

Tytuł: **Implikacje behawioryzmu w badaniach nad sztuczną inteligencją**

Autor: Piotr Kołodziejczyk / pkolodziejczyk@interia.pl

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

Data publikacji: 21 VI 2003

1. Uwagi wstępne

Związek ustaleń psychologii behawiorystycznej z teorią sztucznej inteligencji najbardziej wyraźnie ujawnia się w analizie genezy AI. Skoro bowiem jednym z naczelných zadań sztucznej inteligencji (ujmowanej w jej „mocnej” wersji) miało być stworzenie maszyny porównywalnej z człowiekiem pod względem dokonywanych operacji poznawczych, to działania zmierzające ku realizacji tego zadania *implicite* zakładały charakterystykę operacji poznawczych znamiennej dla człowieka i maszyny matematycznej. Działania te wymagały zatem pewnej podstawy teoretycznej dla uzasadniania formułowanych twierdzeń.

Wydaje się, iż najbardziej naturalnym fundamentem dla uprawomocniania rozstrzygnięć uzyskanych w ramach teorii sztucznej inteligencji była psychologia behawiorystyczna. Konstatacja ta opiera się o dwie racje. Po pierwsze, psychologia behawiorystyczna proponuje prosty schemat wyjaśniania natury procesów poznawczych zachodzących u człowieka. Schemat ten można stosować również do eksplikacji „zachowania” komputerów i na tej podstawie stwierdzać istnienie paraleli pomiędzy człowiekiem a maszyną. Po drugie, w warstwie teoretycznej behawioryzm koresponduje z filozoficznym zapleczem AI – funkcjonalizmem. Podobnie jak filozofia funkcjonalistyczna, również behawioryzm głosi postulat orzekania o zachodzeniu określonych procesów poznawczych poprzez analizę zewnętrznego zachowania się danego systemu. Z tych więc powodów uczynienie behawioryzmu psychologicznym filarem AI może nadawać teorii sztucznej inteligencji oblicze koherentności systemowej. Ustalenia psychologii behawiorystycznej nie są bowiem sprzeczne ani z postulatami funkcjonalizmu, ani z rozstrzygnięciami innych wchodzących w skład AI dyscyplin naukowych. Zasadne więc wydaje się być przebadanie implikacji behawioryzmu na gruncie teorii sztucznej inteligencji.

1. Maszyny a ludzie – warunkowanie zachowania

Mówiąc o doniosłości analizy psychologicznej w projektowaniu sztucznej inteligencji Marvin Minsky stwierdzał:

Punkt zwrotny (w pojmowaniu możliwości programowania maszyn matematycznych, przyp. P. K.) pojawił się gwałtownie w r. 1943 w związku z opublikowaniem trzech artykułów na temat tego, co dziś nazywamy cybernetyką.

Norbert Wiener, Artur Rosenblueth i Julian T. Bigelow z MIT (...) zaproponowali pewne sposoby wbudowania w maszyny celów i dążeń. Warren S. McCulloch (...) i Warren Pitts pokazali, jak maszyny mogą posługiwać się pojęciami logicznymi i abstrakcyjnymi, a K. J. W. Craik (...) podsunął myśl, że maszyny mogą przy rozwiązywaniu problemów stosować modele i analogie. Przy tych nowych założeniach konstruktywnym i potężnym narzędziem opisu maszyn stał się język psychologii.¹

Wypowiadając się na temat związków psychologii behawiorystycznej z teorią i praktyką AI, chciałbym postawić tezę głoszącą, iż korelacja behawioryzmu oraz teorii sztucznej inteligencji zasadza się na opisie zachowania człowieka i „inteligentnych” systemów komputerowych w kategoriach schematu bodźca i reakcji. Teza ta opiera się na następujących (zaczerpniętych z pracy Michaela J. Aptera²) przesłankach:

1. Zarówno o człowieku, jak i maszynie „inteligentnej” można mówić jako o systemie uniwersalnym, czyli systemie przetwarzającym informacje.
2. Systemy te można zatem opisać jako urządzenia otrzymujące informacje ze środowiska zewnętrznego, przetwarzające tę informację oraz przekazujące ją otoczeniu za pomocą urządzeń wyjściowych.

Z tej racji – twierdzi M. J. Apter – modele komputerowe poszczególnych procesów mózgowych nazywa się często „modelami przetwarzania informacji”. Jest dość interesujące, że prawdopodobnie przetworzona w obu przypadkach informacja zostaje zakodowana w formie binarnej. W jednym przypadku spełniają tę funkcję elementy elektroniczne, które mogą przyjmować jeden z dwóch stanów, w drugim zaś – neurony (...), które albo zostają pobudzone, (...) albo nie zostają pobudzone.³

3. Zdolności poznawcze obydwu systemów wynikają z możliwości przeprowadzania wielu prostych operacji.

Przedstawione racje M. J. Aptera (szczególnie przesłanka [2]) pozwalają wnosić, że podobnie jak ludzki umysł, tak i „inteligentne” programy komputerowe traktować można jako systemy, których zachowanie i działanie polega na reagowaniu na dane zestawy bodźców. Pomimo niektórych arbitralnych ustaleń autora *Komputerów...*⁴, zasadniczy zamysł raczej nie budzi zastrzeżeń. Bazując na stwierdzeniach M. J. Aptera chciałbym zaproponować pewną strategię interpretacyjną mającą na celu uzasadnienie wygłoszonej przeze mnie tezy.

