Przegląd Filozoficzny – Nowa Seria
R. 22: 2013, Nr 2 (86), ISSN 1230–1493
DOI: 10.2478/pfns-2013-0034

Marcin Jażyński

Czy inżynier wie, co myśli jego robot?

Słowa kluczowe: *filozofia umysłu, kognitywistyka, umysł, sceptycyzm, sztuczna inteligencja, czarna skrzynka, naturalizm*

1. Wstęp, czyli co ja właściwie zrobiłem?

W jednym z opowiadań Henry'ego Kuttnera, zamieszczonym w zbiorze pod tytułem *Stos kłopotów*[1], występuje pewien wybitny badacz i wynalazca. To naprawdę zdolny człowiek. Jednak ma on tę cechę, że wpada na pomysły tylko wtedy, gdy jest kompletnie pijany. Kiedy się budzi, spostrzega, że znowu coś wymyślił i skonstruował, ale nie ma pojęcia, czym to właściwie jest i co robi. O tym, jak stara się tego dowiedzieć, przeczytajcie sami.

Pod pewnym względem sytuacja kognitywisty jest podobna do zagubienia Kuttnerowskiego bohatera. Kognitywista ma nadzieję na wyjaśnienie tego, co dzieje się w „czarnej skrzynce". Współczesny inżynier umysłu chce to zrobić tylko przy użyciu naturalistycznych metod i narzędzi badawczych. Dane uzyskane dzięki nim uważa za wiarygodne i zupełne. Innych świadectw nie szanuje, bo psychologia to naiwna forma cybernetyki i neurofizjologii. Na filozoficznym zapleczu tego programu znajduje się funkcjonalistyczna teoria umysłu, nastawiona obliczeniowo, oraz tak zwane ucieleśnione poznanie.

Czy kognitywista-konstruktor może się dowiedzieć, jakie myśli kryje czarna skrzynka?

[1] Tę świetną, śmieszną książkę polecił mi Marcin Kałek. Jestem mu za to wdzięczny.

2. Co to robi i myśli?

Kuttnerowski naukowiec zadawał sobie pytania: „Co ja zrobiłem? Co to jest? Do czego to służy?" Zadanie filozoficznie zorientowanego inżyniera może być podobne i da się prosto wyrazić. W wolnej chwili może on zadać sobie pytanie: co robi i myśli mój robot lub moja czarna skrzynka? Na pierwszy rzut oka widać, że robot i czarna skrzynka znacząco się różnią. Otóż wydaje mi się jednak, że pod względem możliwości poznawczego dostępu do nich są takie same. Proponuję więc, by przez chwilę traktować oboje podobnie. Rozważmy dwie sytuacje:

a) W pierwszej inżynier nie jest twórcą robota. Na czym może oprzeć swoje badania? Może obserwować zachowanie czarnej skrzynki, ale może też do niej zajrzeć. Tym, co znajdzie po otwarciu, będzie, mówiąc nieco metaforycznie, działający program. Kłopot inżyniera polega na tym, czy potrafi on jednoznacznie określić, co ten program robi? Na podstawie obserwacji zachowania skrzynki nie może stwierdzić, wedle jakiej instrukcji ona działa, gdyż jej operacje mogą być zgodne z więcej niż jedną regułą.
b) W drugiej sytuacji inżynier jest twórcą robota. Wyposażył go w napisane przez siebie wielozadaniowe programy. Jest to bardzo zdolny inżynier. Stworzył takiego robota, który potrafi reorganizować swój system poznawczy w zależności od zadań, jakie ma zrealizować, i pod wpływem nauki, tj. tego, czego się uczy.

Załóżmy, że wielki inżynier skonstruował robota i wrzucił go w środowisko. Potem odpoczął, a po jakimś czasie chce sprawdzić, jak jego wytwór sobie radzi, co robi i co myśli.

Wydaje mi się, że ma wciąż ten sam kłopot. To znaczy, nie jest w lepszym położeniu niż obserwator czarnej skrzynki. Niechciana sceptycka konkluzja brzmi: zaglądanie do skrzynki nie jest potrzebne, bo nic nie da.

