Semina □ Nr 3
Scientiarum 2004

Ryszard Philipp

# Twierdzenia Gödla a niemechaniczność umysłu (cz. I)

Twierdzenia Gödla, jak również inne twierdzenia z grupy tzw. twierdzeń limitacyjnych, stwarzają możliwość formalizacji sporu o naturę umysłu. Stanowią one mianowicie pewnego rodzaju test, który powinny przejść teorie mechanistyczne[1]. Schemat argumentacji antymechanistycznej z wykorzystaniem twierdzeń limitacyjnych sprowadza się zazwyczaj do następującego rozumowania. Punktem wyjścia jest odnotowanie pewnych, stosunkowo precyzyjnie ustalonych faktów, związanych z działalnością umysłu, mianowicie wytworów umysłu w dziedzinie matematyki. Gdyby umysł był maszyną, maszyna taka powinna być równoważna umysłowi co najmniej pod względem tych wytworów. Antymechanicyści próbują więc wykazać, że maszyny, zgodnie z twierdzeniami limitacyjnymi, podlegają ograniczeniom, podczas gdy umysł takim ograniczeniom nie podlega, o czym świadczą właśnie faktyczne wytwory umysłu. W związku z tym maszyna nie jest w stanie wytworzyć tego, co potrafi wytworzyć umysł i w konsekwencji umysł nie może być maszyną[2]. W omawianym w dalszej części tekstu tzw. argumencie Lucasa, autor argumentu próbuje wykazać, że zgodnie

[1] Por. J. Woleński, „Metamatematyka a filozofia", *Zagadnienia Filozoficzne w Nauce*, 6 (1984), 10; J. Życiński, „Metafilozoficzne następstwa twierdzeń limitacyjnych", *Studia Philosophiae Christianae*, 24 (1988), 1.

[2] Warto zauważyć, że wnioskowanie odwrotne nie jest oczywiste, jak sądzą być może zwolennicy tzw. *testu Turinga*. To, że maszyna byłaby równoważna umysłowi pod względem pewnych lub nawet wszystkich wytworów, nie musi oznaczać, że umysł jest maszyną. Wytworom umysłu mogłyby przykładowo towarzyszyć pewne subiektywne przeżycia (tzw. *qualia*), które nie towarzyszyłyby wytworom maszyny.

z twierdzeniami Gödla jest niemożliwe, aby jakakolwiek maszyna Turinga zweryfikowała wszystkie te twierdzenia arytmetyczne, które umysł ludzki jest w stanie uznać za prawdziwe, co ma właśnie obalić mechanistyczną koncepcję umysłu.

Tradycyjne stanowisko antymechanistyczne w kwestii umysłu wyraża się tym, że uważa się umysł, *resp.* duszę, za byt ontologicznie odrębny od świata zmysłowego. Takie stanowisko wyklucza mechaniczność umysłu *per definitionem*, ponieważ znaczenie terminu mechanizm czy automat sugeruje, że jest to obiekt deterministyczny, należący do świata fizykalnego, podczas gdy duszę uważa się za niematerialną i wolną.

Stanowisko takie określa się tradycyjnie mianem dualizmu. Klasycznymi jego przedstawicielami są Kartezjusz oraz Platon. Stanowisko Platona jest o tyle szczególne, że wielu współczesnych matematyków skłania się w kwestii istnienia obiektów matematycznych do pewnej wersji platonizmu. Platonizm matematyczny przypisuje obiektom matematycznym byt w sensie idei platońskich, z którymi umysł ludzki może mieć bezpośredni (intuicyjny) kontakt. W takim świecie każda formuła matematyczna jest albo prawdziwa albo fałszywa.

Dualizm kartezjański stał się ostatnio etykietą dla wszystkich stanowisk antymechanistycznych, choć jest to nieco nieżyczliwa interpretacja stanowiska Kartezjusza[3]. Kartezjusz zapoczątkował w nowożytnej filozofii tendencję radykalnego odróżniania umysłu i faktów mentalnych od ciała i świata zmysłowego. Jeżeli ktoś, jak to czynili oświeceniowi materialiści, zaprzeczał dualizmowi Kartezjusza, przejmował na siebie obowiązek dowodu obalającego. Współcześnie większość filozofów opowiada się przeciwko dualizmowi w sensie kartezjańskim. Przyjmuje się, niecałkiem słusznie, że ciężar dowodu spoczywa na tych, którzy próbują utrzymać nieredukowalność umysłu do materii. Wymaga tego, by tak rzec, „polityczna poprawność". Jest to o tyle nie *fair*, że zarówno zwo-

[3]Por. J. R. Searle, *Umysł na nowo odkryty*, PIW, Warszawa 1999.

lennicy jak i przeciwnicy mechanicyzmu posługują się w zasadzie argumentami preferencyjnymi o, mniej więcej, podobnej sile.

