Stanisław Krajewski

TWIERDZENIE GÖDLA I JEGO FILOZOFICZNE INTERPRETACJE

(Rozdział drugi)

Twierdzenie Gödla a (nie)mechaniczność umysłu

John R. Lucas w 1959 roku wygłosił odczyt[1], w którym argumentował, że twierdzenie Gödla dyskwalifikuje mechanicyzm, rozumiany jako teza, że umysł jest równoważny skończonej maszynie. Opublikowana wersja tego odczytu, „Minds, Machines, and Gödel" (Lucas [1961]), stała się w literaturze przedmiotu tekstem klasycznym. Jest on podstawowym odnośnikiem da wszystkich, którzy twierdzą, że z twierdzenia Gödla wynika to, że umysł nie ma natury mechanicznej. Jest tak niezależnie od tego, że sam tekst jest niezbyt dobry i pomimo tego, że jego autor nie należy bynajmniej do najwybitniejszych znawców twierdzenia Gödla. Jednak jako pierwszy opisał zagadnienie w miarę szczegółowo. Dlatego będziemy mówić o argumencie Lucasa.

A. Od Posta do Penrose'a

Oczywiście Lucas nie był pierwszy. Wiele osób, zapoznając się z wynikami Gödla, miało i ma nadal poczucie, że dzięki GI i GII dowiedzione jest ograniczenie dotyczące szeroko rozumianych maszyn, czyli komputerów, robotów i ich sieci (Lucas w [1961] używa terminu „maszyny cybernetyczne"), które nie odnosi się do ludzi, a więc w ten sposób udowodnione jest istnienie zasadniczej różnicy pomiędzy ludźmi a maszynami. Idea jest prosta: jeśli jakaś maszyna przedkłada prawdy matematyczne, to nie może przedłożyć zdania Gödla skonstruowanego dla ogółu tych prawd, o ile ma nie popaść w sprzeczność; my natomiast możemy to zdanie udowodnić. A więc – hura! – jesteśmy lepsi niż dowolna maszyna.

1. Prekursorzy

Choć prowadzono takie rozważania już wcześniej, pierwsza drukowana wzmianka znajduje się u Alana Turinga, w fundamentalnej pracy [1950]. Jej celem było pokazanie, że maszyny mogą myśleć, a raczej spełniać role, które kojarzymy z inteligencją. Autor przyznaje jednak ważność argumentacji matematycznej, tzn. opartej na twierdzeniu Gödla lub wprost na twierdzeniu Turinga, która, „sądzi się", dowodzi „takiej niemożności maszyn, jaka intelektu człowieka nie dotyka"[2]. Odczuwamy wyższość i „nie jest to uczucie iluzoryczne", pisze Turing, ale dodaje: „nie przypisywałbym mu większego znaczenia."[3] Co oznacza ta ostatnia

[1] Odczyt dla Oxford Philosophical Society był wygłoszony 30.10.1959; p. np. Lucas [1996], 107. Lucas był wykładowcą filozofii w Oxfordzie.

[2] Cyt. za polskim tłum. w: Chwedeńczuk [1995], 283.

[3] Wg Chwedeńczuk [1995], 284.

uwaga? Wydaje się, że Turing chciał powiedzieć, że warto jest budować roboty, nawet jeśli mają one podlegać jakimś ograniczeniom.[4] Widział te ograniczenia. Wcześniej, w odczycie wygłoszonym w 1947 r. powiedział, że „jeśli oczekujemy od maszyny bezbłędnego działania, nie możemy żądać, by jednocześnie była inteligentna".[5] Sam Turing był bardzo zaangażowany w pierwsze prace dotyczące wykorzystywania komputerów.[6] A o wspomnianych ograniczeniach nic więcej nie napisał. Turing nie był też pierwszą osobą, która się nad tym zastanawiała. Gdy pisał, że „sądzi się", iż twierdzenia o niezupełności pokazują takie ograniczenia maszyn, które nie stosują się do ludzi, odnosił się do dyskusji wcześniejszych. Ponieważ przebywał w Princeton, a Alonzo Church był jego promotorem, musiał znać dyskusje tam prowadzone z udziałem Gödla oraz von Neumanna, Churcha, Bernaysa, Kleene'go, Rossera i innych.[7] Sam Gödel dokonał znacznie subtelniejszej analizy sytuacji i podzielił się nią w 1951, gdy wygłosił wykład im. Gibbsa, ale wydaje się, że niewiele zrozumiano wtedy z jego uwag.[8] (Są one streszczone niżej w II.M.) Zaczęły one szerzej oddziaływać po ponad dwudziestu latach, po publikacji pierwszej książki Hao Wanga, zawierającej zapis rozmów z Gödlem na ten temat. W każdym razie jeszcze przed Turingiem, lub w tym samym czasie, podobne myśli, co on, zapisali Emil Post i Paul Rosenbloom. Potem, ale przed Lucasem, podobne wzmianki uczynili w druku James Newman i Ernest Nagel oraz John Kemeny. Oto ich krótki przegląd.

