostać zrealizowane przez maszyny; maszyny mogą funkcjonować wedle dowolnej funkcji realizowalnej w środowisku fizycznym (o ile nie istnieją konkretne powody, by tak nie było, np. wynikające z niezwykle małej wielkości danego zjawiska lub z efektów kwantowych). Z powyższego wynika, że jeśli myślenie może być opisane w sposób funkcjonalny, to maszyny mogą lub będą mogły realizować taką funkcjonalność w dokładnie ten sam sposób – wedle tego samego algorytmu – jak ludzie i wyższe zwierzęta. Oczywiście taka identyczność ma miejsce na danym poziomie granularności, np. możliwe jest zrobienie ręki posiadającej te same obserwowalne własności co ręka ludzka. Jednak na niższym poziomie granularności, np. gdy rękę przetniemy i zobaczymy, co jest w środku, identyczność nie będzie zachowana. Czy zatem istnieją specyficznie ludzkie funkcje kognitywne?

Istnieje od dawna znana wątpliwość co do adekwatności funkcjonalnej maszyn, dotycząca twórczości czy możliwości dokonywania przez maszyny odkryć. Utrzymuje się jeszcze gdzieniegdzie mniemanie, że tylko ludzie posiadają umiejętności twórcze, zarówno w sferze twórczości artystycznej, jak również twórczości naukowej. Jest to jednakże pogląd błędny. Działania twórcze podlegają w pełni opisowi algorytmicznemu, w którym jawią się jako przekształcanie przez model świata istniejącego (Laszlo). Obecnie doszliśmy do zaawansowanych wyjaśnień, jakiego rodzaju odchylenia – w sensie matematycznym – powodują najbardziej nowatorskie rozwiązania (Takeshi, Thaler). Wiemy także, jakie mechanizmy drugiego stopnia decydują o zaakceptowaniu danych odkryć lub dokonań artystycznych; w nauce mechanizmy te w spo-

sób ogólny opisuje socjologia wiedzy, a w sztuce teoria *artworld* (Dickie). Mamy obecnie tzw. maszyny twórczości (*creativity machines*), które dokonują w sposób skuteczny własnych odkryć i wynalazków stosowalnych w przemyśle (Thaler). Poznajemy też, że pamięć czy asocjacje opisane jako funkcje, np. tzw. *looping mechanisms*, stanowią niezbędny element pozwalający na myślenie odkrywcze czy twórcze. Również emocje, rozumiane jako zwrócenie szczególnej uwagi na niektóre aspekty rzeczywistości, stanowią ważny element myślenia twórczego, ale one także uzyskują w pełni funkcjonalne odpowiedniki (Takeshi). To prowadzi nas do następnego punktu – zagadnienia roli biologicznie zainspirowanych architektur kognitywnych.

2. Biologicznie zainspirowane architektury kognitywne a tradycyjna sztuczna inteligencja

Odchodzi się coraz częściej od założeń twórców dawnej sztucznej inteligencji, wedle których roboty prześcignęły już mózgi zwierząt i ludzi, zatem ich konstruktorzy nie mogą się niczego przydatnego od neurologów nauczyć. Jest wręcz przeciwnie, nowe badania mózgu wykazały, że jest on bardzo zaawansowaną strukturą kognitywną. Stąd wynika obecne zainteresowanie biologicznie zainspirowanymi architekturami kognitywnymi (Samsonovich). Stanowią one nowy trend wykraczający poza tradycyjne metody AI. Założeniem heurystycznym BICA jest dobrze już dzisiaj poparte empirycznie twierdzenie, że centralny układ nerwowy zaawansowanych zwierząt dysponuje technikami kognitywnymi, jakie nie są dostępne tradycyjnym metodom programowania. Różnice między metodami AI i BICA nie ograniczają się do linearnego przetwarzania informacji ani nawet do tworzenia *neural networks*. Badania mózgu wskazują na holistyczne zapisywanie i przetwarzanie informacji poprzez stany energetyczne poszczególnych jego regionów. Tworzenie sztucznej inteligencji w oparciu o te założenia wymagałoby nie tylko zmian w programowaniu, ale przede wszystkim zasadniczych zmian – nie tylko ulepszeń – w zasadach funkcjonowania maszyn myślących.

