Tytuł: **O pewnym nieporozumieniu oceny badań nad sztuczną inteligencją**

Autor: Piotr Kołodziejczyk ; e-mail: pkolodziejczyk@interia.pl

Źródło: http://kognitywistyka.net ; e-mail: mjkasperski@kognitywistyka.net

Data: styczeń 2004

1. Uwagi wstępne

Prezentowany tekst ma charakter krytyczny. W głównej mierze stanowi on refutację oceny badań nad sztuczną inteligencją i komputacyjnego podejścia do badania umysłu dokonanego przez Rogera Penrose'a. Przyczyny napisania tego artykułu są dwojakie. Po pierwsze moją intencją jest rozwinięcie poglądu Włodzisława Ducha traktującego idee Penrose'a jako konsekwencje turingowskiego 'argumentu chowania głowy w piasek'[1]. Po drugie zamierzam wykazać niezasadność szkolnych opracowań tematu bazujących w swej krytyce badań nad sztuczną inteligencją na twierdzeniu Gödla o niezupełności arytmetyki.

Dokonując wstępu do właściwych rozważań chciałbym podkreślić, że właśnie takim szkolnym, wynikającym jak się wydaje z niezrozumienia tematu, podejściem jest krytyka Penrose'a. Pokazując jej ograniczenia przyjąłem następującą strukturę wywodu. W pierwszej części tekstu dokonam analizy relacji pomiędzy badaniami nad umysłem i matematyką oraz postaram się wykazać, że argumenty przeciwko AI oparte na twierdzeniu Gödla można traktować jako nadużycie epistemologiczne i metodologiczne. W części drugiej podejmę próbę ukazania słabych punktów krytyki Penrose'a oraz ich konsekwencji dla badań nad sztuczną inteligencją. W części trzeciej zaprezentuję metodologiczną ocenę pozytywnej części koncepcji Penrose'a oraz jej krytykę.

2. Rola matematyki w krytyce badań nad sztuczną inteligencją

Mówiąc z dużym uproszczeniem, badania nad sztuczną inteligencją można określić mianem analiz algorytmicznych. Oznacza to, że próba konstrukcji sztucznych podmiotów poznawczych opiera się na idei znalezienia takich algorytmów, za pomocą których będzie

[1] Stwierdzenie to odnalazłem w 6-tej części wykładów prof. Ducha, dotyczących podstaw kognitywizmu i filozofii Sztucznej Inteligencji, znajdującej się na jego stronie internetowej.

możliwe symulowanie/implementowanie w systemach sztucznych operacji kognitywnych identycznych ze zdolnościami poznawczymi podmiotów naturalnych. Tego rodzaju wnioski można wysnuć zarówno z analizy zadań, jakie przed swoją maszyną stawiał Turing oraz z opisu celów, jakie przed swoimi badaniami stawiali pionierzy AI. Kwestię doniosłości procedur obliczeniowych dla badań nad AI Stevan Harnard ujął następująco. Jego zdaniem, obliczanie jest to

> niezależna od interpretacji, fizycznie interpretowalna manipulacja symbolami. Tylko w ten sposób okazało się możliwe przeprowadzenie jakiegokolwiek obliczania, o jakim matematycy mogli jedynie marzyć, a także spowodowanie, by maszyny wykonywały wiele różnych inteligentnych czynności, do których wcześniej zdolni byli jedynie ludzie (i zwierzęta). Prawdopodobnie można było więc w tych warunkach całkiem naturalnie wnioskować, że ponieważ: (1) nie wiemy, w jaki sposób podmioty poznawcze poznają i (2) obliczanie może dokonywać tak wiele czynności, do jakich zdolny jest poznający podmiot, to poznanie jest tylko formą obliczania. (...) Ostatecznie, zgodnie z tezą Churcha-Turinga obliczanie jest wyjątkowe i najwyraźniej wszechmocne (...); cokolwiek mogą zrobić systemy fizyczne, mogą to także wykonywać komputery. [Harnard 1995, s. 384][2]