Spór o umysł wydaje się bardziej interesujący na gruncie naturalizmu, niż w paradygmacie klasycznego dualizmu. Zjawiska mentalne uważa się wtedy za pochodne względem zjawisk fizycznych, choć dopuszcza się niekiedy istnienie subiektywnej, nieredukowalnej do obiektywnego poziomu fizykalnego, sfery psychicznej. Stanowisko takie jest określane mianem emergentyzmu. Koncepcja sztucznej inteligencji (zwana czasem silną AI) jawi się w tym kontekście, jako stanowisko skrajnie mechanistyczne i optymistyczne ze względu na wiarę w możliwości nauki[4]. Bardziej umiarkowani naturaliści, na przykład Searle, zgadzają się z tym, że umysł jest quasi–fizykalnym systemem przejawiającym funkcje komputacyjne, niemniej twierdzą, że globalne możliwości umysłu przekraczają możliwości komputerów w sensie von Neumanna. Co więcej, uważa się również, że żaden komputer, rozumiany jako **skończony automat** (ang. *Finite State Machine*), nie osiągnie poziomu istotnie inteligentnych zachowań nawet w przyszłości, bez względu na intensyfikację parametrów, takich jak rozmiar pamięci, częstotliwość taktowania, etc. Dotyczy to również tzw. sztucznych sieci neuronowych, jeżeli przyjmiemy, że także w przyszłości będą działać w sposób deterministyczny, przy czym ich determinizm można wyrazić formułą „*output* równa się *input*" — tyle dostaniemy na wyjściu, ile sami włożymy do systemu na wejściu.

***

Mniej więcej czterdzieści lat po sformułowaniu przez Gödla jego twierdzeń o niezupełności pojawiła się argumentacja J. R. Lucasa[5]. Uzasadnia on tezę, że umysł przewyższa maszynę, czyli,

[4]Zwolennicy *silnej AI* głoszą, że umysł jest automatem, który w zasadzie równoważny jest zwykłym komputerom.

[5]J. R. Lucas, „Minds, Machines and Gödel", *Philosophy* 36, 112–127. Por. również J. R. Lucas, *The Freedom of the Will*, Oxford Univ. Press, Oxford 1970.

wbrew zwolennikom sztucznej inteligencji (AI), nie jest komputerem. Argumentację tę można streścić następująco: twierdzenie Gödla mówi, że w każdej, odpowiednio bogatej (zawierającej arytmetykę liczb naturalnych) i niesprzecznej teorii, istnieje zdanie o liczbach naturalnych, które nie jest w tej teorii dowiedlne, niemniej (w metateorii) można pokazać, że zdanie to jest prawdziwe. Efektywnie dana teoria, tzn. oparta o rozstrzygalny zbiór aksjomatów, a więc również maszyna w sensie Turinga, nie jest w stanie dowieść zdania Gödla, o ile ma pozostać niesprzeczna. Umysł natomiast umie pokazać prawdziwość tego zdania w modelu danej teorii, przy czym zakłada się, jako oczywiste — przynajmniej w intuicyjnym sensie, że umysł jest niesprzeczny. Umysł potrafi więc zweryfikować zdanie Gödla, podczas gdy maszyna tego nie potrafi, co świadczy o tym, że umysł nie może być maszyną. Gdy dopuścimy sprzeczność maszyny, to może ona, zgodnie z prawami logiki, dowieść dowolnego zdania, wtedy jednak nie chcemy uważać jej za równoważną umysłowi[6].

Formalna krytyka argumentów Lucasa oraz innych, podobnie myślących autorów[7], opiera się na dwojakiej strategii. Pierwsza grupa argumentów zmierza do wykazania, że na podstawie samych tylko twierdzeń Gödla nie da się wykluczyć tego, iż możemy być maszynami (niesprzecznymi). Wśród zwolenników tego typu argumentacji wymienić można m. in. samego Gödla, Wanga oraz Benacerrafa. Druga grupa autorów, na przykład Putnam, uważa z kolei, że nie da się wykluczyć sytuacji, iż możemy być maszynami, aczkolwiek sprzecznymi, ponieważ to, iż umysł umie zweryfikować zdanie Gödla, może wynikać właśnie z tego, że jest sprzeczny.

Aby rozważyć argumenty metalogiczne przeciwko stanowisku Lucasa, niezbędna jest dokładniejsza analiza założeń jego rozumowania. Przede wszystkim należy sprecyzować, co będziemy uważać

[6]Nie bierze się tutaj pod uwagę zarzutu, że to, iż umysł umie zweryfikować zdanie Gödla, może wynikać właśnie z tego, iż jest sprzeczny.

[7]Por. R. Penrose, *Nowy umysł cesarza*, PWN, Warszawa 1995 oraz R. Penrose, *Cienie umysłu*, Zysk i S–ka, Poznań 2000.

za maszynę. Intuicyjnie, maszyną jest skończony automat działający w sposób algorytmiczny. Zakłada się więc, że maszyna musi mieć skończoną liczbę możliwych stanów, a jej program musi dać się wyrazić skończonym tekstem. Pomimo tych, zdawałoby się silnych założeń, wgląd w możliwości maszyny nie jest trywialny. Nawet odpowiedź na pozornie proste pytanie, czy maszyna zakończy pracę w skończonym czasie, przerasta możliwości każdej maszyny. Innymi słowy, nie istnieje efektywna procedura rozstrzygająca to pytanie dla dowolnej maszyny Turinga.