To, co pisałem w poprzedniej sekcji o roli *looping mechanism* i o znaczeniu bardzo jasno kwantyfikowalnego stopnia odchylenia od funkcji rzeczywistości, jakie jest niezbędne w procesie odkrycia naukowo-technicznego (zarówno dokonywanego przez ludzi, jak i przez inne systemy kognitywne), jest teraz istotne. Prowadzi nas bowiem do szerszego rozumienia biologicznie zainspirowanej architektury kognitywnej jako struktury posiadającej nie tylko procesy świadome, ale i nieświadome. Bernard Baars w swojej teorii globalnej przestrzeni współpracy (*global workspace theory*) pokazuje, że to, co świadome, stanowi tylko uproszczoną racjonalizację bardziej efektywnych

procesów kognitywnych. Pierwszoosobowa świadomość jest uzasadniona ewolucyjnie jako sposób znakowania różnego rodzaju informacji przynoszonej przez wyspecjalizowane podsystemy układu nerwowego. Podobnie, choć mniej precyzyjnie, podchodzi do świadomości Dan Dennett ze swoją teorią wielu szkiców (*multiple draft theory*). Dennett ma rację, że ani jego podejście, ani Baarsa nie wymaga pierwszoosobowej świadomości, o jakiej piszą Nagel i Chalmers. W następnej sekcji zastanowimy się jednak, czy istnieją powody, aby wprowadzić również taką perspektywę.

3. Pierwszoosobowa jaźń

Wielokrotnie pokazywano, że pierwszoosobowa jaźń nie jest tożsama z jakąkolwiek funkcją kognitywną czy sumą takich funkcji. Głównymi przykładami, czy raczej *case*'ami, jakie mają pokazać niezbędność perspektywy pierwszoosobowej w szeroko zdefiniowanym sensie funkcjonalnym, są: „Czarno-biała Maria" (*black and white Mary*) Franka Jacksona, „Wymóżdżeńcy" (*zombies*) Davida Chalmersa oraz „Chiński pokój" Johna Searle'a. Uważam, w przeciwieństwie do przeważającej większości obrońców jakiejś formy nieredukcjonizmu w filozofii świadomości, że wszystkie te argumenty zawodzą; co więcej twierdzę, że wszystkie argumenty o tej strukturze muszą zawieść (Bołtuć 1988). Tutaj pokażę słabości tych argumentów jedynie skrótowo.

a) „Biała i czarna Maria" to argument, że jeśli ktoś (Maria) byłby wychowany w biało-czarnym środowisku (włączając własne ciało), to nawet gdyby posiadał pełną wiedzę naukową o kolorach (Maria jest światowej sławy ekspertem na temat fizyki kolorów), to mimo wszystko, widząc pierwszy czerwony pomidor, Maria byłaby zdziwiona i dowiedziałaby się czegoś nowego. Zatem, wedle tego argumentu, wiedza fizyczna nie jest pełną wiedzą o świecie. Z istniejących odpowiedzi na ten argument najbardziej odpowiada mi Lewisa i Nemirowa, że Maria uzyskuje nową umiejętność, której nie daje pełna wiedza naukowa – tak jak można wiedzieć wszystko z książek o tenisie, ale nie potrafić w niego grać, albo o pływaniu czy jeździe na rowerze. Trudno powiedzieć, kontynuują autorzy, że umiejętność jazdy na rowerze jest niefizyczna. Ja ze swej strony uważam, że argument Jacksona, który on zresztą w międzyczasie porzucił, opiera się na staromodnym ujęciu nie jednego, ale przynajmniej dwóch założeń. Po pierwsze, Jackson zakłada, że wszystko w świecie fizycznym, naturalistycznym, jest w pełni opisywalne. Wydaje mi się to nieprzekonujące, gdyż można być fizykalistą, a jednak uważać, że pełny opis zachodu słońca w sensie naukowym nie oddaje w pełni tego, co pragną przekazać opisujący go poeci. Świat może być fizykalny, nawet jeżeli pewne jego aspekty umykają