Na podstawie tej, być może przydługiej wypowiedzi, widać, że w badaniach nad sztuczną inteligencją refleksja nad naturą obliczalności pełni fundamentalną rolę. Trzeba jednak zaznaczyć, że pojęcie obliczalności stanowi swoistego rodzaju obusieczną broń. Z jednej strony wykorzystywano je jako podstawę postulowania analogii pomiędzy sztucznymi a naturalnymi podmiotami poznawczymi, z drugiej zaś – w oparciu o nie starano się wykazywać nieprawomocność (w sensie metodologicznym i epistemologicznym) komputacyjnych podejść do badania umysłu.

Jedną z bardziej znanych prób związanych z krytyką komputacjonizmu opartą na konsekwencjach wynikających z twierdzenia Gödla jest wystąpienie Johna Lucasa. Krytyka Lucasa z punktu widzenia tej pracy jest ważna dlatego, że taką samą linię argumentacji w wywodach swych przyjmuje Penrose. Jest zatem oczywiste, że wszelkie zarzuty skierowane przeciwko Lucasowi będą się automatycznie odnosić do krytyki badań nad AI dokonanej przez Penrose'a. Sedno krytyki Lucasa opiera się na założeniu, że porównanie sztucznych i naturalnych podmiotów poznawczych w zakresie wnioskowania dedukcyjnego wystarcza, aby systemom sztucznym odmówić takich własności poznawczych, jakie posiadają systemy naturalne, np. generalizowania, dowodzenia twierdzeń, wnioskowania. Opierając się na twierdzeniu Gödla, Lucas stara się argumentować, że systemy sztuczne nie mają własności poznawczych, gdyż nie są zdolne do dowodzenia zdań samozwrotnych. Odnosząc tą tezę do niedeterministycznych maszyn Turinga, Lucas wskazuje, że systemy izomorficzne z automatami Turinga nie posiadają zdolności rozumienia, a zatem nawet matematyczne kompetencje sztucznych i naturalnych podmiotów kognitywnych nie mogą być utożsamiane [zob. Lucas 1961 oraz por. Kołodziejczyk 2004].

Wydaje się, że strategia krytyki badań nad AI zaproponowana przez Lucasa jest nadzwyczaj myląca. Moim zdaniem zawiera ona szereg nieporozumień natury metodologicznej i epistemologicznej. Po pierwsze, Lucas dokonuje nieuprawnionego metodologicznie zastosowania twierdzenia Gödla. Jest oczywiste, że twierdzenie to dotyczy przede wszystkim obiektów matematycznych. Nie wnikając w tym miejscu w rozstrzygnięcia dotyczące sposobu istnienia przedmiotów matematycznych, trzeba zaznaczyć, że przedmioty te mają inną charakterystyką ontyczną niż systemy AI. Wynika stąd, że twierdzenie Gödla nie ma

[2] Na temat związku badań nad obliczalnością z badaniami nad AI szerzej piszę w moim artykule pt. *Rola teorii obliczalności w badaniach nad sztuczną inteligencją* w: http://kognitywistyka.net.

zastosowania do opisu funkcjonowania sztucznych systemów kognitywnych. Można je używać tylko w procesie analizy niesprzeczności algorytmów składających się na sztuczne systemy poznawcze, nie zaś – w akcie badania wytworów tych systemów. Dodatkowego argumentu przeciwko zasadności krytyki Lucasa dostarczają stwierdzenia Johna L. Castiego i Wernera DePauliego. Analizując związek twierdzenia Gödla z badaniami nad AI piszą oni:

> wystarczy przypomnieć, że twierdzenie Gödla wymaga przyjęcia pewnych założeń. Najważniejsze z nich to założenie, że system formalny (program komputerowy) jest logicznie niesprzeczny. Wydaje się oczywiste, że spełnienie tego warunku w przypadku ludzkiego umysłu jest – delikatnie mówiąc – bardzo wątpliwe; wszyscy z pewnością pamiętamy przykłady własnych, sprzecznych zachowań. Jeśli zaś system jest logicznie sprzeczny, to z twierdzenia Gödla nic nie wynika. [Casti, DePauli, 2003, s. 142]