W 1941 roku jeden z pionierów współczesnej logiki matematycznej Emil Post napisał: „*Maszyna* nigdy nie stworzy zupełnej logiki; albowiem gdy maszyna jest już zbudowana, *my* możemy dowieść twierdzenia, którego ta dowieść nie potrafi."[9] Twierdził, że taką myśl zaczął rozważać już w roku 1924. Potem dopiero uwzględnił wyniki Gödla. Jego tekst ukazał się drukiem znacznie później, w antologii Davisa [1965]. Nie jest to dokładnie teza, że umysł nie jest maszyną, ale sugestię, że o to chodzi, można wyczytać w sformułowaniu: „my możemy dowieść".

W zakończeniu swojego wykładu logiki Rosenbloom napisał, że twierdzenie Gödla pokazuje, iż „pewne problemy nie mogą być rozwiązane przez maszyny, tzn. mózgi [brains] są niezbędne" (Rosenbloom [1950], 208). Człowiek „nie może wyeliminować konieczności użycia inteligencji" (*ibidem*, 163). Nie próbuje jednak dokładniejszej analizy, choć wspomina, że „inteligencja to zdolność do introspekcji" (*ibidem*, 208), a mózg wydaje się zawierać odpowiednie wyposażenie. W podobnym duchu, ale bardzo obszernie i znacznie głębiej, dywaguje Douglas Hofstadter w swoim bestsellerze [1979].

---

[4] Potwierdza to opinia Hodgesa, znawcy Turinga: „Nie zagłębiał się w konsekwencje twierdzenia Gödla i własnych wyników – po prostu przecinał węzeł gordyjski" (Hodges [2002], 301).

[5] W odczycie z 20.02.1947. Zob. Hodges [2002], 301.

[6] P. np. Hodges [2002], szczególnie rozdz. 6. Turing był zresztą prekursorem wielu późniejszych prac w dziedzinie komputerów i sztucznej inteligencji (p. Copeland i Proudfoot [1999]).

[7] Rozmawiał z Churchem, ale nie z Gödlem, Bernaysem, Kleene'm, Rosserem (Feferman [1988], 109).

[8] Tak można sądzić ze wspomnienia Wanga ([1996], 133), który podkreśla, że nie uchwycił wiele z bardzo szybko czytanego tekstu. Inni być może zrozumieli więcej i to rozpowszechniali, bo mam wrażenie, że niektóre późniejsze uwagi Putnama (który od roku 1953 do 1961 pracował na uniwersytecie w Princeton) czy Benacerrafa (też z Princeton) są rozwinięciem poglądów Gödla (w szczególności teza, iż może być tak, że jesteśmy maszyną, ale nie wiemy jaką; por. II.A.2 i II.M). Później sam Wang odegrał największą rolę w upowszechnieniu poglądu Gödla.

[9] Post [1941], w Davis [1965], 417. Wyróżnienia pochodzą od Posta.

