

Tytuł: **Intencjonalność i komputery. Krytyka J. Searle'a mocnej wersji AI**

Autor: Piotr Kołodziejczyk / pkolodziejczyk@interia.pl

Źródło: <http://www.kognitywistyka.net> / mjkasperski@kognitywistyka.net

Data publikacji: 29 VI 2002

1. Uwagi wstępne

Wraz z rozwojem technologii informatycznych, w filozofii umysłu zaobserwować można tendencję do utożsamiania pojęcia *umysł* z zaimplementowanym w mózgu „programem”, pod względem działania analogicznym do programów komputerowych. Pisze w tej kwestii Marek Hetmański:

Umysł człowieka jest wyobrażany jako program działający wobec danych (znaków zakodowanych cyfrowo) pobieranych z otoczenia. Jego istota nie zależy od natury (tworzywa, złożoności) substancji, w którą się wciela.¹

Stąd też, wnioskuje się, że charakterystyka operacji poznawczych podmiotu (np. postrzegania, wnioskowania, rozumienia) jest sprowadzalna do problemów obliczeniowych, czyli algorytmizacji, rozstrzygalności i dowodzenia w systemach formalnych. W takim ujęciu, stany umysłu są sekwencjami obliczeniowymi (jednostkami syntaktycznymi), których koniunkcja stanowi koherentny model formalny zawierający aksjomaty i reguły transformacji tych sekwencji².

W literaturze przedmiotu, twierdzenie o identyczności ludzkiego umysłu i programu komputerowego traktuje się zazwyczaj jako fundament obliczeniowej teorii umysłu (funkcjonalizmu komputerowego). Sam zaś termin *obliczeniowa teoria umysłu* odnosi się, pisze M. Hetmański,

do wielu koncepcji powstałych i rozwijanych w ramach nurtu kognitywistycznego. Pojawiły się one na pograniczu wspólnych badań nauk jak logika, informatyka, nauka o komputerach, nauki o poznawaniu, teorie i badania nad sztuczną inteligencją, także lingwistyka i neuronauk.³

Kognitywizm jest więc programem badawczym mającym na celu zarówno konstrukcję sztucznych systemów rozwiązujących ogólne problemy, jak i próbę odpowiedzi na pytania *stricte* epistemologiczne: o źródła wiedzy poznania czy naturę umysłu. Zadanie pierwsze

¹ M. Hetmański, *Umysł a maszyny. Krytyka obliczeniowej teorii umysłu*, Lublin 2000, s. 42.

² Por. tamże, s. 74.

³ Tamże, s. 43. Por. także, P. Coveney, R. Highfield, *Granice złożoności. Poszukiwania porządku w chaotycznym świecie*, tłum. P. Amsterdamski, Warszawa 1997, ss. 169-170.

przypisuje się z reguły badaniom nad sztuczną inteligencją, drugie – naukom o poznawaniu⁴. Podział ten nie oznacza jednak, że zakres badań Sztucznej Inteligencji i nauk o poznawaniu nie posiada żadnych punktów wspólnych. Wprost przeciwnie, rozstrzygnięcia obydwu dyscyplin bazują na wspólnych założeniach teoretycznych. Założenia te określa się mianem *funkcjonalizmu*.

1. Główne założenia stanowiska funkcjonalistycznego

Funkcjonalizm jest interpretacją klasycznego problemu umysłu i ciała. Podstawowym jego założeniem jest twierdzenie głoszące, że stany umysłu (np. rozumowania, przekonania) są stanami fizycznymi będącymi własnością dowolnego układu fizycznego. Inaczej mówiąc,

dany układ (system) pozostaje w określonym stanie poznawczym wtedy i tylko wtedy, gdy znajduje się w pewnym stanie fizycznym interpretowanym funkcjonalnie (tzn. nie jest on tożsamy wyłącznie ze stanem fizycznym, lecz z jego funkcją); tym szczególnym stanem fizycznym układu jest przetwarzanie informacji.⁵

Termin *funkcjonalizm* do filozofii umysłu wprowadził Hilary Putnam w artykule *Minds and Machines* z roku 1960, w którym, zwracając uwagę na podobieństwo uniwersalnej maszyny Turinga i ludzkiego umysłu, starał się wykazać, że maszyna Turinga może wykonywać takie same operacje myślowe jak poznający i działający podmiot ludzki. Natomiast zasadniczy wykład stanowiska funkcjonalistycznego, (które stało się, jak utrzymuje sam Putnam, ortodoksją we współczesnej filozofii umysłu⁶) przedstawiony został w pracy *Representation and Reality* z 1988 roku. Odwołując się do analizy logiczno-informatycznej, Putnam w pracy tej usiłował wykazać, iż pomimo różnic *hardware’owych* maszyna i człowiek mogą zrealizować ten sam *software* (ciąg instrukcji zawartych w programie) na podstawie danego zbioru reguł obliczeniowych. Nie oznacza to jednak, że ludzkie stany mentalne są stanami obliczeniowymi. Są one jedynie opisywalne za pomocą reguł obliczeniowych. Zatem, w ujęciu Putnama, stany mentalne składające się na umysł ludzki nie są ani programami, ani skończonym zbiorem zawierającym reguły przetwarzania symboli ze względu na fakt, iż stanem mentalnym autor *Representation...* przypisuje własność *obliczeniowej plastyczności*. Znaczący to, że

fizycznie możliwe istoty, które sądzą, że w okolicy jest mnóstwo kotów, czy czegokolwiek innego, mogą mieć niezliczone, najrozmaitsze *programy*.⁷