Konsekwencją takiego postawienia sprawy może być ukazanie, że Lucas bezzasadnie założył jednoznaczną izomorficzność pomiędzy naturalnymi i sztucznymi podmiotami poznawczymi. Mówiąc inaczej, przyjął on, że umysł ludzki i programy komputerowe działają na podstawie dokładnie takich samych algorytmów. Dobrze będzie, więc w tym miejscu przypomnieć, że porównanie ludzkiego umysłu i programów komputerowych jest co najwyżej bardzo wygodną metaforą stosowaną w opisie natury i funkcji procesów poznawczych [por. Dennett 1997]. Nawet, jeśli można osłabić zarzuty odnoszące się do krytyki Lucasa, to można to uczynić tylko poprzez stwierdzenie, że są one zasadne tylko w przypadku ich zastosowania do programu badania sztucznej inteligencji w jego mocnym sformułowaniu. Zarzuty te nie odnoszą się do słabej wersji AI – natomiast sama wersja mocna jest dziś traktowana historycznie [zob. Searle 1999, s. 145]. Dlatego też mocna wersja AI nie wydaje się być badawczo ciekawa.

Mimo, iż lucasowska krytyka paradygmatu obliczeniowego zakrawa na nieporozumienie, to trzeba zaznaczyć, że jest ona wykorzystywana również we współczesnych dyskusjach dotyczących możliwości i zasadności projektowania sztucznych systemów kognitywnych. Poniżej przedstawię rozwinięcie krytyki Lucasa dokonanej przez Penrose'a i postaram się pokazać słabe strony tej argumentacji.

3. Gödel, rozumienie i inne demony Penrose'a

Sedno krytyki badań nad sztuczną inteligencją Penrose, podobnie jak Lucas, opiera na twierdzeniu Gödla. Kontekst odkrycia tego twierdzenia dla refutacji komputacyjnego paradygmatu badania umysłu ujmuje on następująco:

> W czasie nauki na Uniwersytecie w Cambridge słuchałem wykładu z logiki matematycznej. Właśnie on spowodował zmianę moich poglądów na temat mózgu (…). Kiedy dowiedziałem się o pewnych współczesnych koncepcjach w logice, takich jak twierdzenie Gödla (…) uznałem, że ludzki mózg przynajmniej podczas rozwiązywania problemów matematycznych musi działać w sposób niealgorytmiczny. [Penrose 2001]

Przyjęcie tezy o niemożliwości konstrukcji sztucznych podmiotów kognitywnych, czyli niezasadności stanowiska A [zob. Penrose 1997, s. 106] wynika ze stwierdzenia, że pewne problemy z zakresu teorii liczb nie są rozwiązywalne za pomocą algorytmów. Tytułem przykładu można podać w tym miejscu hipotezę Goldbacha, czy tzw. zagadnienie czterech barw. Zdaniem Penrose'a przykłady te wskazują, że rozwiązanie problemów matematycznych wymaga przede wszystkim zaangażowania procesu rozumienia. W przypadku naturalnych

podmiotów poznawczych rozumienie jest, jak podkreśla Penrose, definicyjną cechą ich umysłu. Sztuczne systemy kognitywne nie posiadają zaś, w jego opinii, tej własności. Na zasadność tych stwierdzeń wskazywać ma następująca konsekwencja twierdzenia Gödla:

> żaden zbiór reguł obliczeniowych nie może w pełni scharakteryzować własności liczb naturalnych. A jednak, mimo, że nie istnieje obliczeniowy sposób opisania liczb naturalnych, zna je każde dziecko. [Penrose 1997, s. 119]

Dlaczego? Ponieważ, twierdzi Penrose, dziecko posiada intuicyjną znajomość własności tych liczb. Mówiąc inaczej, rozumie ono, czym są liczby naturalne i dlatego potrafi nimi operować. Natomiast sztuczne podmioty poznawcze nie posiadają zdolności rozumienia, i z tej przyczyny ich sposób działania jest ograniczony, nieporównywalny z zakresem problemów, które jest zdolny rozwiązać podmiot naturalny. Czy tak jest w istocie?