wyczerpującemu opisowi w teorii fizycznej. Po drugie, również twierdzenie, że nieredukcjonizm w filozofii umysłu wymaga odrzucenia naturalizmu czy materializmu, jest nieuzasadnione, co od początku podkreśla Tom Nagel. Nieredukcjonizm perspektywy poznawczej nie prowadzi do odrzucenia teorii materialistycznej (teorii, że wszystko jest materialne), ale tylko wąskiego fizykalizmu (teorii, że wszystko jest opisywalne w języku zasadniczo weryfikacjonistycznej metodologii nauk przyrodniczych, jaka panowała w połowie XX wieku).

b) „Wymóżdżeńcy" to argument, że można sobie wyobrazić, iż istoty takie jak my (czy nawet niektórzy z nas) nie mają świadomości pierwszoosobowej. Można zatem wyobrazić sobie osobę taką jak czytelnik tego artykułu, która nie ma żadnych wewnętrznych przedstawień. Skoro można sobie to wyobrazić, to w jakimś sensie taka różnica istnieje. Ponieważ osoby takie czymś się różnią, zatem pierwszoosobowa świadomość jest zagadnieniem niepustym, a co więcej – każdy z nas ją posiada. Głównym oponentem tego argumentu jest Dan Dennett, który przypomina, że zasadniczym, definicyjnym twierdzeniem materializmu jest twierdzenie, że „nie ma różnicy bez różnicy fizycznej" (*there is no difference without physical difference*). Zatem argument ten *begs the question* przeciwko zwolennikom materializmu czy fizykalizmu. Według mnie Dennett ma tutaj rację, gdyż założenie, iż możliwe są w każdym detalu identyczne osoby, z których jedna ma świadomość pierwszoosobową, a inna nie ma, przyjmuje bez dowodu, że możliwa jest różnica bez różnicy podstaw fizycznych, a to kwestionuje (bez dowodu) słuszność stanowiska, przeciw któremu ten argument ma zostać wykorzystany. Łatwym sposobem zadośćuczynienia zarzutom Dennetta byłoby przyjęcie, że można zdefiniować wymóżdżeńców w sposób niepełny (to moja koncepcja *rough zombies*). Można założyć, że normalny człowiek i wymóżdżeniec są tacy sami na pewnym poziomie granularności (np. na poziomie opisu zewnętrznego), ale przy bliższym badaniu ich mózgi pracują zapewne inaczej. Przedyskutowałem to podejście z Chalmersem w trakcie sympozjum w Arizonie wiosną 2012 roku, ale Chalmers stanowczo się mu opiera, ponieważ wówczas wymóżdżeńcy stają się tylko ogólnym sposobem uzmysłowienia czy wyobrażenia sobie tego, co przynosi perspektywa pierwszoosobowa, a nie stanowią argumentu.

c) „Chiński pokój" Searle'a to historia, w której mamy wyobrazić sobie, że w pewnym pokoju siedzi wiele osób, które dostają informacje po chińsku i postępując wedle ściśle określonych procedur wysyłają odpowiedzi po chińsku, nie mają jednak pojęcia o treści przychodzących czy wychodzących informacji. W sensie funkcjonalnym należy powiedzieć, że „pokój" zna język chiński, zachowuje się bowiem jak kompetentny użytkownik tego języka. Ale w rzeczywistości nikt w danym pokoju nie rozumie chińskiego.

Ten argument ma pokazać dwie rzeczy – po pierwsze, że komputery, które zachowują się jak osoby w tym pokoju, nie rozumieją nic, nawet jeśli funkcjonują tak, jakby rozumiały. Po drugie, że rozumienie to coś więcej niż funkcjonalność. Searle twierdzi, że komputerom i osobom w chińskim pokoju brakuje intencjonalności. W sensie funkcjonalnym możliwa jest odpowiedź na argument Searle'a, że faktycznie osoby w chińskim pokoju nie posiadają znajomości języka chińskiego, bo ich wiedza ogranicza się do jednego kontekstu – syntaktycznego (postępowania zgodnie z danymi regułami) – a wiedza wymaga w sensie semiotycznym także poziomu semantycznego i pragmatycznego (tj. stosunku do tego, o czym jest mowa). Searle jednak twierdzi, że i to nie wystarcza, ponieważ intencjonalność to coś więcej.