3. Problem filozoficzny

Gdyby Bóg wejrzał w nasze dusze, to nie mógłby zobaczyć, o kim myślimy[2]

Czy gdyby zespół inżynierów zaprojektował i skonstruował sztuczny system, który by myślał, mówił w języku naturalnym i zachowywał się w przybliżeniu tak, jak ja, to czy w oparciu o swoją wiedzę dotyczącą jego budowy i funkcjo-

[2] Por. L. Wittgenstein, *Dociekania filozoficzne*, przeł. B. Wolniewicz, Wydawnictwo Naukowe PWN 2000, XI.

nowania inżynierowie mogliby wyjaśniać przebieg jego procesów umysłowych oraz treść stanów jego umysłu?

Powiedziałem, że znajomość działającego programu nie wystarczy, by stwierdzić, „o czym myśli" i co robi robot. Jest tak, ponieważ na podstawie obserwacji zachowań maszyny Turinga nie można rozpoznać, jaki program ona realizuje, to znaczy nie można poznać znaczenia wpisywanych i wymazywanych symboli[3].

Trudność tę można zilustrować następująco: wyobraźmy sobie trzy kompozycje muzyczne, które różnią się od siebie, poza jednym szczególnym fragmentem, jaki został skomponowany i zagrany dokładnie tak samo. Czy słysząc tylko ten fragment, moglibyśmy stwierdzić, którą z nich słyszymy?[4]

Wracając do umysłu, załóżmy, że chcemy się dowiedzieć, jaka jest treść reprezentacji umysłowych robota. Jaką operację wykonuje program? Inaczej mówiąc, chcemy się dowiedzieć, o czym on myśli. Problem w tym, że ten sam program może realizować różne, odmienne operacje i zadania, jak również kilka różnych programów może realizować taką samą operację. Na przykład komputerowa symulacja partii szachów może być nieodróżnialna od symulacji jakiejś potyczki wojennej. Obserwując ją „z zewnątrz", nie wiemy, czy mamy do czynienia z grą, czy z bitwą[5]. Inny przykład: program *Warplan* D.H.D. Warrena może wykonywać trzy operacje w nieodróżnialny od siebie sposób, to jest sterować akcją robota, który chodzi po pokoju i przestawia przedmioty, sterować robotami dokonującymi przemysłowego montażu samochodów lub kompilować wyrażenia matematyczne na kod maszynowy[6].

Podsumowując, problem w tym, że na podstawie obserwacji działania programu nie jesteśmy w stanie jednoznacznie określić, co ten program robi. I analogicznie – nie możemy jednoznacznie zidentyfikować treści myśli czy reprezentacji robota.

4. Uwaga sceptyczna

Wyobraźmy sobie, że komputer zostaje wyposażony w ciało – odpowiedniki organów zmysłowych, urządzenia umożliwiające ruch i zmienianie środowiska. Jest też obdarzony percepcją – spostrzega świat i działa w nim. Zapewne jego wewnętrzne reprezentacje zostaną jakoś powiązane ze środowiskiem i tym samym nabiorą znaczenia. Możemy postawić następujące pytanie: czy ktoś

[3] Jeśli tak, to jak można by zawęzić listę możliwych programów, z których jeden jest rzeczywiście realizowany?

[4] Wpadłem na to dzięki uwagom profesora Jacka Hołówki.

[5] Przykład J. Fodora.

[6] Przykład J. Bobryka.

inny niż on sam i jego konstruktor-programista może znać znaczenie jego myśli?

Odpowiedź trąci sceptycyzmem: analiza wewnętrznych operacji komputera – robociego mózgu – nie powie nam, co robi dany program, bo, jak wcześniej, te same zjawiska – celowe zachowania – można modelować w różny sposób. Wydaje się, że abyśmy mogli poznać znaczenie reprezentacji robota, niezbędna jest wiedza o zamiarach programisty. A jeśli ich nie znamy?

W takiej właśnie sytuacji znajduje się współczesny badacz umysłu. Ma do czynienia z działającym systemem, a nie zna intencji jego projektanta. Tak naprawdę jest jeszcze gorzej. Samego projektanta nie można zapytać o intencje, a nawet gdyby to się udało, i tak nie wiedziałby, co odpowiedzieć, bo wyposażył swojego robota w zdolność do zmiany programów działania. I nawet gdyby wejrzał w jego duszę, nie mógłby zobaczyć, o czym myśli.