Aby odpowiedzieć na zarzuty Penrose'a, ograniczę się do porównania sztucznych i naturalnych systemów kognitywnych tylko do sposobu rozwiązania zagadnień matematycznych. Posłużę się przy tym niektórymi ustaleniami wynikającymi z filozofii nauki Thomasa S. Kuhna. W punkcie wyjścia zastąpię wieloznaczny termin 'rozumienie' pojęciem 'inteligencji'. Zastąpienie to jest zgodne z intencją Penrose'a, który w swych pracach obydwa terminy często stosuje synonimicznie [zob. Penrose 1997, s. 105].

Przyjmując operacyjną definicję inteligencji w odniesieniu do sztucznych i naturalnych podmiotów poznawczych, trzeba przyznać, że zdolność rozwiązywania problemów matematycznych jest warunkowana algorytmami lub heurystykami dostarczonymi danemu systemowi. Podmiotom naturalnym algorytmy/heurystyki dostarczane są zazwyczaj poprzez podawanie im wielu przykładów oraz reguł operowania nimi, np. analogii, indukcji. Podobnie sprawa ma się w przypadku systemów sztucznych. Algorytmy są dostarczane systemowi przez programistę jako twierdzenia bazowe. Natomiast przykłady to nic innego jak dane dostarczane do systemu. Proces rozwiązywania problemów matematycznych można w przypadku obydwu systemów przedstawić jako asymilację danych, ujęcie ich za pomocą znanych uprzednio kategorii/schematów oraz zastosowanie danej reguły operowania danymi. Przykładem zastosowania tego schematu jest proces uczenia się dzieci lub działanie tzw. systemów ekspertowych. Jeśli rozwiązanie problemów matematycznych można przedstawić w postaci algorytmu/heurystyki, to proces ten można oczywiście ująć obliczeniowo. Jest przy tym jasne, że efekty obliczeniowego podejścia do rozwiązywania problemów przez sztuczne i naturalne systemy kognitywne nie zawsze będą poprawne. Znaczy to, że zastosowanie procedur obliczeniowych pojmowanych jako heurystyki zakłada konieczność eliminacji i korekcji błędów mogących się pojawić w procesie rozwiązywania problemów. Kwestii tej zdaje się nie zauważać Penrose. Pisze on:

> jeśli maszyna ma być nieomylna, to nie może być również inteligentna [Penrose 1997, s. 116]

negując tym samym wiele spośród ustaleń robotyki kognitywnej. Choćby pobieżny rzut oka na rozwój tej dyscypliny wskazuje, że istnieją systemy posiadające inteligencję i jednocześnie w sposób nieomylny rozwiązujące postawione przed nimi problemy, np. robot Qrio, pokazany po raz pierwszy końcem 2003 r.[*] Ponadto, założenie, że sztuczne systemy poznawcze muszą

[*] Żeby być ściślejszym: prototypowy model robota *Qrio* firmy Sony był już gotowy i pokazany w roku 2000 na wystawie „Robodex 2000" – nosił on wówczas roboczą nazwę SDR-3X. Później został nieco przemodelowany i powstał wtedy słynniejszy model SDR-4X, na którego konstrukcji oparto właśnie SDR-4X II – czyli *Qrio*. Specyfikacje tychże modeli można prześledzić odwiedzając oficjalną stronę internetową projektów: http://www.sony.net/SonyInfo/QRIO/. Przyp. red. M. Kasperski.

być nieomylne jasno ukazuje, że Penrose nie zna ani współczesnych trendów dominujących w badaniach nad AI, ani nie zdaje sobie sprawy z postępów, jakie w ostatnich latach zaszły zarówno w teoretycznej, jak i praktycznej sferze badań nad sztuczną inteligencją. Jego krytykę głównych założeń AI można traktować więc jako nieporozumienie. Na podobny zarzut jest, jak się zdaje, narażona pozytywna wykładnia poglądów Penrose'a na temat umysłu. Poniżej postaram się podać argumenty na rzecz postawionej tezy.