W najpopularniejszym aż do książki Hofstadtera opracowaniu, opisującym osiągnięcia Gödla, czyli w przeznaczonej dla szerokiego kręgu książeczce Nagela i Newmana [1958][10], autorzy piszą, że „umysł ludzki zdaje się dysponować zasadniczo większymi możliwościami wykonywania operacji niż maszyny, które obecnie potrafimy obmyśleć. (...) struktura i działalność ludzkiego umysłu jest daleko bardziej złożona i subtelna niż budowa i sposób funkcjonowania którejkolwiek z maszyn, jakie dziś potrafimy zaprojektować" (Nagel, Newman [1966], 70-71). Dwukrotnie wymienione zastrzeżenie („potrafimy obmyśleć", „dziś potrafimy zaprojektować") świadczy o staranności sformułowań autorów. Mogłoby się wydawać, że ich podejście podważa tezę o niemechaniczności umysłu, bo dopuszcza pojawienie się maszyn w jakimś innym nieznanym dotychczas sensie, do których nie stosowałyby się wyniki typu Gödla, a które mogłyby być równoważne umysłowi. Jednak autorzy powstrzymują się przed takimi wnioskami. Potwierdza to ich odpowiedź na zawartą w recenzji z książki krytykę Putnama, wedle której chodzi „zwyczajnie i po prostu" o błędne zastosowanie twierdzenia Gödla.[11] Mówią mianowicie, że Putnam „dogmatycznie" zakłada, że wszelki wyobrażalny dowód niesprzeczności maszyny (hipotetycznie równoważnej umysłowi) może być też skonstruowany przez tę maszynę.[12] To znaczy, że według nich pewne umysłowe możliwości mogą być uznane za niemechaniczne. Już ta wczesna polemika unaocznia, że stosunek do argumentu nazwanego potem argumentem Lucasa może zależeć od przyjętego podstawowego założenia, a mianowicie, czy da się maszynowo naśladować rozumowania tworzone przez umysł, czy też nie.

W książce o filozofii nauki z roku 1959 Kemeny pisze, powołując się na Gödla, Turinga i Churcha, że mając dowolną maszynę „możemy zawsze sformułować problem tego rodzaju, że może go ona zrozumieć, ale nie może rozwiązać". Jest to bardzo luźne sformułowanie, nadzwyczaj antropomorficzne, ale sugeruje to samo, co Lucas. Autor jest jednak ostrożny. Pisze bowiem dalej, że maszyna musi być ograniczona, lecz „nie jest rzeczą jasną, czy człowiek, dysponując odpowiednią ilością czasu, nie mógłby rozwiązać wszystkich tych problemów". Pyta też, tak jakby rozmawiał o tym z Nagelem i Newmanem, czy rzeczywiście „niemożliwe są maszyny", które nie podlegałyby ograniczeniom opisanym w twierdzeniu Gödla, i dodaje, iż jeśli istnieją, „to muszą być w pewnym sensie żywe" (Kemeny [1967], 226).

Wiarę w takie maszyny, które są w stanie osiągnąć „intuicyjną" wiedzę o metateorii[13], wyraża artykuł Smarta [1960]. Nie jest to pogłębiony tekst, ale jest warty wzmianki jako pierwsza niezależna od Lucasa publikacja, przedstawiająca nieco szerzej argument wychodzący z twierdzenia Gödla. Czyni to jednak tylko po to, by go skrytykować w przekonaniu, że „ludzie są skomplikowanymi maszynami", a „przemyślne" maszyny zdolne do wglądu („insight") są możliwe.[14] Kierunek myślenia Smarta idzie więc w stronę przeciwną w stosunku do argumentacji Lucasa.

Książka Kemeny'ego jest popularyzacją filozofii naukowej opartej na logice i matematyce. W latach pięćdziesiątych pogląd o antymechanicystycznych konsekwencjach

---

[10] Jej pierwsza wersja ukazała się w 1956 roku jako artykuł w *Scientific American*.

[11] „misapplication of Gödel's theorem, pure and simple" (Putnam [1960a], 207). Rzecz w tym, że możemy nie być w stanie dowieść niesprzeczności odpowiedniej maszyny, podobnie jak ona sama (por. II.G.1).

[12] Nagel i Newman [1961], 211.

[13] Smart [1960] używa terminu „ultimate syntax language" (s. 109), korzystając z terminologii Carnapa, na oznaczenie metateorii, która zawiera teorie z dodanymi kolejno zdaniami Gödla.

[14] Smart [1960], 107-9.

twierdzeń limitacyjnych był już najwyraźniej rozpatrywany wśród filozofów analitycznych jako naturalny, choć zapewne mało kto gotów był przysiąc, że nie kryje się w tym jakiś błąd. To właśnie Lucas wystąpił z przekonaniem, że matematyczny wniosek o wyższości człowieka nad maszyną, a nawet nad materią, jest niewątpliwy.