Działania robota mogą być zgodne z różnymi instrukcjami lub – używając słownika Wittgensteina – jego zachowania mogą być zgodne z różnymi regułami działania. Obserwacja programu nie pozwoli jednoznacznie określić, jaką operację program wykonuje. Natomiast obserwacja zewnętrznego zachowania robota nie pozwoli na jednoznaczne określenie, którą regułę działania on stosuje. Czyżbyśmy nie mogli się dowiedzieć, co on w ogóle robi? A może jesteśmy na zbyt ogólnym poziomie opisu umysłu i działania? Jednak zejście na niższy poziom nic nie da. Zobaczymy wtedy syntaktyczne manipulacje na symbolach lub zmiany potencjału czynnościowego, ale nie będziemy widzieli, co one znaczą.

W przypadku tak zwanego ucieleśnionego poznania mamy ten sam kłopot: jak wyróżnić i określić treść powstających scenariuszy działania oraz decyzję o wyborze jednego z nich?

5. Eksperyment myślowy Davidsona

Analogiczne problemy pojawiają się w przypadku prób opisu i wyjaśnienia psychiki człowieka – treści myśli, reprezentacji i denotacji. Czy wiedza o fizycznych zdarzeniach w mózgu pociąga za sobą wiedzę o zdarzeniach mentalnych? Czy dysponując pełną wiedzą o procesach mózgowych możemy coś powiedzieć o stanach mentalnych, w szczególności o treści postaw propozycjonalnych?

5.1. Art

Sformułujmy dwa wygórowane założenia. Ich realizacja byłaby ideałem fizykalisty. W rzeczywistości jest to tylko jego wyznanie wiary:

a) Mamy pełną wiedzę o procesach, jakie przebiegają w naszym mózgu i układzie nerwowym.
b) Jest ona sformułowana w słowniku fizykalnym.

Następnie wyobraźmy sobie, że w oparciu o naszą znajomość fizyki skonstruowaliśmy sztucznego człowieka – Arta. Jest on zbudowany z tego samego materiału, co my. Wszystkie procesy fizjologiczne i neurologiczne, jakie zachodzą w nas, przebiegają także w Arcie. Struktura fizyczna i funkcjonalna mózgu i ciała Arta jest taka sama jak w przypadku ludzkiego mózgu i ciała. Zachowanie Arta jest pod wszystkimi względami takie samo jak zachowanie człowieka. Nie możemy stwierdzić, patrząc na jego działania, że powstał w inny sposób niż my. Art wykonuje wszystkie czynności, mówi i reaguje na bodźce tak jak my. Dobrze też rozumiemy procesy zachodzące w mózgu Arta i potrafimy je identyfikować i opisywać w terminach czysto fizykalnych. W tym momencie powraca nasze kluczowe pytanie: co możemy powiedzieć o stanach mentalnych Arta, opierając się jedynie na naszej szerokiej fizycznej wiedzy o budowie i procesach przebiegających w jego mózgu? Problem dotyczy możliwości i sposobu, w jaki mielibyśmy przypisywać postawom propozycjonalnym, takim jak żywienie przekonań i pragnień, ich propozycjonalną treść. Stany umysłowe, jakimi są postawy propozycjonalne, różnią się od siebie przede wszystkim pod względem treści, tj. co do tego, czego dotyczą przekonania lub na co są skierowane zamiary. Dzięki znajomości treści moglibyśmy rozróżniać i identyfikować poszczególne postawy i wyjaśniać zachowanie. Zatem czy posługując się pojęciami fizykalnymi możemy rozróżniać i opisywać treści stanów psychicznych Arta?

5.2. Nadzieja inżyniera

Przy założeniu identyczności stanów psychicznych i fizycznych, wiedząc wystarczająco dużo o mózgu, powinniśmy umieć opisać stany, jakie w języku psychologii nazywamy stanami mentalnymi Arta, i oczywiście powinniśmy to zrobić w terminologii naturalistycznej. Zapewne oznaczałoby to, że mentalistyczna psychologia może zostać z powodzeniem sprowadzona do cybernetyki, a potem do fizyki, i niczego na tym nie stracimy.