¹ M. Minsky, *Sztuczna inteligencja*, ss. 290-304, tłum. D. Gajkowicz, w: *Dziś i jutro maszyn cyfrowych*, red. J. McCarthy, tłum. D. Gajkowicz i inni, PWN, Warszawa 1969, ss. 290-291.

² Por. M. J. Apter, *Komputery a psychika. Symulacja zachowania*, tłum. K. Niemiec, PWN, Warszawa 1973, ss. 16-19.

³ Tamże, s. 17.

⁴ Mam tutaj na myśli przede wszystkim twierdzenie o binarnym charakterze kodowanej przez komputer i człowieka informacji. O ile w odniesieniu do komputerów twierdzenie to jest prawdziwe, to w przypadku analizy przetwarzania informacji przez mózg ludzki ma ono co najmniej wątpliwy status. Myśl tę trafnie oddał John Z. Young pisząc: „Algorytmy języków komputerowych są programami dyktującymi szczegółowo, krok po kroku, operacje logiczne prowadzące do rozwiązania problemu. Niemożliwe są jakiegokolwiek skróty. Niewykluczone, że mózg działa inaczej. Być może, wielość funkcjonalnych połączeń pomiędzy jego częściami pozwala w jakiś sposób dostrzegać analogie między różnymi sytuacjami i przeskakiwać do konkluzji z pominięciem pewnych etapów żmudnego procesu analizowania i sprawdzania wszystkich możliwości. Dopóki nie wiemy, na czym to polega, spekulacje nie na wiele się zdadzą, a nawet próby szczegółowych porównań między mózgiem a komputerem nie mogą być bardzo przydatne. Dobrze zaprogramowane maszyny matematyczne potrafią zmieniać tok swych czynności i swoje cele stosownie do rezultatów poprzednich decyzji. Jak to się dzieje, że miliony komórek mózgowych, mimo to, że pracują powoli dochodzą do rozwiązań, które komputer osiąga dzięki błyskawicznemu wielu możliwości?”; J. Z. Young, *Programy mózgu*, tłum. H. Bartoszewicz, PWN, Warszawa 1984, s. 374.

Przyjmując, iż człowiek i komputer są systemami przetwarzającymi informacje należy założyć również, że owo przetwarzanie odbywa się w pewnym języku. Mówiąc o językowych determinantach możliwości przetwarzania informacji przyjmuję, że funkcjonowanie tych systemów opiera się o proces dekodowania otrzymanej informacji, czyli proces jej zrozumienia i adaptacji. Na kanwie wygłoszonych powyżej stwierdzeń powstaje więc pytanie:

jaki jest związek pomiędzy językową podstawą działania systemów przetwarzających informacje a opisem zachowania tych systemów za pomocą schematu *bodziec-reakcja*?

Punktem wyjścia w odpowiedzi na postawione pytanie będzie przeformułowanie jednego z cytowanych twierdzeń M. J. Aptera (dokładnie – twierdzenia [2]). Można bowiem powiedzieć, że funkcjonowanie omawianych systemów warunkowane jest dostarczeniem im pewnego bodźca (w tym przypadku informacji), operowaniem tym bodźcem i reakcją na niego (rozwiązaniem określonego problemu)⁵. Widocznym jest zatem, że procesy przetwarzania informacji przez systemy inteligentne można wyjaśnić dzięki zastosowaniu operacyjnej definicji pojęć psychologicznych. W ujęciu Stanleya S. Stevensa psychologiczna zasada operacjonizmu zostaje wysłowiona następująco:

Zasady operacjonizmu dostarczają procedury, dzięki której pojęcia psychologii cechować będzie ścisłość. Procedura ta polega na odniesieniu każdego pojęcia – dla jego zdefiniowania – do konkretnych operacji, w których w rezultacie się je uzyskuje, oraz na odrzuceniu wszystkich pojęć mających za podstawę niemożliwe do wykonania operacje. Doktryna operacjonizmu uznaje wyraźnie fakt, że pojęcie czy zdanie, ma znaczenie empiryczne tylko wtedy, gdy może być obiektem określonych, konkretnych operacji.⁶

Transponując ten postulat w obszar badań nad sztuczną inteligencją okazuje się, iż problem eksplikacji zachowania się systemów „inteligentnych” polega na określeniu „inteligentnego” sposobu reagowania systemu na dany zestaw bodźców. Eksponując ten fakt należy podkreślić, że behawiorystyczny opis działania i zachowania się systemów przetwarzających informacje jest zasadny tylko w przypadku analizy językowego zachowania się tych systemów⁷.