Ważną cechą maszyn jest determinizm, tzn. działanie maszyny jest w pełni opisane przez jej program, dane wejściowe oraz stan w jakim się znajduje. Można więc przyjąć, że maszyna puszczona w ruch w tej samej konfiguracji, będzie zawsze działać identycznie. Z punktu widzenia obserwatora zewnętrznego czas maszyny jest dyskretny, maszyna „wie" jedynie, że zawsze znajduje się w jakimś stanie wewnętrznym. Pytanie o to, jak długo trwa przejście od jednego stanu drugiego, jest z punktu widzenia maszyny bezsensowne. Na koniec wypada zgodzić się co do tego, że jeśli umysł jest w stanie wykonać pewne operacje w skończonym czasie, to maszyna, o ile ma być równoważna umysłowi, musi wykonać te same operacje, względnie osiągnąć te same wyniki, również w skończonym czasie, aczkolwiek dowolnie długim.

Matematyczną formalizacją opisanego powyżej automatu jest tzw. maszyna Turinga, która jest idealizacją „zwykłego" komputera i jest mu równoważna pod względem efektów działania[8]. Takie rozumienie automatu pokrywa się z intuicyjnym pojęciem algorytmu. Zgodnie z tzw. *tezą Churcha*, intuicyjne pojęcie algorytmu nie jest szersze od matematycznego pojęcia maszyny Turinga. Z kolei klasa obliczanych przez nią funkcji równoważna jest tzw.

[8]Dotyczy to również *sztucznych sieci neuronowych*. Zachowanie takiej sieci można symulować na zwykłym komputerze. Nie jest jednak oczywiste, czy sieci takie nie „wymkną się" z ograniczeń determinizmu, gdy zaczną wykorzystywać ewentualne niedeterministyczne efekty, które, jak uważają niektórzy, są możliwe na gruncie współczesnych lub przyszłych teorii fizycznych (por. R. Penrose, *Nowy umysł cesarza*, op. cit.).

klasie funkcji rekurencyjnych[9]. Inaczej mówiąc, wszystko to, co według naszych intuicji można osiągnąć w sposób automatyczny, można również osiągnąć używając odpowiednio zaprogramowanej maszyny Turinga. Turing pokazał, że nawet wprowadzenie elementu quasi–indeterministycznego, np. w postaci losowania zachowania maszyny w następnym kroku ze skończonej liczby możliwości, nie wyprowadza poza zwykłe, deterministyczne maszyny.

Maszyna Turinga jest równoważna pewnemu systemowi formalnemu, wyrażonemu w języku I rzędu. System taki jest dany przez zbiór aksjomatów i reguł dowodzenia, co do którego zakładamy, że jest rozstrzygalny[10], tzn. istnieje efektywna procedura, pozwalająca odróżnić aksjomaty od pozostałych wyrażeń języka. W tak zdefiniowanym języku zbiór wszystkich formuł, które można udowodnić w oparciu o aksjomaty i reguły tego systemu, jest efektywnie przeliczalny[11], tzn. istnieje algorytm, który umie rozpoznać dowody formalne wśród wszystkich (skończonych) tekstów tego języka. Nie jest natomiast efektywnie przeliczalny zbiór formuł nie będących twierdzeniami systemu. Inaczej mówiąc, jeżeli dowolna formuła jest twierdzeniem systemu, to maszyna potrafi znaleźć dowód tej formuły w skończonym czasie. Maszyna nie potrafi natomiast orzec w skończonym czasie o dowolnej formule, że nie posiada ona dowodu w tym systemie. W dalszych rozważaniach maszyna będzie *per definitionem* uważana za niesprzeczną, gdy odpowiadająca jej teoria elementarna jest niesprzeczna.

Gödel pokazał, że dowolny tekst (wyrażenie sensowne), utworzony w języku zbudowanym w oparciu o logikę elementarną (I rzę-

[9]Istnieją jeszcze inne równoważne matematyczne definicje pojęcia obliczalności, np. $\lambda$–rachunek Churcha. Fakt istnienia wielu, powstałych niezależnie od siebie, równoważnych definicji obliczalności uważany jest za dobre uzasadnienie *tezy Churcha*.

[10]Zgodnie z twierdzeniem Craiga założenie to można osłabić. Wystarczy, aby zbiór aksjomatów był efektywnie przeliczalny.