4. Zamiast zakończenia – kwantowa nieobliczalność, czyli magia w fizyce

Tłem dla sformułowania swojej koncepcji umysłu Penrose czyni wnioski wynikające z mechaniki kwantowej. Na ich podstawie formułuje tezę o nieobliczalności umysłu, zgodnie z którą akty i procesy umysłowe nie dają się ująć za pomocą reguł obliczeniowych. Natomiast w warstwie pozytywnej twierdzi on, że działanie umysłu jest warunkowane kwantowymi reakcjami zachodzącymi w mikrotubulach. Pomijając problem jednoznacznego określenia związku między procesami umysłowymi a mikrotubulami, warto podkreślić, że koncepcja ta jest skazana na niepowodzenie, ponieważ zgodnie z nią nie jest możliwe trwałe zaistnienie dowolnego procesu umysłowego, gdyż czasy koherencji procesów kwantowych zachodzące w mikrotubulach są niezwykle krótkie. Zatem, w świetle podejścia Penrose'a do zagadnienia umysłu, nie sposób wyjaśnić ani zagadnienia konstytucji cech i świadomości, ani sposobu, w jaki procesy umysłowe łączą się ze sobą na przykład w aktach percepcji czy podejmowania decyzji. Stanowisko Penrose'a można więc, za Duchem, określić jako 'kompletne manowce'. Zbyt wiele w jego ramach niejasności i niedomówień. Można mieć tylko nadzieje, że kolejna praca Penrose'a przyniesie w końcu usystematyzowaną wykładnię jego koncepcji pozbawioną argumentacji opartej na twierdzeniu Gödla, której trudność udało mi się, jak sądzę, wykazać.

Literatura wykorzystana:

[1] J. L. Casti, W. DePauli, *Gödel. Życie i logika*, tłum. P. Amsterdamski, Wyd. CiS, Warszawa, 2003.

[2] D. Dennett, *Natura umysłów*, tłum. W. Turopolski, Wyd. CiS, Warszawa, 1997.

[3] W. Duch, *Wykłady z filozofii kognitywnej*, źródło: Internet.

[4] S. Harnard, *Computation Is Just Interpretable Manipulation. Cognition Isn't*, w: "Minds and Machines", Nr 4/1995, ss. 379-390.

[5] P. Kołodziejczyk, *Funkcjonalizm jako ontologia sztucznej inteligencji*, w: *Byt i jego pojęcie*, red. A. L. Zachariasz, Wyd. UR, Rzeszów, 2004.

[6] P. Kołodziejczyk, *Rola teorii obliczalności w badaniach nad sztuczną inteligencją*, w: http://kognitywistyka.net.

[7] J. Lucas, *Minds, Machines and Gödel*, w: "Philosophy", Nr 137/1961, ss. 112-137.

[8] R. Penrose, *Makroświat, mikroświat i ludzki umysł*, tłum. P. Amsterdamski, Wyd. Prószyński i S-ka, Warszawa, 1997.

[9] R. Penrose, *Mózg nie potrafi myśleć algorytmicznie!* (wywiad udzielony redakcji "Wiedza i Życie"), w: "Wiedza i Życie", Nr 9/2001.

[10] J. Searle, *Świadomość, inwersja wyjaśnień i nauki kognitywne*, tłum. E. Hunca, w: *Modele umysłu*, red. Z. Chlewiński, PWN, Warszawa, 1999, ss. 144-177.