Sądzę, że zarówno w przypadku Searla owo „coś więcej", czego nie wyczerpują zaawansowane mechanizmy funkcjonalne, to nieredukowalna perspektywa pierwszoosobowa. Podobnie ma się sprawa z biało-czarną Marią i wymóżdżeńcami Chalmersa: we wszystkich wypadkach argumenty funkcjonalne mogą w pewnym stopniu naprowadzić intuicje, ale tylko Nagel zdaje się w pełni rozumieć, zwłaszcza w książce *The View from Nowhere*, iż podmiotowość pierwszoosobowa nie poddaje się opisowi funkcjonalnemu. Stanowi ona bowiem podmiotowy warunek wszelkiego poznania; twierdzenie to, oparte po części na podejściu Nagla, a po części Husserla, a nawet Fichtego, wyjaśniłem w języku polskim (w pracy Bołtuć 2008) i w angielskim (Bołtuć 2009). W dalszej części niniejszego artykułu nie bronię tezy o nieredukcjonizmie świadomości, pytam raczej w trybie warunkowym, czy jeżeli przyjmiemy nieredukcjonizm, to maszyny mogą posiadać niereduktywną świadomość.

4. Teza inżynieryjna o teoretycznej możliwości pierwszoosobowej jaźni maszyn

Argument ma następujący charakter:

1) Traktuję świadomość – w tym nieredukowalną świadomość dostępną tylko z perspektywy pierwszej osoby – jako proces naturalny.
2) Jeżeli świadomość jest procesem naturalnym, to możliwe jest poznanie zasad jej generowania (tj. mechanizmu, jaki ją generuje).
3) Jeżeli poznamy w szczegółach sposób generowania świadomości pierwszoosobowej, to będzie to zapewne jakiś algorytm inżynierski (zapewne nie jest to program komputerowy, który można urzeczywistnić na maszynie obliczeniowej, ale raczej program drugiego stopnia, jakiego można użyć do zaprogramowania robota – np. w systemie AutoCad – żeby zbudował urządzenie generujące świadomość).

4) Tego rodzaju algorytmy są zwykle możliwe do zrealizowania w różnych substancjach. Nie ma więc powodu, aby tego rodzaju świadomości nie zbudować jako części robota.
5) Kwestia weryfikowalności, czy dana maszyna miałaby świadomość pierwszoosobową, ma zawsze charakter ogólnej dedukcji, czyli rozumowania zawodnego. Nie jest to jednak nic specjalnego, bo także problem „innych umysłów" ludzkich rozwiązywalny jest tylko indukcyjnie (np. na zasadzie *inference to the best explanation*).

Omówię pokrótce te punkty: W punkcie 1 przyjmuję założenie naturalizmu. Nie będę go szczegółowo uzasadniał, ale ogólnie mówiąc, dzieje rozwoju nauki wskazują na to, że zawsze rozwiązania nienaturalistyczne (takie jak *vis vitalis*) w miarę rozwoju nauki udaje się sformułować bardziej precyzyjnie w języku nauk przyrodniczych. Z jakiegoś powodu wielu przeciwników naturalizmu traktuje świadomość fenomenalną jako ostatni bastion nienaturalizmu w epistemologii, a co więcej, wielu naturalistów (nawet tak wybitnych jak Dennett) zdaje się to założenie podzielać. Jest to jednak założenie bezpodstawne, jakieś skojarzenie z pojęciem duszy. Otóż nie ma powodu, żeby perspektywa pierwszoosobowa na świat wymagała skojarzeń teologicznych; w tej kwestii Nagel wydaje się mieć najtrzeźwiejsze podejście.

W punkcie 2 opieram się także na prawidłowościach rozwoju nauki. Świat jest poznawalny i dotychczas wszystkie kwestie, które zdawały się onegdaj „niepoznawalne z zasady", okazują się jak najbardziej poznawalnymi, i to zwykle na wielu różnych poziomach wyjaśniania.