[11]Dowód jest to na mocy definicji skończony ciąg formuł, z których każda jest albo aksjomatem, albo przesłanką, albo powstaje z poprzednich na mocy reguł dedukcji oraz kończący się formułą, która ma być dowiedziona.

du), ze skończonym zbiorem symboli pozalogicznych[12], można w efektywny sposób zakodować, w sposób jedno–jednoznaczny, liczbą naturalną. Tak więc wszystkich tekstów, utworzonych z wyrażeń tego języka nie może być więcej niż liczb naturalnych. Efektywna procedura odwrotna pozwala odtworzyć teksty na podstawie ich kodów, przy czym zbiór wszystkich kodów tekstów danego języka jest rozstrzygalny (procedura „umie" stwierdzić czy dana liczba naturalna jest kodem jakiegoś tekstu, czy nie). Ponieważ dowolna dedukcja, będąca na mocy definicji skończonym ciągiem wyrażeń języka, jest również tekstem, zatem liczba naturalna przyporządkowana tej dedukcji, zwana numerem Gödla, musi znaleźć się na liście wszystkich liczb naturalnych. Można ponadto w sposób efektywny stwierdzić czy dany tekst jest poprawnym dowodem, czyli wszystkie dowody formalne można efektywnie ustawić w ciąg. Możemy zatem rozważania o systemach formalnych sprowadzić do rozważań o liczbach naturalnych — zamiast o formułach wystarczy mówić o liczbach. Ponadto pewne interesujące własności metasystemowe, jak np. dowodliwość, dają się wyrazić poprzez relacje zachodzące między liczbami naturalnymi. W szczególności ograniczenia formalne arytmetyki będą „przenosiły się" na wyrażony w niej system[13].

---

[12]Założenie to również można osłabić, zbiór symboli pozalogicznych może być nieskończony, wystarczy, by był rozstrzygalny, por. R. Murawski, op. cit., ss. 84 nn.

[13]Warto zwrócić w tym miejscu uwagę na pewien interesujący fakt. Otóż jeżeli zgodzimy się, że każdy algorytm musi dać się wyrazić przy pomocy skończonego tekstu, to wynika z tego, iż wszystkich algorytmów nie może być więcej niż liczb naturalnych, czyli przeliczalnie wiele. Również wszystkich maszyn Turinga jest nie więcej niż przeliczalnie wiele. Każdy algorytm jest formalnie równoważny pewnej funkcji naturalnej ($\mathbb{N} \longmapsto \mathbb{N}$), wiadomo jednak, że wszystkich funkcji naturalnych jest, zgodnie z twierdzeniem Cantora, nieprzeliczalne wiele, czyli istotnie więcej. Wynika stąd, że pewne funkcje naturalne nie dają się obliczyć w sposób algorytmiczny. Pewne procesy myślowe (prowadzące od pewnych tekstów do innych tekstów, a zatem dające się pomyśleć jako funkcje naturalne) nie mogą być więc przeprowadzone w sposób algorytmiczny.

Maszyna Turinga, aby być modelem umysłu, musi być dostatecznie „bogata". Żądamy co najmniej tyle, by zawierała ona arytmetykę, czyli była równoważna umysłowi pod względem zdolności arytmetycznych. Maszyna taka powinna dla przykładu umieć odpowiadać na pytanie, czy konkretna formuła arytmetyczna jest dla niej twierdzeniem, czy nie. Moglibyśmy umówić się, że formuła jest dowiedziona przez maszynę wtedy, gdy po wprowadzeniu danej formuły „na wejście" formule tej będzie towarzyszyć jakiś ustalony znak „na wyjściu", np. zapalenie się zielonej lampki. Wszystkie (i tylko te) formuły, które pojawią się na wyjściu systemu w obecności zielonego światła, mogą być uznane za twierdzenia maszyny.

***

Traktując tezę o mechaniczności umysłu poważnie, można próbować uważać umysł za efekt ewolucji prostszych systemów qusi–inteligentnych. Mechanizm nie obdarzony świadomością może „stwierdzić" jedynie proste fakty, przy czym nie jest w stanie dokonać aksjologicznej oceny owych faktów. Powtarzające się, długie czasy poszukiwania rozwiązań, mierzalne „zegarem biologicznym" systemu, mogły spowodować pojawienie się dodatkowych reguł działania, początkowo nieświadomych, pozwalających reagować w sytuacjach przedłużających się poszukiwań. Jednym z takich mechanizmów mógł być mechanizm pozwalający oszacować czas poszukiwań, aby zareagować w sytuacji niekorzystnej z punktu widzenia egzystencji systemu. Jest to całkiem rozsądna hipoteza — systemy biologiczne w przyrodzie podlegają obiektywnej presji co do skuteczności swoich działań w sytuacjach decyzyjnych. Mechanizmy podejmowania decyzji w przypadkach nierozwiązywalnych, a taką byłoby właśnie dla systemu formalnego zdanie niezależne, są niezbędne dla przetrwania systemu. Jednym z istotnych mechanizmów, być może jednym z kamieni milowych na drodze do uzyskania samoświadomości, mogłaby być umiejętność oceniania możliwych rozwiązań z punktu widzenia dobra systemu przeciwstawionego otaczającej go rzeczywistości. Oceniając pewną sytu-

ację jako krytyczną (ze względu na czas poszukiwania decyzji), system przechodziłby do rozważań metajęzykowych nad samą tą sytuacją. W naszym przypadku taką sytuacją krytyczną dla systemu są właśnie zdania Gödla ($G_T$). Mechanicysta będzie argumentował, że metajęzykowe rozwiązanie problemu dowodliwości zdań $G_T$ nie wymaga pozamechanicznych zdolności, lecz powstało w wyniku przystosowania się organizmów (algorytmów) do rzeczywistości.