Zgodnie z powyższym ujęciem wydaje się, że funkcjonalizm należy interpretować jako pogląd mówiący, iż dany układ fizyczny za sprawą różnorodności swego działania determinuje wielość i różnorodność swoich funkcjonalnych realizacji⁸. Procedury obliczeniowe są zaś wyrażeniem aktualnej realizacji danego stanu mentalnego. Takie postawienie sprawy sugeruje, że twierdzenie o identyczności działania ludzkiego umysłu i programu komputerowego jest wysoce problematyczne. Problematyczność ta wynika z niemożności przyporządkowania danemu stanowi mentalnemu jednego tylko opisującego go formalizmu. W przypadku stanowiska Putnama mówić więc można o dwóch rodzajach

⁴ Zob. M. Gardner, *The Mind's New Science. A History of the Cognitive Revolution*, New York 1985, s. 9.

⁵ M. Hetmański, *Umysł...*, s. 69.

⁶ Zob. H. Putnam, *Representation and Reality*, Cambridge Mass. 1988, s. XI.

⁷ Tamże, s. 139.

⁸ Zob. M. Hetmański, *Umysł...*, s. 73.

funkcjonalizmu: skrajnym (zaprezentowanym w *Minds and Machines*) oraz umiarkowanym (z okresu *Representation and Reality*). Godnym podkreślenia jest w tym miejscu fakt, iż główna teza funkcjonalizmu skrajnego (głosząca, że jeśli uniwersalna maszyna Turinga zostałaby wyposażona w organy zmysłowe i możliwość wypowiedzania się o swoich stanach wewnętrznych, to jej operacje poznawcze w niewielkim stopniu różniłyby się od możliwości człowieka⁹) stała się ideą przewodnią we wczesnych badaniach nad sztuczną inteligencją. Do poglądów Putnama nawiązywali bowiem tak prominentni badacze AI jak: Marvin Minsky, Alan Newell, John Holland czy Herbert Simon¹⁰. Na podstawie stwierdzeń Putnama wysnuto wniosek, że do zrozumienia działania ludzkiego umysłu konieczne jest użycie komputera jako modelu funkcjonowania samego umysłu. Wniosek ten z kolei zezwolił na sformułowanie operacyjnej definicji inteligencji¹¹.

W badaniach nad sztuczną inteligencją, inteligencję pojmuje się zazwyczaj jako ogólną własność wszystkich systemów poznawczych, które w podobnych sytuacjach rozwiązywania danych problemów wykazują powtarzalne właściwości. **W przypadku człowieka za działanie inteligentne uznaje się działanie oparte o pewien zbiór reguł prowadzących do rozwiązania zadanego problemu, natomiast w przypadku komputerów inteligencja polega na wykonaniu określonego programu.** Zatem, działanie programu komputerowego traktuje się jako inteligentne wtedy, gdy jest ono podobne do działania człowieka, które uznaje się za inteligentne. Relację pomiędzy ludzką i maszynową inteligencją dobitnie wyraził M. Minsky pisząc:

Wobec tego „czym właściwie jest” inteligencja? Z mojego punktu widzenia jest to raczej zagadnienie estetyki albo szacunku dla nauki. Dla mnie „inteligencja” oznacza niewiele więcej niż kompleks działań, z którymi mamy do czynienia, ale którego nie rozumiemy (...). Ale nasza niezdolność do wskazania siedziby inteligencji nie powinna nas prowadzić do wniosku, że wobec tego maszyny programowalne nie mogą myśleć, gdyż jeśli dla człowieka, tak jak dla maszyny rozumiemy wreszcie strukturę i program, to uczucie tajemniczości (i samouwielbienia) zniknie.¹²

W świetle powyższej wypowiedzi jest widocznym, że inteligencja nie jest rozumiana substancjalnie, lecz funkcjonalnie. Funkcjonalizm ten wynika z akceptacji stwierdzenia głoszącego, iż jeśli dany układ poznawczy działa zgodnie z regułami, na podstawie których człowiekowi przypisuje się zachowanie inteligentne, to układowi temu należy również przypisać własność inteligencji (niezależnie od analizy elementów, z których jest on zbudowany). Na konieczność asubstancjalnego definiowania terminu *inteligencja* wyraźnie wskazywał także inny klasyk AI, H. Simon pisząc:

Zrozumieliśmy, że inteligencja nie jest sprawą substancji – protoplazmy, szkła czy drutu – lecz formy, którą substancja przyjmuje i procesów, które w niej zachodzą. Korzeniami inteligencji są symbole z ich denotacyjną siłą i podatnością na manipulację. Symbole mogą zaś być wytwarzane prawie ze wszystkiego, co może zostać zgromadzone, połączone i zorganizowane. Inteligencją jest umysł implementowany w każdy wymodelowany rodzaj materii.¹³

⁹ Por. tamże, s. 71.

¹⁰ Por. P. Coveney, R. Highfield, *Granice...*, ss. 170-173.