5.3. Przykład

Zanosi się na deszcz. Art mówi: „Będzie padać”. Chcemy wiedzieć, o czym on myśli. Jesteśmy świadkami jego zachowania werbalnego, a oprócz tego wiemy, jakie procesy fizyczne zachodzą w mózgu Arta w chwili, kiedy wypowiada on zdanie „Będzie padać”. To znaczy: wiemy, że takiemu zachowaniu werbalnemu towarzyszą takie a takie procesy neurofizjologiczne, albo: kiedy

następowało pierwsze – wypowiedź Arta, że będzie padać – równocześnie następowało drugie – w jego mózgu przebiegał szereg pewnych znanych nam zdarzeń fizycznych. Czy znając wszystkie potrzebne fizyczne dane, dotyczące tego, co się dzieje w mózgu Arta, kiedy mówi o deszczu, moglibyśmy stwierdzić, co Art ma w tej chwili na myśli? Jeśli się nie mylimy, powinno być tak, że jeżeli wiadomo dostatecznie dużo o tym, co się dzieje w mózgu Arta w danej chwili, wiadomo tym samym wszystko, co potrzebne, by odpowiednio do stanu neurofizjologicznego przypisać konkretne przekonanie o określonej treści. Problem w tym, że jak się wydaje, nie sposób jednoznacznie przypisać pewnej jednej treści mentalnej do stanu mózgu. Dlaczego tak może być?

Przyjmijmy, że w tej sytuacji zachodzi w mózgu Arta szereg procesów neurofizjologicznych. Otóż może być przecież tak, że Art wtedy myśli, że deszcz, który pada, jest ładny, lub równie dobrze, że zapomniał pieniędzy i nie pójdzie na kawę, lub że przedwczoraj nie padało. Wiadomo, że w mózgu Arta funkcjonuje pewien mechanizm neurofizjologiczny, lecz na tej podstawie nie sposób jednoznacznie określić treści myśli Arta. Oprócz tego może być przecież i tak, że innym razem Art będzie myśleć o deszczu, a ten mechanizm nie będzie działał.

Możemy stwierdzić, że pewne zdarzenie fizyczne wywołało pewien stan mentalny, oraz wiadomo nam, że jest on identyczny z pewnym fizycznym zdarzeniem (czy stanem) w mózgu Arta, ale mimo to nie potrafimy nic pewnego powiedzieć o treści tego stanu mentalnego. Bowiem nie jest tak, że zdarzenie fizyczne lub stan mózgu jednoznacznie wyróżnia pojedynczy stan mentalny o takiej a nie innej określonej treści. Nie możemy w ten sposób jednoznacznie zidentyfikować i wyróżnić jego treści spośród wszystkich możliwych. Nie pomoże nam tu dodatkowa znajomość fizycznego stanu mózgu Arta z chwili, w której patrzył on na chmury. Nadal nie będziemy w stanie przypisać mu tego a nie innego przekonania. Korzystając tylko z naszej fizycznej wiedzy o przebiegu procesów mózgowych Arta, nie możemy wiele powiedzieć o treści jego przekonań. Co najwyżej poprzez badania moglibyśmy stwierdzić, że Art w danej chwili myśli o czymś mniej lub bardziej intensywnie.

5.4. Dlaczego tak może być?

Po pierwsze, twierdzenie o identyczności nie jest dobrą podstawą dla opisu niektórych istotnych własności stanów psychicznych. Identyczność stanów mózgu i stanów umysłu nie wystarcza, byśmy mogli zidentyfikować pewien dany stan fizyczny mózgu Arta z konkretnym przekonaniem o określonej jednoznacznie treści. Po drugie, nawet jeśli osłabimy nasze fizykalne zapędy do wywoływania, a nie identyczności, nadal na podstawie wiedzy, że dane zdarzenie fizyczne wywołuje pewien stan mentalny, nie będziemy mogli jednoznacznie opisać