Jednym z kluczowych punktów w sporze o konsekwencje twierdzeń Gödla dla filozofii umysłu jest problem niesprzeczności, zarówno umysłu jak i maszyny. Maszyna, twierdzi się, aby dorównywać umysłowi musi być niesprzeczna, ponieważ umysł jest niesprzeczny, choć argumenty za niesprzecznością umysłu mogą być tylko nieformalne[14]. Z drugiej strony, każdy system, w którym istnieje zdanie niedowodliwe, nie może być formalnie sprzeczny, ponieważ, przez kontrapozycję, gdyby był sprzeczny, to nie istniałoby zdanie formalnie niedowodliwe w tym systemie. W rozważanej argumentacji za wyższością umysłu postępujemy jednak odwrotnie. Chcemy mianowicie wykazać, że właśnie umiemy udowodnić zdanie Gödla, niedowodliwe w rozważanym systemie formalnym, reprezentowanym przez maszynę. Musimy zatem niezależnie założyć naszą niesprzeczność. Aby argumentacja była konkluzywna, trzeba nie tylko wykazać, że umysł potrafi dowieść zdania Gödla, lecz dodatkowo tego, że jest niesprzeczny. Jak wynika z drugiego twierdzenia Gödla, żadna maszyna nie jest w stanie dowieść własnej niesprzeczności, ponieważ nie jest możliwe dowiedzenie niesprzeczności systemu formalnego w tym samym systemie. Ponieważ umysł również nie potrafi formalnie dowieść własnej niesprzeczności, nie możemy w punkcie wyjścia rozważań wykluczyć, że jest sprzeczną maszyną. Brak formalnego dowodu na niesprzeczność umysłu ogranicza więc rolę twierdzeń Gödla w rozważanym sporze.

[14]Możliwy jest jedynie formalny dowód niesprzeczności pewnego fragmentu umysłu w oparciu o środki „wykraczające" poza ten fragment.

Intuicyjnie nasza niesprzeczność wydaje nam się oczywista — nie uznajemy przecież dowolnego stwierdzenia wierząc, że przynajmniej niektóre z nich są nieprawomocne, nieprawdziwe, czy w jakikolwiek inny sposób zdyskwalifikowane. Z drugiej strony, często zdarzają się przypadki, i to nie tylko w codziennym użyciu umysłu lecz również w praktyce matematycznej, że niedozwolone wnioskowania nie od razu są rozpoznawane[15]. Niemniej jednak postulat niesprzeczności jest „ideą regulatywną". Wydaje się, że umysł może być sprzeczny najwyżej potencjalnie — nigdy aktualnie, w sposób jawny, przynajmniej w matematyce. Matematycy deklarują, że będą zwalczać sprzeczność zarówno w swoich poglądach, jak i w poglądach innych, co więcej, jak nikt inny się do tego stosują. Być może sama ta deklaracja wystarczyłaby już jako nieformalny dowód niesprzeczności umysłu gdyby nie to, że nie można wykluczyć sytuacji, w której rozpoznanie sprzeczności okaże się niewykonalne z powodu komplikacji teorii[16]. Wiele osób głosi, świadomie lub nie, sprzeczne poglądy, nic sobie z tego nie robiąc. Matematyk głoszący sprzeczne poglądy zostałby zdyskwalifikowany przez społeczność matematyków. Matematycy przyjmują za oczywistą tezę, iż sprzeczność umysłu powinna oznaczać eksterminację również w każdej innej dziedzinie. Przejawia się tutaj *implicite* religijne wręcz przekonanie, iż „świat preferuje niesprzeczne umysły". Teza taka może być jednak potwierdzona najwyżej w sposób empiryczny, tzn. ewentualnie przez fakt, że interakcja umysłów z rzeczywistością fizykalną odbywa się w sposób niesprzeczny[17].

Choć wielu autorów sądzi inaczej, nie można wykluczyć, że obowiązuje nas pewnego rodzaju psychologiczna zasada nie-

[15]Wystarczy wspomnieć choćby historię dowodu tzw. wielkiego twierdzenia Fermata, zob. A. D. Aczel, *Wielkie twierdzenie Fermata. Rozwiązanie zagadki starego matematycznego problemu*, Prószyński i S–ka, Warszawa 1998.

[16]Por. Krajewski, op. cit., s. 111.