W punkcie 3 podałem w nawiasie wyjaśnienie, które dla porządku powinno zostać umieszczone dopiero w tym miejscu, ale zrobiłem to celowo. Mianowicie dlatego, że od dawna wielu programistów i teoretyków sztucznej świadomości twierdzi uparcie, że świadomość to jakiś sposób przetwarzania informacji. Jest to pogląd zrozumiały, gdy mówimy o świadomości jako o sposobie myślenia, np. myśleniu na wysokim poziomie złożoności. Przykładowo, zaspokojenie wymogów testu Turinga rzeczywiście wymaga li tylko napisania dobrego programu. Natomiast świadomość pierwszoosobowa to jest proces naturalny, emergentny, podobniejszy do funkcjonowania hologramu. Trudno powiedzieć, że maszyna tworząca hologram jest tylko programem, natomiast – jak wszystko w naturze – podlega ona opisowi za pomocą funkcji matematycznych, zatem może być zbudowana. Komputer może opisać „funkcję inżynierską", wedle której generator świadomości mógłby zostać zbudowany przez mniej lub bardziej skomplikowanego robota. To ostatnie sformułowanie jest nietrywialne, może się bowiem okazać, że świadomość pierwszoosobowa może powstać tylko w systemach opartych o chemię organiczną, czy że polega na trudnych do wywołania efektach kwantowych, co przesuwa naszą dyskusję do pkt 4. Nie ma powodu sądzić, że tego rodzaju struktury nie są możliwe do

skonstruowania, ale niekiedy może to być proces skomplikowany, a w przypadku struktur organicznych nie zawsze musi istnieć jasna granica pomiędzy bioinżynierią a hodowlą tkanek czy narządów. Ale są to zagadnienia szczegółowe, które pozostawiam do dalszej dyskusji.

Warto zauważyć, iż naukowe poznanie świadomości jest dopiero w powijakach. Dopiero od kilku lat istnieją naukowe teorie świadomości, z których jedna (albo grupa teorii wyjaśniających jej różne aspekty) może przynajmniej dostarczyć kierunku naukowego wyjaśnienia świadomości. Warto tutaj wymienić teorię globalnej przestrzeni (*global workspace*), która wyjaśnia rolę czy funkcję, jaką spełnia świadomość. Jeśli chodzi o teorie dotyczące bezpośrednio świadomości pierwszoosobowej, to należy do nich szereg teorii, które upatrują centrum świadomości w podwzgórzu; inne wyjaśniają świadomość jako przesunięcie w fazie funkcjonowania hemisfer mózgu; Hameroff twierdzi, że specjalną rolę spełniają efekty kwantowe zachodzące w mikrotubulach. Nie chodzi o to, byśmy na obecnym etapie rozwoju nauki mogli określić, która z tych teorii jest zgodna z prawdą, ale o to, że naukowe teorie zmierzające do wyjaśnienia zasad powstawania świadomości pierwszoosobowej istnieją. Nie jest to zatem problem naukowo pusty.

Dochodzimy teraz do punktu 5, najbardziej filozoficznego. Jak pokazał Kartezjusz, w sensie absolutnym nie wiemy o świecie właściwie niczego, ale oczywiście tradycja empiryzmu europejskiego okazała się o wiele bardziej produktywna. Jakkolwiek nie wiemy nic w sposób całkiem pewny, to wiele rzeczy wiemy w sposób indukcyjnie zadowalający (np. na zasadzie *induction to the best explanation*). Na tej zasadzie mamy dobre powody, żeby sądzić, iż inni ludzie, a także inne zwierzęta, są świadomi w sensie pierwszoosobowym. Badania FMRI potwierdzają (indukcyjnie) te intuicje. Jeżeli będziemy wiedzieli, na jakim konkretnie mechanizmie opiera się funkcjonowanie świadomości u ludzi i zwierząt, to będziemy zapewne w stanie określić satysfakcjonujące wyznaczniki funkcjonowania takiego mechanizmu. Jak wszelka wiedza empiryczna, będzie to z pewnością wiedza zawodna, ale nie ma powodu sądzić, że będzie to szczególny problem tego rodzaju maszyn.