¹¹ Zob. H. Gardner, *The Mind's...*, s. 9.

¹² M. Minsky, *Na drodze do stworzenia sztucznej inteligencji*, tłum. D. Gajkowicz, ss. 378-424, w: *Maszyny matematyczne i myślenie*, red. E. Feigenbaum, J. Feldman, Warszawa 1972, s. 420.

¹³ Cyt. za: M. Hetmański, *Umysł...*, s. 52.

Stąd też, jeżeli inteligencja przysługuje zarówno komputerom, jak i człowiekowi, to o jej istocie nie decyduje samo działanie, ale efekt działania. Ważniejszym jest bowiem, podkreślają teoretycy AI, samo rozwiązanie problemu, niż materiał (struktura fizyczna) realizująca ten proces.

Warto podkreślić, że ze względu na wieloznaczność terminu *inteligencja* w obrębie badań nad AI wyróżnia się jej mocną i słabą wersję. Rozróżnienie to zostało wprowadzone przez Johna Searle'a w pracy *Minds, Brains and Science*. Teoria mocnej wersji Sztucznej Inteligencji (silne AI) zakłada, że

umysł jest tym da mózgu, czym program dla komputera.¹⁴

Natomiast wersja słaba (słabe AI) głosi, że „inteligentne” programy komputerowe stanowią środki testujące koncepcje opisujące inteligentne działanie i zachowanie człowieka¹⁵.

Mocna wersja sztucznej inteligencji jest reprezentowana m.in. przez M. Minsky'ego, J. Hollanda, J. McCarthy'ego i D. Hofstadtera. Przedstawiciele tego stanowiska twierdzą, że stany funkcjonalne (obliczanie, przetwarzanie informacji, manipulowanie symbolami) mają taką samą naturę niezależnie od budowy układów, które je realizują. W ujęciu tym, abstrahuje się więc zarówno od genezy umysłu i programu komputerowego. Akcentuje się zaś wyłącznie funkcjonalne powiązania – odpowiednio z mózgiem i sprzętem komputerowym. W wersji tej przyjmuje się również, że możliwe jest bardzo dokładne modelowanie przez program komputerowy większości inteligentnych procesów poznawczych człowieka. Modelowanie to odbywa się za pomocą implementacji modeli procesów poznawczych w postaci programów wykonywanych przez komputery o ogromnej mocy obliczeniowej i pamięci operacyjnej. Zdaniem M. Hetmańskiego, z konkluzji tej

wyciąga się następnie wniosek w postaci ogólnej i radykalnej tezy (stanowiącej właśnie istotę silnej wersji Sztucznej Inteligencji), iż komputery mają stany umysłu w takim samym sensie, w jakim ma je człowiek.¹⁶

Z kolei do zwolenników słabej wersji AI należy większość nowszego pokolenia badaczy Sztucznej Inteligencji. Przedstawiciele tego nurtu unikają zazwyczaj formułowania ontologicznych tez o równoważności programów komputerowych i ludzkiego umysłu. Formułują oni raczej tezy o metodologicznej stosowalności programów komputerowych w analizie ludzkich czynności poznawczych posiadających znamiona zachowania inteligentnego. Żaden program – twierdzą – nie zapewni odtworzenia bogatej struktury ludzkiego umysłu. W tym ujęciu, użycie komputerów ma jednak sens, ponieważ stanowi ono użyteczne narzędzie poznawcze dla neurofizjologów, psychologów i innych teoretyków badających umysł.

Moim zdaniem – pisze J. Searle – słabe AI, użycie komputerów do modelowania lub symulowania procesów mentalnych wyszło z tej debaty bez szwanku, nie tylko istnieje, ale nawet w swoim konekcyjnym wcieleniu – kwitnie.¹⁷

¹⁴ J. Searle, *Umysł, mózg i nauka*, tłum. J. Bobryk, Warszawa 1994, s. 25.

¹⁵ Zob. tamże, s. 27.

¹⁶ M. Hetmański, *Umysł...*, s. 53.

¹⁷ J. Searle, *Umysł na nowo odkryty*, tłum. T. Baszniak, Warszawa 1999, s. 145.

Warto dodać, że zarówno w mocnej, jak i słabej wersji Sztucznej Inteligencji podejmuje się problematykę, która z punktu widzenia tego artykułu wydaje się być najbardziej istotna. Idzie mianowicie o problematykę wyznaczaną zagadnieniem intencjonalności. Począwszy bowiem od wczesnych lat osiemdziesiątych XX w. badacze AI podejmują próbę powiązania teorii intencjonalności ze sposobem opisu procesów poznawczych charakterystycznych dla prac syntetyzujących refleksję psychologiczną z teoriami Sztucznej Inteligencji¹⁸. Poniżej przedstawiam i analizuję niektóre rozstrzygnięcia oscylujące wokół kwestii semantycznych w badaniach nad AI.

2. Problematyka intencjonalności w mocnej wersji AI

W początkach badań nad sztuczną inteligencją problematyka intencjonalności w ogóle nie pojawiała się w rozważaniach teoretyków AI. Zdawać się może, że sytuacja ta podyktowana była akceptacją tezy o symbolicznej mediatyzacji ludzkich procesów poznawczych. Badania nad sztuczną inteligencją stanowią bowiem dyscyplinę, która opowiada się za rozstrzygnięciami semiotycznymi Ch. Peirce'a, w szczególności zaś – za twierdzeniem, że ludzka pamięć semantyczna jest systemem wzajemnie interpretujących się elementów.