[17]W sferze ogólnie pojętej kultury rzeczywistość weryfikuje raczej, jak się wydaje, umiarkowaną zdolność do sprzeczności w przekonaniach jako korzystniejszą z punktu widzenia sukcesu ewolucyjnego. Zdolność tę próbuje się ostatnio określać mianem inteligencji emocjonalnej.

sprzeczności[18]. Teza taka może mieć oczywiście charakter jedynie empiryczny. Niemożliwość istnienia dwóch sprzecznych przekonań w tym samym czasie, w jednym umyśle, może mieć więc przyczyny poza(onto)logiczne. Niektórzy uważają[19], że mogą równocześnie mieć dwa sprzeczne przekonania w swoim umyśle, choć dla innych wydaje się to niemożliwe. Niełatwo jest zweryfikować stan posiadania takich dwóch sprzecznych przekonań, ponieważ nie jest to fakt intersubiektywnie komunikowalny. Należy podkreślić, że czym innym jest mieć dwa bezpośrednio sprzeczne przekonania, a czym innym jest rozważanie dwóch sprzecznych przekonań[20]. Ponadto sprzeczność może być mniej lub bardziej bezpośrednia. Niewiele osób jest prawdopodobnie skłonnych przyjąć, że dwa razy dwa równa się cztery i równocześnie, że dwa razy dwa nie równa się cztery, lecz wielu ludzi przyjmuje za słuszne tezy, których uzasadnienie opiera się na błędnym wnioskowaniu, przez co wspomniane tezy mogą pozostawać w sprzeczności z innymi wcześniej uzyskanymi twierdzeniami. W tym drugim przypadku sprzeczność jest, można powiedzieć, „zapośredniczona" i w pewnym sensie nieświadoma. Obiektami dobrze nadającymi się do testowania umiejętności posiadania sprzecznych przekonań mogą być antynomie, na przykład jakieś szczególnie drastyczne sformułowanie antynomii kłamcy lub któraś z kantowskich antynomii czystego rozumu[21]. Nie jest wykluczone, że zdolność posiadania niesprzecznych przekonań jest stopniowalna (może przejawiać się w różnym stopniu u róż-

[18]Rozróżnienie na ontologiczną, logiczną i psychologiczną zasadę niesprzeczności pochodzi od Łukasiewicza, por. J. Łukasiewicz, *O zasadzie sprzeczności u Arystotelesa*, PWN, Warszawa 1987.

[19]Przykładowo Łukasiewicz, por. op. cit.

[20]Tak, jak czym innym jest logika parakonsystentna a czym innym metajęzyk, w którym o niej mówimy.

[21]Antynomie czystego rozumu Kanta dotyczą kwestii kosmologicznych. Dla przykładu pierwszą antynomię tworzy następująca para twierdzeń: „świat posiada początek w czasie, a przestrzennie jest również ograniczony" oraz „świat nie ma początku i nie ma granic w przestrzeni, lecz jest nieskończony zarówno co do czasu, jak i przestrzeni", I. Kant, „Antynomia czystego rozumu" w: *Krytyka czystego rozumu*, przekład R. Ingarden, Antyk, Kęty 2001.

nych ludzi) i uwarunkowana neurofizjologiczną strukturą mózgu. Jeśli zgodzimy się, że przynajmniej w interakcji z rzeczywistością fizykalną, posiadanie niesprzecznej reprezentacji tej rzeczywistości sprzyja sukcesowi ewolucyjnemu jednostki, można oczekiwać, iż w mózgach zarówno człowieka jak i zwierząt wykształciły się odpowiednie struktury sprzyjające niesprzecznemu myśleniu.

Współczesne badania nad mózgiem potwierdzają tezę, iż praca mózgu wiąże się ze zmianą rozkładu pewnej mierzalnej wielkości fizycznej. Gdyby przyjąć, że myślenie jest uwarunkowane stanami mózgu, procesy myślowe mogłyby być zdefiniowane przez zmiany tej wielkości, na przykład przez funkcję potencjału, analogicznie do opisu rozkładu potencjału elektrycznego w pewnej przestrzeni (określony obszar mózgu mógłby być rozpatrywany jako wirtualne źródło potencjału). Jest prawdopodobne, że dwóm sprzecznym przekonaniom odpowiadałyby antysymetryczne rozkłady potencjałów. Posiadanie dwóch sprzecznych przekonań byłoby więc z energetycznego punktu widzenia stanem wysoce niestabilnym. Nie jest wykluczone, że mózgi mogą różnić się pewnymi parametrami, analogicznie jak dwa elektryczne źródła napięcia różnią się oporem wewnętrznym. Mózgi charakteryzujące się wysokim „oporem wewnętrznym" byłyby bardziej stabilne, bardziej odporne na krytyczne połączenia, w tym wypadku również na sprzeczność. Duża odporność na sprzeczność mogłaby powodować jednak silne osłabienie „sygnału", a w związku z tym utrudniać przeprowadzanie długich wnioskowań. Różnice indywidualne, związane ze zdolnościami przeprowadzania rozumowań matematycznych, są oczywistym faktem empirycznym. Każdy chyba, kto zajmował się matematyką, a w szczególności analizował dowody matematyczne, wie dobrze, że rozumienie poszczególnych kroków dowodu to nie to samo, co rozumienie dowodu jako całości. W oparciu o powyższy model działania mózgu można zaryzykować hipotezę, że *rozumienie* pewnego wnioskowania wymaga *jednoczesnego* i *stabilnego* uaktywnienia odpowiednich obszarów mózgu, odpowiadających poszczególnym krokom dowodowym. Zdolność powtórnego