czy określić, jaki to stan mentalny. Rozumiem to tak, że dane twierdzenie fizykalne o pewnym stanie mózgu (lub zdanie fizykalne o związku przyczynowym fizyczne-mentalne) nie determinuje jednoznacznie zdania o określonej, takiej a nie innej treści propozycjonalnej. Nie jest to wystarczające dla wykazania, że możemy wyjaśniać i przewidywać zdarzenia psychiczne, tak jak przewidujemy zdarzenia fizyczne, ani że opisy zdarzeń psychicznych są redukowalne do opisów zdarzeń fizycznych. Myślę, że równie dobrze można by ten przykład zinterpretować kognitywistycznie. Art może być robotem, o jakim pisałem wcześniej. W miejsce silnego twierdzenia o identyczności stanów mózgu i umysłu pojawi się nieco słabsze czy bardziej subtelne funkcjonalistyczne twierdzenie, mówiące, że umysł jest tym, co robi mózg (czy jakiś inny fizyczny realizator), a to, co robi, polega na realizacji pewnego złożonego programu lub wielu programów. Obawiam się, że powyższe wątpliwości wymierzone w radykalny fizykalizm stosują się też do kognitywistycznego naturalizmu. Funkcjonalny związek między programem i jego realizatorem może nie wystarczyć do tego, byśmy mogli zidentyfikować stan funkcjonalny mózgu Arta z konkretnym przekonaniem o określonej jednoznacznie treści. A na podstawie wiedzy o działającym programie nie możemy jednoznacznie określić, w jakim stanie mentalnym jest Art, czy też tego, co myśli nasz robot, ponieważ nie potrafimy jednoznacznie wyróżnić rzeczywistej treści jego stanu umysłu spośród możliwych.

6. Skąd wiemy, o czym myślą i co robią inni?

6.1. Davidson

Dlaczego w ogóle jesteśmy skłonni przypisywać Artowi postawy propozycjonalne? Interpretując jakiś ruch Arta jako działanie spowodowane pewnym przekonaniem, przypisujemy mu bardzo złożony system wzajemnie powiązanych stanów i zdarzeń psychicznych. W istocie – mówi Donald Davidson – interpretacja czegoś jako intencjonalnego zachowania zależy od znajomości takiego systemu. Wiedza neurobiologiczna, czy, jak w tym przypadku, wiedza o funkcjonalnej organizacji robota, może być pomocna przy opisie procesów umysłowych, ale w przypadku postaw propozycjonalnych jest niewystarczająca. Nie można w oparciu tylko o nią sporządzić wiarygodnego wyjaśnienia postaw propozycjonalnych wraz z treścią przekonań. Niezbędne wydaje się korzystanie z metod mentalistycznych. Poznanie i wyjaśnianie postaw propozycjonalnych obejmuje empatię, interpretację i znajomość zachowań, w tym językowych. Wymaga więc o wiele bogatszego arsenału środków niż te, jakimi dysponuje inżynier z mojego przykładu. Interpretacja z kolei musi się opierać

na wielu szerokich założeniach dotyczących innych przekonań Arta. Jak pisze Davidson: „Najlepsza droga do dokładnego zidentyfikowania intencji i przekonań wiedzie przez teorię zachowań językowych. (...) Jeśli mam rację, to szczegółowa wiedza o fizyce czy fizjologii mózgu, a w istocie o całym człowieku, nie ułatwiłaby tego rodzaju interpretacji, jakiej wymaga stosowanie wyszukanych pojęć psychologicznych. Ustalenie, co człowiek-maszyna rozumie przez to, co mówi, nie byłoby łatwiejsze od zinterpretowania słów człowieka”[7].