Wybór ten – pisze Jerzy Bobryk – zdaje się nie być podyktowany ani wyrażoną wprost znajomością prac Peirce'a, ani też świadomym odrzuceniem Husserlowskiej kategorii intencjonalności. Założenie o upośrednionym charakterze poznania wynika raczej z popularności metafory komputerowej i ścisłych związków teorii psychologicznej z informatyką.¹⁹

Kontynuując ten wątek stwierdzić należy, że z twierdzenia o upośrednionym charakterze procesów poznawczych wynika arbitralne (moim zdaniem) ustalenie charakteru wewnętrznych reprezentacji i relacji między operacjami poznawczymi. Akceptacja twierdzeń wygłoszonych przez Peirce'a sprawia, że reprezentacje wewnętrzne określa się poprzez zbiór algorytmów, które testuje się poprzez zestawienie rezultatów zastosowania tych algorytmów z wynikami działań ludzkich. W początkach badań nad sztuczną inteligencją nie podejmowano jednak refleksji nad własnościami tych reprezentacji, ich przedmiotowym odniesieniem czy sposobem skierowania się systemu poznającego na obiekty wobec niego transcendentne.

Pierwszym krokiem ku asymilacji w ramy teorii AI zagadnień wyznaczanych pojęciem *intencjonalności* stały się analizy Zenona Pylyshyna. Wprawdzie w jego najbardziej znanej pracy – *Computation and Cognition*, pojawia się tylko termin *znaczenie* (nie mówi się zaś niczego o intencjonalności), to analiza pojęcia znaczenia *explicite* wyznacza refleksję nad zagadnieniem intencjonalności²⁰. Pylyshyn przyjmuje bowiem, że dany proces jest intencjonalny wtedy, gdy posiada on treść semantyczną. Treść semantyczna jest zaś wynikiem operacji obliczeniowych dokonywanych przez poznający system (umysł lub program komputerowy). Zdaniem Pylyshyna, podobnie jak program, tak i umysł jest swoistą „maszyną syntaktyczną”, której podstawowym zadaniem jest obliczanie, czyli przetwarzanie symboli. Ową „maszynę syntaktyczną” charakteryzują więc własności formalne. Jest godnym podkreślenia, iż według autora *Computation...* zarówno umysł, jak i program komputerowy posiada także własności semantyczne. Postawienie tej tezy wynika z założenia, że

¹⁸ Por. J. Bobryk, *Przyczynowość i intencjonalność*, Warszawa 1992, s. 63.

¹⁹ Tamże, s. 92.

²⁰ Por. tamże, s. 96.

podstawowy dla „maszyny syntaktycznej” proces – proces obliczeniowy – posiada swoje przedmiotowe odniesienie.

Proces obliczeniowy – pisze Pylyshyn – to taki, którego zachowanie traktuje się jako zależne od przedstawieniowej lub semantycznej treści jego stanów. Definicja ta wynika z faktu istnienia poziomu strukturalnego określanego jako „poziom symboliczny”. (...) Poziom ten posiada dwie istotne własności. Po pierwsze – formalna, syntaktyczna struktura poszczególnych zdarzeń (znaków) symbolicznych odnosi się do rzeczywistych fizycznych różnic w systemie; różnic wywołujących odpowiednie zachowanie się systemu. Po drugie – formalna struktura symboli ukazuje wszystkie ważne różnice semantyczne, wobec których (...) system reaguje w przypadku zastosowania pewnych semantycznie zinterpretowanych reguł formułujących nowe struktury symboliczne. Za pomocą takich środków (formalnych reguł przekształcania symboli wraz z ich semantyczną interpretacją, przyp. P. K.) uzyskuje się możliwość opisu zachowania systemu jako reagującego na treść swoich przedstawień – i to w sposób całkowicie zgodny z materializmem.²¹

W świetle przytoczonych stwierdzeń jest widocznym, że realizacja funkcji obliczeniowych systemu poznającego implikuje odniesienie się tego systemu ku swym fizycznym własnościom. Fakt ten, w opinii J. Bobryka, potwierdza, że w koncepcji Pylyshyna istnieją rozważania nad pojęciem intencjonalności. Założenie, iż zawarte w strukturze systemu treści semantyczne reprezentują fizyczne własności „maszyny syntaktycznej” implikuje, że treści te są skierowane na owe własności. Są one więc intencjonalne²².

Na podstawie rekonstrukcji koncepcji Pylyshyna można powiedzieć, że jej doniosłość polega na powiązaniu w teorii Sztucznej Inteligencji problematyki semantycznej z zagadnieniem intencjonalności. Nietrudno jednak zauważyć braki tego ujęcia. Niepełność omówionej teorii polega na arbitralnym ograniczeniu rozważań do zespołu zagadnień wyznaczanych przez analizy syntaktyczne. Dlatego też współcześnie koncepcję Pylyshyna traktuje w sposób historyczny. Trudno mówić o dyskusjach na jej temat²³.