5. Podsumowanie

Program BICA pozwala na potencjalne stworzenie robotów wyposażonych w pierwszoosobową świadomość – jeżeli spełnione są następujące warunki: 1) świadomość pierwszoosobowa jest procesem naturalnym, 2) wszystkie procesy naturalne są poznawalne, 3) poznawalność procesów naturalnych sprowadza się do możliwości podania algorytmu ich funkcjonowania, 4) jeżeli wiemy dokładnie, jak dany proces naturalny funkcjonuje, to w zasadzie możemy go

odtworzyć. A zatem możliwe jest stworzenie robota posiadającego pierwszoosobową świadomość.

Powodzenie takiego zadania nie jest bezpośrednio weryfikowalne, co wynika z „problemu uprzywilejowanego dostępu". Na zakończenie wspomnę o zagadnieniu, które pojawia się w dyskusjach w Polsce, a które bardziej szczegółowo omawiam gdzie indziej (Bołtuć 2010). Chodzi o kwestię tego, czy świadomość pierwszoosobowa posiada wartość funkcjonalną, w przeciwnym razie byłaby ona bowiem w pełni epifenomenalna. Jeśli pierwszoosobowa świadomość ułatwia kodowanie różnorakich danych (używając koloru, dźwięku, zapachu itp.), to możliwe jest zapewne zweryfikowanie istnienia takiej świadomości za pomocą jej cech funkcjonalnych. Pojawia się wątpliwość, czy poziom fenomenalnego postrzegania jakości postrzeżeniowych posiada znaczenie funkcjonalne, skoro jakości te są epifenomenem nerwowych odpowiedników świadomości (*neural correlates of consciousness*). Jednak pewne poziomy opisu pozwalają na odczytanie niesionych przez nie znaczeń, zaś inne znaczenia te jedynie przenoszą. Na przykład tekst składa się z liter, ale ma sens dopiero, gdy litery są czytane na poziomie słów lub zdań; badanie sekwencji liter nie pozwoli zwykle na wygenerowanie tekstu. Także obraz nie z każdej perspektywy pozwala na ujęcie *gestalt*-u. Podobnie jest z poziomem fenomenalnym, dla którego *korelaty nerwowe* nie stanowią nośników znaczenia, a tylko podłoże procesu emergentnego. Czysta świadomość jest częścią układu podmiotowego, który może być analizowany jako zespół podmiotowo-przedmiotowy różniący się od innych przedmiotów występowaniem cechy podmiotowości przejawiającej się w jego funkcjonowaniu (Bołtuć 2007). Wymóżdżeńcy też mogą się tak zachowywać, zachowanie jest bowiem jedynie pewnym algorytmem, ale to podmiotowość pomaga wybrać z nieskończonej liczby możliwości te, które podmiotowi zdają się słuszne i są zgodne ze specyfiką jego budowy. Oczywiście ta specyfika budowy też postępuje zgodnie z algorytmem, co umożliwia tworzenie metaprogramów takich jak *creativity machines* i maszyny będące funkcjonalnymi odpowiednikami emocji.

Nie sądzę, żeby istniała jakaś funkcja, która jest specyficznie podmiotowa i nie jest replikowalna przez sztuczną inteligencję. Byłoby to sprzeczne z tezą Churcha-Turinga, aczkolwiek istnieją różne interpretacje tej tezy. Program BICA zakłada, że istnieje wiele funkcjonalności kognitywnych organizmów biologicznych, od których możemy się uczyć konstruując maszyny myślące, nie zakłada on natomiast jakiejś zasadniczej funkcji, której maszyny nie mogłyby zreplikować. W artykule tym argumentuję na rzecz tezy, że nie zawsze chodzi tu o *software*; w przypadku funkcji świadomości pierwszoosobowej jej zreplikowanie wymagałoby zapewne na poziomie *hardware* zbudowania generatora świadomości pierwszoosobowej. Zadanie to jest niezależne od odpowiedzi na pytanie, czy taki generator przynosiłby jakiś pożytek praktyczny, w sensie