Pogląd o obaleniu mechanicyzmu nie jest jednak w żadnym wypadku przyjmowany powszechnie. W wyniku refleksji nad problemem wątpliwości miał Post: „Konkluzja, że człowiek nie jest maszyną, jest nieuzasadniona [invalid]. Wszystko, co możemy stwierdzić, to to, że człowiek nie może stworzyć maszyny, która może myśleć jak on [do all the thinking he can].”[15] Potem szereg autorów pokazywał słabości rozumowania w stylu Lucasa. To samo jest celem niniejszego rozdziału. Właściwie wśród logików matematycznych dominuje pogląd o nietrafności tego rozumowania. Jest tak – pominąwszy nawet pogląd samego Gödla – począwszy od Hilarego Putnama, którego pierwsza wzmianka krytyczna (w [1960] i w [1960a], a więc jeszcze przed publikacją Lucasa) została uznana przez Boolosa[16] za „klasyczną”, poprzez opinię Quine’a,[17] analizy Benacerrafa w [1967] i Wanga w [1974], po recenzję Putnama [1995] z wersji argumentu rozwiniętej przez Penrose’a. (Prace te są uwzględnione w krytyce argumentu Lucasa, zawartej w dalszym ciągu tego rozdziału).

Roger Penrose jest wybitnym fizykiem matematycznym, który rozwinął własną wersję argumentu Lucasa w książkach *The emperor’s new mind* ([1989], [1995]) i *Shadows of the mind* ([1994], [2000]). Wedle Putnama, Penrose popełnia zwykłą „pomyłkę matematyczną.” Rzecz zasługuje jednak na uwagę (p. niżej II.L) z dwóch względów. Po pierwsze, wiele osób właśnie z jego popularnych książek poznaje jakąś wersję argumentu Lucasa, a po drugie, Penrose wydaje się być zbyt poważnym matematykiem, by nie zareagował na wytknięcie zwykłego błędu formalnego. Nadal uznaje, że nie popełnił błędu, ale że – jak to sformułował Putnam – „prowadzi spór filozoficzny ze społecznością logiczną” (Putnam [1995], 370).

Argument z twierdzenia Gödla wywiera nadal „mistyczny” urok. Temu urokowi ulega zresztą wielu filozofujących naukowców, a coraz częściej i innych autorów, którzy powołują się na Gödla, by wygłaszać ogólne tezy już nie tylko o umyśle, ale o granicach racjonalności, o niepoznawalności świata itp. (O tym jest rozdz. IV.)

2. Dwie metody krytyki argumentu Lucasa

Pomimo dość daleko posuniętej zgody (wśród logików), że należy odrzucić argument Lucasa, trzeba przyznać, że ma miejsce pewna bulwersująca okoliczność. Mianowicie nie ma jednej drogi wykazywania błędu w argumentach w stylu Lucasa (i Penrose’a), a są co najmniej dwa podstawowe podejścia. Te dwie metody ataku dobrze streścił John Burgess. Otóż dla jednych „błąd leży w pominięciu możliwości, że może *być* tak, iż procedura [mechaniczna generująca prawdziwe zdania matematyczne] generuje tylko takie stwierdzenia matematyczne, których prawdziwość możemy stwierdzić [we can see to be true], ale nie mamy dostatecznie jasnego wglądu w to, co ta procedura generuje, by być w stanie

---

[15] Post [1941] w: Davis [1965], 423; Post pisze też: „Aby to [że nie można wnioskować, iż człowiek nie jest maszyną] zilustrować, możemy zauważyć, że dałoby się skonstruować rodzaj człowieka-maszyny, który by udowodnił podobne twierdzenie dla swoich aktów mentalnych.”

[16] We wstępie do pracy Gödla [1951], zamieszczonym w [CW3], 294.

[17] Smart [1960], s. 109, pisze, że Quine powiedział mu, iż argument oparty na twierdzeniu Gödla można odrzucić „na poziomie czysto matematycznym”. Chodzi o to, że tworzenie formuły gödlowskiej z opisu aksjomatów czy odpowiedniej maszyny Turinga jest efektywne. Ta (nadzwyczaj istotna) okoliczność jest ujęta poniżej w II.J.