Z odmienną sytuacją mamy do czynienia w przypadku rozstrzygnięć uzyskanych przez Daniela Dennetta. Może się wydawać, że zaproponowana przez niego korelacja twierdzeń o intencjonalności z badawczym programem AI jest charakterystyczna dla całego nurtu kognitywistycznego. Taktyka badawcza Dennetta jest bowiem oparta o wnioski płynące ze stanowiska funkcjonalistycznego.

Formułując swoją koncepcję, Dennett wychodzi od klasycznej definicji intencjonalności: **intencjonalność jest kierowaniem się aktów świadomości ku obiektom innym niż one same**. Inaczej mówiąc,

coś, co przejawia intencjonalność, zawiera reprezentację czegoś innego.²⁴

Jednakże, podobieństwo teorii Dennetta z koncepcjami klasycznymi kończy się właśnie na akceptacji przytoczonej definicji. Autor *The Intentional Stance* przyjmuje bowiem, że intencjonalność jest własnością nie tylko umysłu ludzkiego. Przysługuje ona również innym

²¹ Z. Pylyshyn, *Computation and Cognition. Toward a foundation for Cognitive Science*, Cambridge Mass. 1984, s. 74.

²² Por. J. Bobryk, *Przyczynowość...*, s. 96.

²³ Zob. tamże, s. 94-95.

²⁴ D. Dennett, *Natura umysłów*, tłum. W. Turopolski, Warszawa 1997, s. 49.

złożonym systemom przetwarzającym informacje (np. zwierzętom, komputerom). Systemy te nazywa *systemami intencjonalnymi*. W *Naturze umysłów* czytamy na ten temat:

Systemy intencjonalne to z definicji wszystkie te (i tylko te) byty, których zachowanie da się przewidzieć/wyjaśnić z nastawienia intencjonalnego. Samopowielające się makrocząsteczki, termostaty (...), nietoperze, ludzie, komputery szachowe – wszystko to są systemy intencjonalne.²⁵

Użyty w definicji systemów intencjonalnych termin *nastawienie intencjonalne* zezwala, zdaniem Dennetta, na traktowaniu zachowania danego systemu tak, jak gdyby był on podmiotem racjonalnym. Podmiotem, który „wybiera”, „pragnie” czy „jest przekonany”. Nastawienie intencjonalne jest więc

postawą czy perspektywą, którą rutynowo przyjmujemy wobec siebie, zatem przyjęcie tego nastawienia wydaje się zamierzonym antropomorfizowaniem.²⁶

Co więcej, akceptacja tezy o nastawieniu intencjonalnym jest, twierdzi Dennett, kluczem do rozwiązania podstawowych problemów filozofii umysłu, ponieważ umożliwia ona eksplikację sposobów i motywów działania, postępowania i zachowania danego systemu w abstrakcji od jego fizycznej struktury. Teza o nastawieniu intencjonalnym jest zatem tezą o wyraźnie funkcjonalistycznej proveniencji. Fundamentalnym założeniem Dennetta, dzielonym z wieloma teoretykami umysłu, jest przekonanie, że

o tym, czy coś jest umysłem (przekonaniem, bólem, lękiem), **nie decyduje to, z czego się składa, lecz to, co potrafi zrobić.**²⁷ (Wytłuszczenie P. K.).

W związku z tym, uprawnione wydaje się stwierdzenie, że o tym, czy system jest intencjonalny nie decyduje jego budowa fizyczna, lecz działanie i zachowanie analogiczne do ludzkiego działania i zachowania intencjonalnego. Takie ujęcie prowadzi z kolei do dyskredytacji wprowadzonego przez J. Searle’a rozróżnienia na *intencjonalność właściwą* i *jak-gdyby-intencjonalność*²⁸. Zdaniem Dennetta, Searle popełnia kardynalny błąd przypisując intencjonalność wewnętrzną wyłącznie ludzkim podmiotom. Opisując sytuację, w której program komputerowy cechuje się własnością rozumienia języka naturalnego, autor *Brainstorms* dowodzi, że nie jest możliwe jednoznaczne wykazanie, iż generowanie zdań sensownych (intencjonalność wtórna) jest pochodną wewnętrznych stanów intencjonalnych maszyny. Istnieje bowiem możliwość, że komputer działa intencjonalnie

na mocy „jedynie” pochodnej intencjonalności, wmontowanej w jego konstrukcję – najpierw przez jej konstruktorów, a następnie w miarę jak nabywałaby coraz więcej informacji o swoim otoczeniu, w procesie samokreacji.²⁹

Warto również zaznaczyć, że przypisywanie przez Searle’a wewnętrznej intencjonalności podmiotom ludzkim jest być może błędne. Intencjonalność wewnętrzna stanowi, zdaniem Dennetta własność biologicznie pierwotnych systemów (atomów, neuronów, tkanek). Dlatego też, zbudowany z mikroelementów podmiot ludzki charakteryzuje się intencjonalnością pochodną wobec intencjonalności podstawowych składników. Píše Dennett:

²⁵ Tamże, s. 47.

²⁶ Tamże, s. 40.

²⁷ Tamże, s. 83.

²⁸ W kwestii tego rozróżnienia zob. J. Searle, *Umysł na nowo...*, ss. 114-118.

²⁹ D. Dennett, *Natura...*, s. 69.