przeprowadzenia dowodu wymagałaby odtworzenia odpowiedniej struktury połączeń. Sprzeczność we wnioskowaniu powodowałaby niestabilność rozumowania. Mózgi cechujące się większą odpornością na sprzeczność nie byłyby zdolne do przeprowadzania długich wnioskowań, ponieważ sygnał mógłby być silniej tłumiony podczas przejścia przez kolejne obszary i dla długich wnioskowań brakowałoby energii. Warto zauważyć, że *myślenie* o dwóch sprzecznych przekonaniach, w przeciwieństwie do *posiadania* dwóch sprzecznych przekonań, nie musi w powyższym modelu prowadzić do niestabilności. Na zakończenie tych rozważań należy podkreślić, że są to jedynie dość ostrożne spekulacje, niemniej dzisiejsze metody badania mózgu[22] stwarzają warunki, by taką hipotezę uczynić przedmiotem sensownego programu badawczego.

***

Zakładając, że badany system formalny jest niesprzeczny, można, na mocy twierdzenia Gödla, wskazać formułę $G_T$, która nie jest dowiedlna w tym systemie. Wykluczona jest zatem sytuacja, że maszyna zweryfikuje zdanie $G_T$, pojawiające się na jej wyjściu[23]. Nie wypowiadamy żadnych uwag na temat tego, czy maszyna rozumie zdanie $G_T$, czy nie, jak również nie wymagamy od niej deklaracji co do prawdziwości zdań. Jest to o tyle istotne, że zdanie $G_T$ jest tak skonstruowane, że mówi o sobie, iż nie jest dowiedlne w rozważanym systemie formalnym, czyli fakt jego niedowodliwości świadczy o jego prawdziwości. Umysł uznaje to za oczywiste, w oparciu o rozumienie znaczenia tego zdania[24]. Ponieważ maszyna tego nie potrafi, ma to świadczyć o jej podrzędności w stosunku

[22]Na przykład metoda rezonansu magnetycznego.

[23]Twierdzenie Gödla mówi nawet więcej. Przy pewnych dodatkowych założeniach (które nie są konieczne, co wykazał Rosser, J. Rosser „Extensions of Some Theorems of Gödel and Church", *Journal of Symbolic Logic*, 1 (1936), 87–91, por. również R. Murawski, op. cit., s. 94), maszyna nie zweryfikuje również zdania $\neg G_T$.

[24]Prawdziwość zdania Gödla można uzasadnić bardziej formalnie. Jeżeli badany system formalny jest niesprzeczny, to na mocy twierdzenia o pełności, ma model, w którym każde zdanie, należące do języka tego systemu, jest bądź

do umysłu — umiejętność wykraczania poza system ma świadczyć o niemechaniczności umysłu. Należy tutaj zwrócić uwagę na pewną istotną rzecz. Jeżeli rzeczywiście zawieszamy tezę o niemechaniczności umysłu na potrzeby argumentacji, to samo przypisywanie umysłowi wglądu w prawdziwość zdania Gödla należy traktować ostrożnie. Przede wszystkim zdolność orzekania prawdziwości może być spowodowana faktem, iż umysł jest sprzeczny. Jeżeli umysł nie jest sprzeczny, to prawdziwość zdania Gödla wynika rzeczywiście ze spojrzenia na zdanie Gödla z poziomu metajęzyka i odwołania się do semantyki. Prawdziwość zdania Gödla nie może być jednak rozumiana absolutnie, ponieważ nie we wszystkich modelach, w których prawdziwe są wszystkie inne twierdzenia produkowane przez maszynę, zdanie to jest prawdziwe. To, że umysł uznaje za prawdziwe zdanie $G_T$ a nie jego zaprzeczenie, jest ściśle rzecz biorąc arbitralne — może wynikać przykładowo z powodów pozalogicznych (np. pragmatycznych, rozumianych w kontekście ewolucyjnym). Sama zdolność orzekania prawdziwości zdania $G_T$ nie wyklucza mechaniczności umysłu, orzekanie prawdziwości może być po prostu odpowiednikiem zapalania zielonej lampki. Nie jest wykluczone, że uznawanie przez nas pewnych zdań za prawdziwe odbywa się na mocy nieznanych nam reguł, w związku z czym sam fakt *operowania* terminami semantycznymi nie musi świadczyć o niemechaniczności umysłu; nie można wykluczyć sytuacji, że umiejętność rozumienia zdań zostanie „zautomatyzowana". Arbitralne stwierdzenie, że maszyna dowodzi *mechanicznie* a umysł dowodzi przez *zrozumienie*, prowadzi do błędnego koła, ponieważ z góry zakłada się niemechaniczność umysłu[25]. Odmawianie zdolności semantycznych maszynie należy traktować zatem jako argument preferencyjny, a nie formalny.

---

prawdziwe, bądź fałszywe. W naturalnym modelach, zdanie to będzie prawdziwe. Prawdziwość zdania $G_T$ nie ma charakteru „absolutnego", ponieważ istnieją modele, w których to zdanie jest fałszywe.