6.2. Wittgenstein

W *Dociekaniach* Wittgenstein porównuje maszynę do czytania z osobą, która czyta[8]. Rzekomo jest między nimi zasadnicza różnica. Polega ona na tym, że tylko w przypadku osoby jest to rzeczywiste czytanie, a to dlatego, że czytanie polegać ma na świadomym przeżyciu psychicznym, „że się czyta”. Takie doznanie nie tyle ma towarzyszyć czynności czytania, ale ją wyznaczać, tj. być kryterium tego, że się czyta. Dlatego też osoba czyta, a maszyna nie. Ale – mówi Wittgenstein – to przeżycie mogłoby występować, a osoba nie czytałaby, lub mogłoby go nie być podczas lektury tej osoby (na przykład byłaby ona pod wpływem trucizny mózgu zmieniającej podłoże przeżyć). Jeśli tak, to przeżycie, „że się czyta”, nie konstytuuje czytania. Nie jest dla niego kryterium. Wydaje się, że to swoiste doznanie może towarzyszyć czytaniu, ale nie musi. Czytanie na nim nie polega. Zatem czy maszyna z *Dociekań* czyta, czy nie czyta? Jeśli nie czyta, to co robi? Dlaczego mielibyśmy sądzić, że nie czyta? Jeżeli czytanie miałoby polegać na świadomym przeżywaniu, to czym jest takie przeżycie? Być może jest to gotowość do skupienia uwagi na tym, co się robi. Dlaczego jednak maszyna nie mogłaby być wyposażona w możliwość sterowania swoją uwaga, kierowania jej na swoje własne stany, i możliwością zapamiętywania tego, porównywania z innymi stanami i raportowania o tym, co robi? Widzieliśmy jednak, że to specyficzne przeżywanie czytania mogłoby nie nastąpić mimo czynności czytania, lub mogłoby się pojawić, kiedy nie czytam. Przeżycie to nie wyznacza, w sensie kryterialnym, czytania. Czytamy niezależnie od tego, czy żywimy takie doznania, czy nie. Jeśli tak, to dlaczego mielibyśmy odmawiać maszynie tej zdolności? Wydaje się, że w omawianym przypadku maszyna czyta, i powinienem to stwierdzić na podstawie obserwacji jej zachowania, tak jak stwierdzam to w przypadku czytających uczniów, którym nie zaglądam przecież do umysłów. Zresztą – parafrazując słowa Wittgensteina – nawet gdybym tam zajrzał, nie mógłbym stwierdzić obecności lub braku doznania czytania. Uwagi Wittgensteina w tej kwestii są

[7] D. Davidson, *Eseje o prawdzie, języku i umyśle*, przeł. B. Stanosz, Warszawa 1992.
[8] L. Wittgenstein, *Dociekania filozoficzne*, dz. cyt., § 158 i n.

niejednoznaczne i zastanawiające. Z jednej strony, w *Dociekaniach* czytamy, że: „Zmiana, która nastąpiła, była zmianą jego zachowania" i „Czy nie polega to tylko na naszej niedostatecznej znajomości zjawisk zachodzących w mózgu i układzie nerwowym. Gdybyśmy je znali lepiej, to dostrzeglibyśmy, jakie połączenia wytworzone zostały w wyniku ćwiczenia, a wtedy zajrzawszy do mózgu, moglibyśmy rzec: «Teraz przeczytał ów wyraz, teraz powstało połączenie, na którym polega czytanie»", a nieco wcześniej, odnośnie maszyny: „Dopiero gdy to a to w maszynie zrobiono – te a te części zostały połączone – zaczęła ona czytać"[9]. Może więc podstawę zjawiska psychicznego, jakim jest czytanie, stanowią stany mózgu, lub w przypadku maszyny, fizyczne stany sztucznego systemu? I jeśli tak, to uda się wyjaśnić zjawisko psychiczne, jakim jest czytanie, poprzez sprowadzenie go do pewnego zjawiska neurofizjologicznego? Takie uwagi mogą sugerować kryterialną rolę stanów fizycznych odpowiedzialnych za stany umysłowe i odpowiednio kryterialną rolę pojęć fizykalistycznych w kwestii opisu i wyjaśniania życia umysłowego, do którego należy między innymi przypisywanie i interpretowanie myśli i pragnień innych oraz rozpoznawanie własnych. Stanowisko Wittgensteina jest dla mnie tyleż niejasne, co pociągające. Aby zilustrować swoje zagubienie, zacytuję jeszcze dwa fragmenty *Dociekań:* „Gdyby Bóg zajrzał do mojej głowy, nie wiedziałby, o czym myślę" i „Mogłoby się okazać, że podczas operacji otworzono by moją czaszkę i byłaby ona pusta". Co to może znaczyć? Uwaga ta odnosi się do naszych interpretacji tego, co myślą i mówią inni za pomocą pojęć fizykalnych (i funkcjonalnych), a raczej do nieistotności takiej interpretacji. Być może Wittgenstein ma na myśli to, że pojęcia te nie grają żadnej roli w naszych charakterystykach stanów mentalnych innych ludzi. Aby określić, co myślą inni (także roboty), stosujemy kryteria behawioralne, a żeby określić i powiedzieć, co myślimy my sami, nie jest nam potrzebna wiedza o mózgu i być może nie są nam potrzebne żadne kryteria. Wedle Wittgensteina, predykaty psychologiczne są związane z zachowaniem. Lecz w jaki sposób? Wydaje się, że wypowiedzi psychologiczne nie są równoważne behawioralnym, ale określenia stanów psychicznych zawierają ich behawioralne objawy, które być może pełnią rolę warunków, pod jakimi możemy komuś przypisać pewien stan mentalny. Nie znaczy to, że stany umysłowe są identyczne ze stanami neurofizjologicznymi (lub w przypadku robota, ze stanami jakiegoś fizycznego realizatora umysłu), nie znaczy to też, że między nimi zachodzi związek przyczynowo-skutkowy, tj. w oparciu o domniemania przyczynowości stanów mózgu wobec stanów mentalnych i ich manifestacji w zachowaniu nie można poprawnie wyjaśnić, czyli zredukować, umysłu. Jednak nie znaczy to również, że można sensownie oddzielać stany psychiczne od stanów fizycznych i behawioralnych.