stwierdzić, że tak właśnie jest [see that this is the case].” Dla innych błąd leży w tym, że „nawet jeśli widzimy, iż ta procedura generuje tylko takie stwierdzenia matematyczne, których prawdziwość, jak uważamy, da się stwierdzić [we think we see are true], to może być racjonalne uznanie ludzkiej omylności i powstrzymanie się od konkluzji, że ta procedura generuje tylko takie stwierdzenia matematyczne, które są faktycznie prawdziwe.”[18] Mówiąc bardziej obrazowo, pierwsza metoda ataku to okazanie, że nie da się wykluczyć, iż jesteśmy maszynami niesprzecznymi, ale nie wiemy tego, a druga – że nie da się wykluczyć, iż jesteśmy maszynami sprzecznymi. Ta pierwsza metoda została wprowadzona przez Gödla (p. II.M poniżej), ta druga – choć też wzmiankowana przez Gödla – przez Putnama (p. II.F i G).

Taka niejednoznaczność powoduje, że żadna krytyka nie wydaje się konkluzywna. Problem w tym, że pierwsza metoda zakłada, że jesteśmy niesprzeczni, a druga dopuszcza, że wręcz przeciwnie. Ponieważ są one niezgodne ze sobą, zwolennik Lucasa może to wygrywać, by wskazać, iż sprawa nie jest zamknięta: przeciwnicy nie mogą się pogodzić . Jednak wzięte razem stanowią one mocny kontrargument przeciw Lucasowi: albo jesteśmy niesprzeczni, albo nie, ale w obu przypadkach Lucas nie ma racji.

Celem niniejszego rozdziału jest uwzględnienie obu podejść, a ponadto ostateczne odrzucenie argumentu Lucasa na jeszcze innej drodze: bez zakładania czegokolwiek na temat naszej (Lucasa) niesprzeczności pokażemy (w II.K), jak każda metoda w stylu Lucasa prowadzi do błędnego koła lub popadnięcia w sprzeczność.

Nadzwyczaj ważne jest bez wątpienia to, że wniosku w stylu Lucasa nie formułuje sam Gödel. Mówiąc najogólniej, wierzy on w niemechaniczność umysłu, ale twierdzi, iż wniosek tej treści wynika z jego twierdzeń dopiero w połączeniu z *dodatkowymi* założeniami. Jak wspomniałem, jego argumentacja zawarta w [1951], została opublikowana po latach, a w całości dopiero w 1995,[19] a więc nie miała początkowo wpływu na wielu komentatorów, w szczególności na Lucasa. Stopniowo musiała przenikać poprzez rozmowy, które prowadzili z Gödlem wpływowi logicy, w szczególności Georg Kreisel i Wang, ale na początku miała wpływ głównie na niektórych autorów amerykańskich. Należy jednak dodać, iż nawet po podaniu do publicznej wiadomości poglądów Gödla, co miało miejsce pierwszy raz w książce Wanga [1974], Lucas nie zmienił stanowiska.[20] Nie jest to może dziwne, bo filozofowie nie zmieniają poglądów chętniej niż ktokolwiek inny. Jednak nie budzi zaufania to, że brak u niego wzmianki o poglądach Gödla. Penrose cytuje trochę Gödla, ale uważa, że da się ująć rzecz inaczej. Nie chodzi o to, że którykolwiek z nich miał z miejsca ulec autorytetowi Gödla, ale właśnie o rzetelność dyskusji merytorycznej.

Osobiste poglądy Gödla są dla nas szczególnie istotne, nie tylko ze względu na zamierzenie tej książki. To on rozpoczął te wszystkie rozważania, to on głębiej niż inni badał ich konsekwencje, to on wreszcie poświęcił wiele pracy konsekwencjom swoich twierdzeń dla poglądów na naturę umysłu. Jego przenikliwość jest najwyższej próby. Dlatego, choć nie należy w żadnej sprawie ulegać magii nazwiska, poglądom Gödla poświęcić wypada osobny podrozdział (II.M poniżej). To zakończy rozważania zawarte w tym rozdziale, ale łatwiej jest je śledzić, gdy ma się w pamięci stanowisko Gödla, wyrażone w [1951], które wedle streszczenia przekazanego w roku 1972 Wangowi[21] jest następujące:

---

[18] We wstępie do III części książki Boolosa [1998], 351.

[19] [CW3] i niezależnie w Rodriguez-Consuegra [1995].

[20] Por. Lucas [1996], [1997], [1998].

[21] Wg Wanga: [1974], 324, oraz [1996], 184-5. (Mniej dosłowne tłumaczenie zawarte w książce Penrose’a [2000], 168, mówi o „skończonej” teorii liczb; chodzi tu o zdania finitystyczne [finitary, finit], w tym klasy Π_1.)