Pochodzimy od robotów i składamy się z robotów, i cała intencjonalność, jaką możemy się cieszyć, jest pochodna od bardziej fundamentalnej intencjonalności milionów elementarnych systemów intencjonalnych.³⁰

Omawiając koncepcję intencjonalności w ujęciu Dennetta należy podkreślić, że pojęcia systemu i nastawienia intencjonalnego implikują przede wszystkim rozstrzygnięcia metodologiczne. Zaproponowane przez autora *Natury umysłów* rozumienie tych terminów ma na celu wykazanie, że pojęcia te stosuje się do użytecznego opisu pewnej klasy zjawisk. Nie jest to, tak jak np. w przypadku teorii Franza Brentana, rozwiązanie ontologiczne, którego skutkiem jest podzielenie rzeczywistości na sferę obiektów fizycznych i sferę zjawisk psychicznych. Terminy te wprowadza się tylko w celu deskrypcji zachowania się pewnych układów w terminach mentalistycznych³¹.

Konkludując powyższe rozważania powiedzieć należy, że cechują się one niepełnością i arbitralnością. Stwierdzenie to jest wynikiem akceptacji krytyki mocnej wersji AI dokonanej przez J. Searle'a. W poniższym paragrafie rekonstruję sedno tej krytyki i podejmuję próbę wykazania jej zasadności.

3. Johna Searle'a krytyka mocnej wersji AI

Nauki kognitywne i ich najprężniej rozwijająca się dziedzina – badania nad sztuczną inteligencją są, zdaniem Searle'a, analizą procesów przetwarzania informacji w oparciu o formalne (syntaktyczne) reguły operowania symbolami.

Gdybyśmy chcieli – pisze Searle – streścić program badawczy kognitywizmu powiedzielibyśmy: myślenie jest procesem przetwarzania informacji, zaś przetwarzanie informacji jest niczym innym, jak manipulowaniem symbolami. Czynnikiem to komputery, dlatego najlepszym sposobem badania myślenia (używa się tu raczej słowa poznanie) jest badanie obliczeniowych programów manipulowania symbolami niezależnie od tego, czy zachodzą one w komputerach, czy w mózgu. Dlatego, zgodnie z tym poglądem, zadaniem nauk o poznawaniu jest opisywanie mózgu nie na poziomie komórek nerwowych, ani na poziomie świadomych stanów psychicznych, lecz na poziomie, na którym działa on jako system przetwarzania informacji.³²

Searle nadmienia również, że genezy programu kognitywistycznego należy doszukiwać się w klasycznej pracy Alana Turinga *Computing Machinery and Intelligence* z roku 1950, z której to wynikają dwa twierdzenia stanowiące (używając terminologii Imre Lakatosa) *twardy rdzeń* mocnej wersji AI. Pierwszym z nich jest twierdzenie (lub teza) Churcha-Turinga głoszące, że dla każdego algorytmu istnieje pewna maszyna Turinga (program komputerowy) mogąca wykonać ten algorytm. Twierdzenie drugie (twierdzenie Turinga) mówi, że istnieje uniwersalna maszyna Turinga, która jest zdolna do symulowania dowolnej innej maszyny Turinga. Istotę tego twierdzenia trafnie wyraził Andrew Hodges pisząc:

Wyrażenie „maszyna Turinga” jest analogiczne do wyrażenia „wydrukowana książka”, gdyż odnosi się do klasy potencjalnie wielu egzemplarzy. W obrębie tej klasy pewne maszyny Turinga są „uniwersalne”, te mianowicie, które dysponują dostatecznym

³⁰ Tamże, s. 70.

³¹ Por. J. Bobryk, *Przyczynowość...*, s. 96.

³² J. Searle, *Umysł, mózg...*, s. 39.

stopniem złożoności, aby zinterpretować i wykonać tablicę zachowań dowolnej innej maszyny Turinga.³³

Wydaje się zatem, iż koniunkcja dwu powyższych twierdzeń pozwala zwolennikom mocnej wersji AI przyjąć, że wszystkie ludzkie procesy poznawcze mogą być implementowane w programach komputerowych za pomocą pewnego zbioru procedur algorytmicznych. Zgodnie bowiem z opisem Turinga, zauważa Searle,

zarówno ja, ludzki komputer, jak i komputer mechaniczny, realizujemy ten sam algorytm. Ja realizuję go świadomie, a komputer mechaniczny nieświadomie. Otóż wydaje się, że istnieją podstawy, by przypuszczać, iż w moim mózgu może nieświadomie zachodzić bardzo wiele innych procesów mentalnych, które również mają charakter obliczeniowy. Jeśli tak jest w istocie, to możemy ustalić, w jaki sposób działa mózg, przeprowadzając symulację tych samych procesów na komputerze cyfrowym. Podobnie jak otrzymujemy symulację procesów związanych z wykonywaniem pisemnego obliczenia ilorazu, moglibyśmy uzyskać również symulację procesów uwikłanych w rozumienie języka, percepcję wzrokową, kategoryzację itd..³⁴

Zwolennicy mocnej wersji Sztucznej Inteligencji starają się zatem wykazać, że w umyśle nie istnieje nic warunkowanego biologicznie. W ich ujęciu, mózg stanowi rodzaj maszyny przetwarzającej informację, w której realizują się programy wytwarzające zachowanie inteligentne. Zdaniem przedstawicieli mocnej wersji AI, każdy odpowiednio zaprogramowany system będzie więc tak samo, jak ludzki podmiot, posiadać umysł. Pogląd ten wynikający z założenia, że operacje poznawcze podmiotu są określane w sposób *stricte* formalny implikuje także, iż każdy system wyposażony w zdolność manipulowania symbolami posiada analogiczne/identyczne z ludzkimi zdolności poznawcze³⁵.