Teoretycy badań nad AI, inteligencję pojmują zazwyczaj jako zdolność rozwiązywania problemów. Innymi słowy, inteligentne zachowanie danego systemu polega na przeprowadzeniu szeregu operacji mających na celu osiągnięcie zamierzonego celu. Operacje te można z kolei rozłożyć na ciąg bodźców dostarczanych systemowi w postaci informacji zakodowanych w danym języku. Definiując zaś zbiór dostarczonych systemowi informacji jako zespół operacji koniecznych do rozwiązania danego problemu, w przypadku próby rozstrzygnięcia określonej sytuacji problemowej zachowanie człowieka i „inteligentnego” systemu komputerowego uznać można za tożsame. Opierając się na operacyjnej metodzie definiowania pojęć psychologicznych warto odnotować, że działanie systemów przetwarzających informacje polega na asymilowaniu danych informacji i operowaniu nimi.

⁵ Por. M. A. Arbib, *Mózg, maszyny, matematyka*, tłum. B. Stanosz, PWN, Warszawa 1968, ss. 26-30.

⁶ S. S. Stevens, *Operacyjne definiowanie pojęć psychologicznych*, s. 107, w: *Behawioryzm i psychologia świadomości*, red. J. Siuta, K. Krzyżewski, Wydawnictwo UJ, Kraków 2000, ss. 107-115.

⁷ Zob. S. S. Tomkins, *Simulation of Personality. The Interrelationship Between Affect, Memory, Thinking, Perception and Action*, ss. 44-56, w: *Computer Simulation of Personality*, red. S. S. Topmkins, S. Messick, John Wiley and Sons, New York-London 1963, ss. 3-57. Por. także J. Kammersgaard, *Four Different Perspectives on Human-Computer Interaction*, ss. 49-51, w: *Human-Computer Interaction*, red. J. Preece, L. Keller, Prentice Hall Inc., Engelwood Cliffs 1990, ss. 42-63.

Ważnym jest również fakt, że samo działanie poznawcze omawianych systemów podpada pod behawiorystyczny schemat $S \rightarrow R$ ⁸.

Nietrudno zatem – pisze M. J. Apter – (...) zaprogramować komputer tak, aby generował zachowanie będące odpowiednikiem warunkowania klasycznego i warunkowania instrumentalnego. Tak więc (...), wystarczy zaprogramować komputer tak, aby, po pierwsze, produkował daną liczbę wyjściową, gdy tylko otrzyma daną liczbę wejściową (reprezentuje to odruch bezwarunkowy), po drugie, aby rejestrował w pamięci wszystkie, poprzednie i następne liczby wejściowe, a po trzecie, aby wytwarzał pierwotną liczbę wyjściową w odpowiedzi na wszystkie liczby wejściowe, które (jak wynika z danych zarejestrowanych w pamięci) zwykle pojawiały się przed pierwotną (bezwarunkową) liczbą wejściową lub zaraz po niej. (...) W ten sposób, (...) można za pomocą komputera odtworzyć wiernie zjawiska warunkowania klasycznego.⁹

2. Wnioski

Abstrahując od rozważań natury technicznej, dodać należy, że behawiorystyczny opis zachowania „inteligentnych” systemów przetwarzających informacje implikuje szereg trudności teoretycznych. Trudności te można przedstawić jako dwie grupy zagadnień:

1. Opis zachowania i działania danego systemu oparty na schemacie *bodziec-reakcja* jest niekompletny. W jego ramach nie jest możliwe holistyczne wyjaśnienie funkcjonowania danego systemu poznającego. Ujęcie behawiorystyczne umożliwia co najwyżej eksplikację reakcji na pojedynczą informację (bodziec). Nie tłumaczy ono natomiast sekwencyjnego działania określonego systemu. Zatem, zgodnie ze schematem $S \rightarrow R$ nie można uzasadnić natury działania poznawczego omawianych systemów.
2. Z faktu, że zachowanie „inteligentnych” systemów przetwarzania informacji jest zazwyczaj zachowaniem językowym (werbalnym) wynika, iż u podstaw wyjaśnienia sposobu działania tych systemów leży *implicite* sformułowanie pewnej teorii języka zawierającej konieczne warunki, które winny zostać spełnione, aby werbalne zachowanie człowieka i komputera uznać za nieodróżnialne. Behawioryzm nie proponuje teorii zadawalającej. Sztandarowe twierdzenie behawioryzmu, głoszące, że działanie językowe jest zachowaniem nawykowym nie odnosi się do zagadnienia formalnych warunków poprawności danej wypowiedzi. Na jego podstawie nie można również odpowiedzieć na pytanie o możliwość konstrukcji koncepcji języka opisującej sposób używania języka przez człowieka i komputer.

Z tych więc przyczyn teoria behawiorystyczna jest wysoce dyskusyjna.

⁸ Por. R. P. Abelson, *Computer Simulation of „Hot” Cognition*, ss. 288-298, w: *Computer...*, ss. 277-298.

⁹ M. J. Apter, *Komputery...*, s. 77.