> Na podstawie tego, co zostało udowodnione do tej pory, pozostaje możliwe, iż może istnieć maszyna do dowodzenia twierdzeń (którą może nawet da się odkryć empirycznie), która faktycznie *jest* równoważna intuicji matematycznej, ale nie da się *dowieść* tego, że tak jest, ani tego, że dostarcza ona tylko *poprawnych* [prawdziwych, correct] twierdzeń finitystycznej teorii liczb.

(Cały bieżący rozdział może być uznany za przypis do tego stwierdzenia. W pewnym sensie można uważać, że takim przypisem jest cała niniejsza książka.)

B. Tło: mechanicyzm

Argument Lucasa skierowany jest przeciwko tezie mechanicyzmu. Historycznie rzecz biorąc, mechanicyzm powstał w okresie Oświecenia. Przedtem najdalej w tym kierunku posunął się Kartezjusz, który twierdził, że zwierzęta są maszynami, ale ludzie nie, bo „nie znajdą się ludzie (…), którzy by nie byli zdolni zestawić razem rozmaitych słów i ułożyć z nich sensownych wypowiedzi, które czyniłyby zrozumiałymi dla innych ich myśli; odwrotnie zaś, nie ma żadnego zwierzęcia (…), które by dokazało tego samego" (Descartes [1981], 66-67). Kartezjusz był zarazem przekonany, że nie da się stworzyć mechanizmu naśladującego czynności specyficznie ludzkie: „aczkolwiek maszyny wypełniają wiele czynności równie dobrze, a może nawet lepiej niż niejeden z nas, to jednak nie wykonałyby z pewnością pewnych odmiennych czynności, dzięki czemu dałoby się wykryć, że nie działały posługując się wiedzą" (Descartes [1981], 66). Głównym kryterium odróżniania była mowa i myślenie, ale zapewne nie należy tego oddzielać od jego idei nieskończoności, która transcenduje *cogito*, a więc łączy nas, ludzi, ze sferą, która jest poza wszelkimi mechanizmami. Jednak sto lat później uważający się za jego kontynuatora lekarz La Mettrie w pracy „Człowiek-maszyna" postawił argument Kartezjusza na głowie: twierdził, że i człowiek może być uważany za maszynę.[22] Chodziło mu zarówno o ciało jak i o umysł. Ciało ludzkie to „zegar ogromny" zbudowany „kunsztownie i umiejętnie." Nic dziwnego, że mowa jest o zegarze, bo to był najbardziej skomplikowany sztuczny mechanizm wtedy znany. Natomiast „myślenie jest tak dalece nieodłączne od materii zorganizowanej, że wydaje się ono jej właściwością w równym stopniu jak elektryczność, zdolność ruchu, nieprzenikliwość, rozciągłość itd." (La Mettrie [1953], 81). To, czy możliwe jest stworzenie maszyny, będącej jak człowiek, czy wręcz będącej człowiekiem, było wówczas tylko kwestią wiary. Jest zagadnieniem otwartym i dzisiaj.[23] Nieco później Condillac wyobrażał sobie posąg, który wyposaża się w kolejne ludzkie zmysły. Był „prekursorem cybernetyki (…), ponieważ badał logiczne właściwości 'układów' bez względu na to, czy są 'ożywione', czy nie" (Apter [1973], 14). Można by też rzec, że było to prekursorskie w stosunku do robotyki. Etapem w rozwoju mechanicyzmu w odniesieniu do umysłu była też psychologia behawioralna. Nie jest dziwne, że o ile sto lat temu porównywano mózg do centrali telefonicznej, to w czasach współczesnych pojawiło się porównanie do komputera.[24]

---

[22] Praca [1747] ukazała się z datą 1748 na karcie tytułowej. Por. La Mettrie [1953], XXX.

[23] Sprawa mechaniczności zwierząt wydaje się prostsza i obecnie, gdy wiemy, że znaczna większość genów jest wspólna ludziom i zwierzętom, trudno jest widzieć w pozaumysłowej działalności zwierząt mechanizmy inne niż u ludzi. Również intencjonalność nie wydaje się być absolutnym kryterium rozróżniającym. Nawet sama fizjologiczna zdolność do mówienia jest może wynikiem tylko drobnej mutacji genetycznej (w sierpniu 2002, gdy kończyłem ten tekst, podano, że gen FOXP2, którego nie mają szympansy jest, być może, odpowiedzialny za zdolność do mówienia). To wszystko nie przesądza kwestii rozumienia.