Przedstawione powyżej stwierdzenia będące podstawą mocnej wersji Sztucznej Inteligencji są, według Searle'a, błędne. Aby wykazać te błędy, autor *Intentionality* posługuje się własnym eksperymentem myślowym, w literaturze przedmiotu znanym jako *argument chińskiego pokoju*³⁶. Konsekwencją tego argumentu jest twierdzenie, że ludzkie myślenie jest niewspółmierne z operacjami wykonywanymi przez komputer, ponieważ procesu myślenia nie można przedstawić jako zbioru reguł syntaktycznych (obliczeniowych). Ludzkie myślenie jest bowiem wynikiem stosowania nie tylko pewnych reguł syntaktycznych, ale związane jest również z semantyczną interpretacją tych reguł. Píše Searle:

Przyjęcie, że programy można opisywać czysto formalne lub syntaktycznie ma zgubne skutki dla poglądu, że proces umysłowy i proces wykonywania jakiegoś programu są czymś identycznym (...). Mieć umysł, to coś więcej, niż realizować formalne, czy syntaktyczne operacje. Nasze stany umysłowe, na mocy definicji mają zawsze jakąś treść.³⁷

Jeśli więc myśl jest zawsze o czymś, dany ciąg symboli, aby stać się myślą, musi posiadać znaczenie. Umysł ludzki działa zatem nie tylko w oparciu o syntaktykę. Jego działanie determinuje także semantyka.

³³ A. Hodges, *Turing*, tłum. J. Nowotniak, Warszawa 1998, s. 30.

³⁴ J. Searle, *Umysł na nowo...*, s. 267.

³⁵ Por. J. Searle, *Umysł, mózg...*, s. 26.

³⁶ Treść tego argumentu wyłożona jest w pracy *Umysł, mózg...*, ss. 28-29.

³⁷ Tamże, s. 28.

Przyjmując, że ontologicznym wyróżnikiem umysłu jest posiadanie przez niego treści semantycznych, Searle *implicite* wskazuje, że umysł od programu komputerowego odróżnia *modus* intencjonalności. O ile ludzkiemu umysłowi przysługuje *intencjonalność wewnętrzna*, to program komputerowy posiada co najwyżej *jak-gdyby-intencjonalność*. W takim ujęciu, umysł faktycznie myśli, doznaje czy postrzega, natomiast komputer zachowuje się tak jak gdyby myślał, doznawał i postrzegał. Umysł ludzki posiada zatem własność psychiczności, program cechy tej nie posiada³⁸.

Odróżnienie rodzajów intencjonalności przypisywanych człowiekowi i maszynie nie wystarcza jednak, by orzec, że programy komputerowe nie są równoważne ludzkiemu umysłowi pod względem dokonywania operacji poznawczych. Omawiana różnica nie stanowi więc niepodważalnego argumentu przeciwko twierdzeniu, iż maszyny posiadają stany intencjonalne. Jednakże istnieją co najmniej dwa argumenty na rzecz stanowiska Searle'a. Argumenty te przedstawić można następująco:

3.1. „Reguły obliczeniowe, za pomocą których komputery dokonują określonych operacji poznawczych nie są wewnętrznym atrybutem rzeczywistości fizycznej”

Wbrew fundamentalnej tezie funkcjonalizmu głoszącej, że procesy poznawcze mogą realizować się na różnym podłożu fizycznym³⁹, łatwo wskazać, że reguły syntaktyczne nie są własnością fizyczną w takim samym sensie jak *pęd*, *masa* i *grawitacja*. Możliwość wielorakiej realizacji stanów funkcjonalnych systemu nie implikuje zatem charakterystyki relacji przyczynowych zachodzących pomiędzy „stanami wewnętrznymi” danego programu. Jeśli bowiem program ten może być realizowany na różnym podłożu fizycznym, to strukturę programu, twierdzą zwolennicy mocnej wersji AI, określa wyłącznie zbiór reguł obliczeniowych. Searle wskazuje natomiast, że

możliwość wielorakiej realizacji obliczeniowo równoważnych procesów w różnych środowiskach fizycznych oznacza nie tylko, że owe procesy są abstrakcyjne, ale także, że nie są one wewnętrznymi atrybutami tego systemu. Są zależne od zewnętrznej interpretacji.⁴⁰

Dlatego też, tak jak *argument chińskiego pokoju* obalał twierdzenie, że semantyka stanowi immanentną własność składni, tak omawiana teza dowodzi, że syntaktyczna charakterystyka stanów maszyny cyfrowej jest określana przez pewną podmiotową interpretację. Mówiąc, że coś pełni rolę procesu obliczeniowego, mówi się o wiele więcej niż to, że występuje dana konfiguracja zdarzeń (procesów) fizycznych warunkująca realizację procesu obliczeniowego. Zakłada się także, że obliczeniowy charakter danego procesu jest przypisywany przez podmiot⁴¹. Przypisanie to jest zaś możliwe wówczas, gdy podmiot jest przekonany, posiada dowód, że pewien proces realizuje się w oparciu o reguły obliczeniowe. Mówiąc inaczej, ujęcie Searle'a pozwala stwierdzić, iż o tym, czy dany proces jest realizacją pewnych reguł formalnych decyduje bycie podmiotu w danym stanie intencjonalnym. Píše w tej kwestii Searle:

Fizyczny stan pewnego systemu jest stanem obliczeniowym tylko ze względu na pewną obliczeniową rolę, funkcję, czy interpretację, którą temu stanowi przypisujemy

³⁸ Por. tamże, s. 45.