[9] L. Wittgenstein, *Dociekania filozoficzne*, dz. cyt., § 157 i 158.

7. Paradoksalna konkluzja

Davidson i Witgenstein wskazują na problem, jaki ma każda naturalistyczna teoria umysłu lub program badawczy o charakterze naturalistycznym, także kognitywistyczny. Kłopot, o jakim mówię, przypomina zagadkę sceptyka. Jeżeli nie potrafię określić, co robi program w czarnej skrzynce, nie potrafię też powiedzieć, co robi czarna skrzynka (i robot), ponieważ ich zachowanie da się uzgodnić z więcej niż jedną instrukcją. Podobny problem ma Kuttnerowski naukowiec. Zaglądanie do wnętrza czarnej skrzynki nie pomoże mu w odpowiedzi na pytanie: do czego to służy? Nie dość, że zaglądanie do niej nic nie da, to nie jest potrzebne, bo inżynier-konstruktor nie jest w lepszym położeniu niż obserwator albo rozmówca robota.

Na zakończenie chcę powiedzieć, że nie podoba mi się ta konkluzja. Na pewno się gdzieś pomyliłem. Tylko gdzie?

Bibliografia

Bobryk J., *Locus umysłu*, Zakład Narodowy im. Ossolińskich, Wrocław 1987.

Davidson D., *Eseje o prawdzie, języku i umyśle*, przeł. B. Stanosz, Wydawnictwo Naukowe PWN, Warszawa 1992.

Wittgenstein L., *Dociekania filozoficzne*, przeł. B. Wolniewicz, Wydawnictwo Naukowe PWN, Warszawa 2000.

Streszczenie

W jednym z opowiadań Henry'ego Kuttnera występuje wybitny badacz i wynalazca, który wpada na pomysły tylko wtedy, gdy jest kompletnie pijany. Kiedy się budzi, spostrzega, że znowu coś wymyślił i skonstruował, ale nie ma pojęcia, czym to właściwie jest i co robi. Pod pewnym względem sytuacja kognitywisty jest podobna do zagubienia Kuttnerowskiego bohatera. Kognitywista ma nadzieję na wyjaśnienie, jakie myśli kryje czarna skrzynka lub o czym myśli jego robot. Ale zaglądanie do czarnej skrzynki nic nie da, a nawet nie jest potrzebne, bo inżynier-konstruktor nie jest w lepszym położeniu niż obserwator albo rozmówca robota. Nie podoba mi się ta konkluzja. Na pewno gdzieś się pomyliłem. Tylko gdzie?