[24] Analizę tła historycznego mechanicyzmu w perspektywie problematyki, która na tu interesuje, podaje Webb w [1980]. P. też Heller i Życiński [1988].

1. Sztuczna inteligencja

Współczesną wersją mechanicyzmu, a w każdym razie mechanicyzmu w odniesieniu do umysłu, jest ideologia sztucznej inteligencji. Jej zamierzenie jest, by użyć sformułowania Minsky'ego, osiągnięcie tego, by komputery wykonywały czynności, które od nas wymagają inteligentnego działania. Napotykamy dwie interpretacje: może chodzić tylko o odtworzenie efektów naszych działań (słabsza teza) lub o odtworzenie struktury myślenia, sposobu, w jaki działa umysł (mocniejsza teza).[25] Mocniejsza teza sztucznej inteligencji – w skrócie AI, od „artificial intelligence" – stanowi, że nasze umysły działają jak komputery. Tu też da się odróżnić dwie wersje: struktura działania umysłu to jedno, a posiadanie umiejętności specyficznie poznawczych, zdolności do rozumienia czy do aktów intencjonalnych – to, być może, co innego. Być może nawet naśladowanie struktury działania umysłu nie musi powodować rozumienia, wytwarzać semantyki. Umysł w całym swym, znanym nam z introspekcji, bogactwie intencjonalności i semantyki, może być bowiem nierozerwalnie związany nie tylko z mózgiem jako układem neuronów, ale z naszym ciałem, z jego innymi aspektami fizycznymi, chemicznymi, biologicznymi, a także ponadjednostkowymi (społecznymi). Wtedy naśladowanie umysłu nie byłoby tylko kwestią wzmocnienia mocy komputerów. Do rozróżniania działania umysłu od symulowanego działania umysłu (czyli rozróżniania słabej i mocnej AI w sensie Searle'a) skłania się coraz więcej osób, już nie tylko spośród sceptyków, ale nawet spośród osób zaangażowanych w tworzenie sztucznej inteligencji.[26] Niezależnie od ważności tych kwestii, analiza tych różnic nie jest w tym miejscu konieczna, ponieważ argument z twierdzenia Gödla ma obracać w niwecz nawet najsłabszą tezę AI, dotyczącą możliwości symulowania wyników, a więc i wszystkie mocniejsze wersje. Z tego samego powodu nie musimy się kłopotać tym, że definicja umysłu nie wydaje się możliwa. Uwzględnimy bowiem niektóre, znane nam z doświadczenia, efekty jego działania, a nie potrzebujemy znać jego istoty. Żadne opisy i znane z tradycji rozważania nie dają definitywnej koncepcji umysłu, ani jasności, czym on jest, choć mimo to można o nim stwierdzić pewne fakty pozytywne, np., że umysł jest cechą ludzi. Stwierdzenie faktów negatywnych, tego czym umysł nie jest, wydaje się łatwiejsze. Wymaga bowiem przyjęcia tylko niektórych niewątpliwych jego możliwości; wśród nich jest zdolność do zrozumienia twierdzenia Gödla.

Z drugiej strony, ponieważ rozważamy ewentualne obalenie tezy, że umysł jest mechaniczny lub może być naśladowany przez maszynę, należałoby zdefiniować, czym jest maszyna. Na przykład nie zaakceptowalibyśmy maszyny, w której wnętrzu ukryty jest sterujący nią karzeł. Chodzi o komputery, ale może też takie ich wersje, o których na razie nie mamy pojęcia. Czym więc jest maszyna? Podanie definicji również nie jest proste, choć

---

[25] To rozróżnienie nie jest tożsame z rozróżnieniem opisywanym np. przez J. Searle'a, i wspomnianym poniżej w tym samym akapicie, wedle którego mocna AI to teza, że umysł *jest* programem, a słaba – że komputer tylko symuluje działanie umysłu (por. Searle [1991]). W dalszym ciągu będę używał wyrażenia „mocna teza AI" w sensie Searle'a, a na potrzeby niniejszego podrozdziału używam sformułowania „słabsza" i „mocniejsza teza AI".