dostarczania nowych funkcji kognitywnych (aczkolwiek miałoby to na pewno znaczenie w sensie motywowania takiego projektu). Zapewne byłby on jednym ze sposobów na uzyskiwanie funkcji poznawczych właściwych bytom świadomym. Z pewnością złożony program może dostarczyć takich funkcji prościej. Znaczenie podmiotu pierwszoosobowego jako „lustra świata", jak określali to filozofowie renesansowi, ma raczej charakter aksjologiczny – bez podmiotu trudno jest mówić o znaczeniu czegokolwiek; ale to już temat na inny artykuł.

Bibliografia

Block N. (1995), *On a Confusion about a Function of Consciousness*, „Brain and Behavioral Sciences" 18, 2, s. 227–247.

Block N (2002), *Searle's Argument Against Cognitive Science*, w: J. Preston, M. Bishop (eds.), *Views into the Chinese Room: New Essays on Searle and Artificial Intelligence*, Oxford University Press, s. 70–79.

Bołtuć P. (1998), *Reductionism and Qualia*, „Epistemologia" 4, s. 111–130.

Bołtuć P. (2007), *Czy subiektywność epistemiczna posiada znaczenie moralne?*, „Analiza i Egzystencja" 6, s. 5–23.

Bołtuć P. (2009), *The Philosophical Problem in Machine Consciousness*, „International Journal of Machine Consciousness" 1, s. 155–176.

Bołtuć P. (2010), *A Philosopher's Take on Machine Consciousness*, w: V.E. Guliciuc (ed.), *Philosophy of Engineering and the Artifact in the Digital Age*, Cambridge Scholar's Press, s. 49–66.

Bołtuć P. (2012), *The Engineering Thesis in Machine Consciousness*, „Techne" 2, s. 187–207.

Chalmers D. (2003), *Consciousness and Its Place in Nature*, w: S. Stich, F. Warfield (eds.), *Blackwell Guide to Philosophy of Mind*, Blackwell.

Crick F. (1984), *Function of the Thalamic Reticular Complex: The Searchlight Hypothesis*, „Proceedings of the National Academy of Sciences, 81, 14, s. 4586–4590.

Dennett D. (1994), *The Practical Requirements for Making a Conscious Robot*, „Philosophical Transactions of the Royal Society", A 349, s. 133–146.

Franklin S., Baars B.J., Ramamurthy U. (2008), *A Phenomenally Conscious Robot?*, „APA Newsletter on Philosophy and Computers" 1.

Ikegami T. (2005), *Chaotic Itinerancy, Active Perception and Mental Imagery*, „Proceedings of the Symposium on Next Generation Approaches to Machine Consciousness", April, University of Hartfordshire, s. 36–39.

Nagel T. (1986), *The View from Nowhere*, Oxford University Press.

Searle J. (1980), *Minds, Brains, and Programs*, „Behavioral and Brain Sciences" 3, 3, s. 417–457.
Turing A. (1950), *Computing Machinery and Intelligence*, „Mind" LIX, 236, s. 433–460.

Streszczenie

Biologiczne Zainspirowane Architektury Kognitywne (BICA) to projekt zmierzający do zbudowania architektury kognitywnej osadzonej psychologicznie i neurobiologicznie na architekturze ludzkiego mózgu. Zawiera ona takie procesy funkcjonalne, jak emocje i pętle wspomnieniowo-asocjacyjne. Uważam, że jedną z niestandardowych biologicznie zainspirowanych architektur, która wykracza poza poziom przetwarzania informacji, jest stworzenie perspektywy pierwszoosobowej (jaźni). Skoro wysoko rozwinięte zwierzęta mają jaźń i skoro mózg jest przedmiotem naturalnym, to należy się spodziewać, że kiedyś zrozumiemy jego funkcjonowanie także w kwestii tworzenia jaźni pierwszoosobowej. Oznaczałoby to teoretyczne poznanie mechanizmu pozwalającego na zaprojektowanie maszyny posiadającej świadomość pierwszoosobową.