[25]Wydaje się, że podobny zarzut można postawić komuś, kto powołuje się na tzw. *argument chińskiego pokoju*.

Zgodnie z twierdzeniem Tarskiego o niedefiniowalności prawdy dla dostatecznie bogatych i niesprzecznych systemów formalnych, żadnej własności, dotyczącej **wszystkich** formuł pewnego języka I rzędu, nie da się reprezentować w tym systemie przy pomocy jednej (tej samej) formuły[26]. Na mocy twierdzenia o reprezentowalności wynika stąd dalej, że czynność orzekania takiej uniwersalnej własności, którą w rozważanym przypadku jest prawdziwość, nie może być „zautomatyzowana" w tym systemie. Zdolność orzekania prawdziwości o wszystkich zdaniach pewnego języka formalnego oznacza zdolność wyjścia poza system. Gdyby umysł potrafił zweryfikować dowolne zdanie odpowiednio bogatego języka elementarnego, świadczyłoby to o jego niemechaniczności. Jest jednak inaczej. Wierzymy wprawdzie, że każde zdanie pewnego języka formalnego jest albo prawdziwe, albo fałszywe w modelu odpowiednim dla tego języka, czyli zakładamy dwuwartościowość, co przejawia się w definicji spełniania dla formuły $\neg\alpha$ w modelu tego języka, niemniej istnieją sytuacje, w których orzekanie prawdziwości nie jest dla nas trywialne. Przykładem może być tzw. *hipoteza continuum*, która jest zdaniem niezależnym w teorii ZFC. Orzeczenie prawdziwości w tym przypadku nie jest tak łatwe, jak w przypadku zdania Gödla. Nasza zdolność intuicyjnego rozpoznawania prawdy wydaje się więc ograniczona. Dlatego też przekonanie o tym, że potrafimy orzec prawdziwość bądź fałszywość każdego zdania, które nie jest dowiedlne w systemie formalnym, należy traktować z dużą ostrożnością. Jeżeli nie potrafimy przypisać wartości logicznej każdemu zdaniu pewnego języka sformalizowanego (na przykład języka teorii mnogości), to nasza przewaga nad maszyną może być złudzeniem. Wiemy, że każda niesprzeczna i odpowiednio bogata teoria, wyrażona w języku elementarnym i oparta o rozstrzygalny zbiór aksjomatów, nie jest rozstrzygalna,

[26]Twierdzenie to zachodzi również dla systemów zupełnych, tzn. takich, w których dla każdej formuły $\varphi$, w systemie jest dowiedlna formuła $\varphi$ lub jej zaprzeczenie.

czyli w języku teorii istnieją zdania niezależne od tej teorii[27]. Wierzymy również, że istnieje niesprzeczna i zupełna teoria pewnego modelu, do której należą wszystkie i tylko prawdziwe zdania języka, którego interpretacją jest ten model, przy czym teorii tej nie da się zaksjomatyzować w sposób rozstrzygalny[28]. Przekonanie, że umysł jest w stanie dokonać wglądu w tę teorię ma jednak charakter metafizyczny a nie formalny. Przykładem zdania, co do którego nasz wgląd zawodzi, jest wspomniana *hipoteza continuum.* Paradoksalnie, w przypadku gdy umiemy podać niekwestionowalny dowód jakiegoś twierdzenia, to, jak uważają niektórzy, staje się on efektywny, a wtedy, jak zauważa Webb, „na podstawie faktu, że procedury efektywne mogą być symulowane przez maszyny Turinga, wnioskujemy, że dowody te także może symulować maszyna"[29]. Wniosek z powyższych rozważań jest następujący: o ile nawet istnieje „dwuwartościowy" platoński świat matematyki, to umysł niekoniecznie ma do niego dostęp, a jeżeli świat ten jest mitem i matematyka jest jedynie „modelem umysłu", interpretacją pewnego języka, to okazuje się, że ten model jest *wirtualny*, że jest jedynie ideą regulatywną, ponieważ to, co jest w stanie stworzyć umysł, nie może być kompletne. Nie można zatem wykluczyć, że nasza matematyka jest w zasięgu możliwości jakiegoś superautomatu. Ponadto, postulaty, którymi kierujemy się przy tworzeniu teorii matematycznych, na przykład niesprzeczność, mogą być uwarunkowane empirycznie, choć nie oznacza to wcale, że rzeczywistość „sama w sobie" jest niesprzeczna.

---

[27] Mówi o tym twierdzenie Churcha.

[28] Teoria taka istnieje na mocy twierdzenia Lindenbauma. W dowodzie wykorzystuje się nieefektywne metody teorii mnogości (należy przy tym pamiętać, że *nieefektywne* metody dowodu nie oznaczają tego samego co *nieefektywność* w teorii rekursji), choć dla języków skończonych ($|\ J\ | = \aleph_0$) twierdzenie można dowieść efektywnie.

[29] J. C. Webb, *Mechanism, Mentalism and Metamathematics*, Synthese Libr. vol. 137, Reidel, Dodrecht, 1980, cytat za: S. Krajewski, op. cit., s. 132.