³⁹ Zob. Z. Pylyshyn, *Computation...*, s. 57.

⁴⁰ J. Searle, *Umysł na nowo...*, s. 275.

⁴¹ Por. tamże, s. 277.

(...). Stany obliczeniowe nie są bowiem czymś, co odkrywamy w obrębie zjawisk fizycznych, lecz są przypisywane zjawiskom fizycznym.⁴²

3.2. „Składnia nie ma własności eksplikacji przyczynowej”

Twierdzenie to wynika z faktu, że w ramach mocnej wersji AI błędne nie jest bowiem uwzględniany problem przyczynowości intencjonalnej.

Ludzki komputer – pisze Searle – świadomie przestrzega reguł i faktów wyjaśnia jego zachowania, natomiast komputer mechaniczny nie przestrzega w sensie dosłownym żadnych reguł (...). Komputer mechaniczny nie może przestrzegać reguł, ponieważ nie zawiera intencjonalnych treści, będących wewnętrznym składnikiem systemu, którego przyczynowe oddziaływania generują zachowanie.⁴³

Dzieje się tak, gdyż składnia nie posiada zdolności wyjaśniania kauzalnego. W przypadku działania programów komputerowych o wyjaśnianiu przyczynowym można mówić tylko w tym sensie, że na podstawie wiedzy o pewnych konfiguracjach systemowych orzeka się o istnieniu innych konfiguracji – tych mianowicie, które są przyczyną określonego działania programu. Warto jednakże podkreślić, iż programy komputerowe nie są zdolne do wyjaśnienia przyczynowych interakcji powodujących dane ich działanie, ponieważ nie jest możliwe (na obecnym etapie wiedzy) skonstruowanie komputera zachowującego się w sposób świadomy. Przyczynowe działanie programów „inteligentnych” wyjaśnia się zatem poprzez odwołanie się do intencjonalności programisty ze względu na fakt, że

istota ludzka świadomie przestrzega reguł przeprowadzania określonych obliczeń i fakt ten wyjaśnia przyczynowo wykonywane przez nią czynności. Kiedy jednak programujemy komputer mechaniczny w taki sposób, aby przeprowadzał te same obliczenia, przypisywanie interpretacji obliczeniowej zależy od nas.⁴⁴

Przyczynowe wyjaśnianie procesów obliczeniowych zależy więc od posiadania świadomości (i co się z tym wiąże – własności intencjonalności) przez system realizujący te procedury. Programom komputerowym zaś, można przypisać co najwyżej konfigurację wykazującą formalne podobieństwo do „programów” realizowanych przez ludzki podmiot. Na przykład, systemy „rozumiejące” język naturalny nie wyjaśniają przyczyny, dla której następujące po sobie wyrażenia tworzą zdanie sensowne, ponieważ generowanie zdań opiera się o zaimplementowany zbiór reguł logiczno-gramatycznych (np. różnicy porządku słów, końcówek fleksyjnych itd.) w postaci pewnych warunków formalnych⁴⁵.

4. Podsumowanie

Przestawione argumenty wskazują na dwie ważne konsekwencje dla badań nad sztuczną inteligencją. Po pierwsze, mocna wersja AI jest teoretycznie niespójna. Próba wyjaśnienia natury ludzkiego umysłu wyłącznie na podstawie jego własności syntaktycznych wydaje się być niepełna. Nie podnosi się tu bowiem problematyki fizycznej reprezentacji symboli czy referencji procesów poznawczych danego systemu. Po drugie natomiast, ważnym elementem krytyki dokonanej przez Searle’a jest wskazanie, że problematyka wyznaczana zagadnieniem

⁴² Tamże, s. 276.

⁴³ Tamże, s. 284.

⁴⁴ Tamże, s. 284.

⁴⁵ Zob. L. Bolc, J. Zaremba, *Wprowadzenie do uczenia się maszyn*, Warszawa 1992, ss. 93-102.

intencjonalności niweczy próby sformułowania koherentnej naturalistycznej teorii umysłu. Wniosek ten prowadzi z kolei do konstatacji, iż konstrukcja spójnej koncepcji intencjonalności w ramach teorii Sztucznej Inteligencji mogłaby rozwiązać wiele problemów, z którymi dziedzina ta obecnie się styka⁴⁶. Lecz to już jest zagadnienie na zgoła inny artykuł.

⁴⁶ Por. D. Jaquette, *Searle's Intentionality Thesis*, s. 267-275, w: "Synthese", Nr (80)/1989, s. 272.