[26] Obok takich krytyków jak Dreyfus (p. [1972]) czy Searle, można w pewnym sensie stawiać nawet „guru" AI, Minsky'ego, który uważa, że można zrozumieć umysł, ale „mimo opinii zaciekłego redukcjonisty, należy tak naprawdę do antyredukcjonistów" (Horgan [1999], 237). Warto też wspomnieć ewolucję Putnama, który w [1960] dopuszczał, że umysł może być programem, w [1987] i [1988] uznał, że stany umysłowe są nie tylko różne od fizycznych, ale i od obliczeniowych, że „'poziom intencjonalny' po prostu nie daje się zredukować ani do 'poziomu obliczeniowego', ani do 'poziomu fizycznego'" ([1998], 340), a więc, że umysły to nie programy.

zapewne łatwiejsze niż zdefiniowanie umysłu. Uprzedzając rozważania z II.D.1 można od razu stwierdzić, że jesteśmy w o tyle dobrej sytuacji, że możemy odwołać się do Tezy Churcha. Maszyny, które przetwarzają informacje (o takie nam chodzi, bo oczywiście nie o takie maszyny, jak silnik samochodowy czy obrabiarka), czymkolwiek są, mają dawać produkt, który da się opisać funkcją rekurencyjną. Wszystkie znane próby definiowania maszyn dają pojęcie równoważne funkcjom rekurencyjnym i maszynom Turinga. Oczywiście mowa o równoważności co do wyników, a nie co do sposobu funkcjonowania, ale szczęśliwie o to właśnie chodzi w obalanej tezie słabej AI.

2. Różne stopnie tezy o mechaniczności umysłu

Teza mechanicyzmu może być odniesiona do różnych zakresów. W najprostszym sformułowaniu pełna mocniejsza teza brzmi jak chciał La Mettrie:

(Mech) Człowiek jest maszyną.

Węższą wersją jest teza mówiąca tylko o umyśle:

($\mathrm{Mech_U}$) Umysł ludzki jest maszyną.

Jest to teza AI. Jeszcze węższa wersja powstaje przez ograniczenie do matematyki:

($\mathrm{Mech_M}$) Działanie umysłu w zakresie matematyki jest mechaniczne.

Najbardziej ograniczona wersja, tu rozważana, dotyczy samej arytmetyki liczb naturalnych:

($\mathrm{Mech_A}$) Działanie umysłu w zakresie arytmetyki jest mechaniczne.

Jak już wspomnieliśmy, każda z tych tez może być sformułowana w wersji słabszej, która nie mówi o działaniu człowieka i jego umysłu, jak tezy powyższe, ale tylko o *wynikach* tego działania. Słabsza wersja tezy mechanicyzmu nie mówi więc, że człowiek, umysł, choćby w zakresie matematyki, a nawet arytmetyki, jest jakąś maszyną, ale dopuszcza możliwość, że mamy do czynienia z czymś w istocie niemechanicznym; utrzymuje zarazem, że używając odpowiedniej maszyny da się symulować człowieka, czy jego umysł tak, by otrzymywać dokładnie te same wyniki. Otrzymujemy więc odpowiednie słabsze tezy:

(mech) Człowiek może być symulowany przez maszynę (robota).

($\mathrm{mech_U}$) Umysł może być symulowany przez maszynę.

($\mathrm{mech_M}$) Działanie umysłu w zakresie matematyki można symulować mechaniczne.

($\mathrm{mech_A}$) Działanie umysłu w zakresie arytmetyki można symulować mechaniczne.

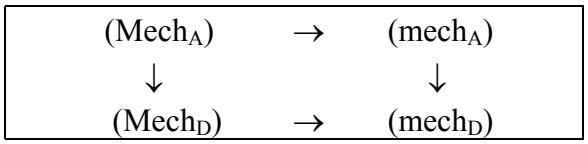
Można powiedzieć, że ($\mathrm{mech_U}$) to słabsza teza AI.

Do tych tez można by dodać jeszcze słabsze i bardziej ograniczone wersje wynikające z naszej wiedzy o postaci formuł gödlowskich. Tak więc ($\mathrm{Mech_D}$) i ($\mathrm{mech_D}$) odnosiłby się do działania umysłu w zakresie stwierdzania nieistnienia rozwiązań równań diofantycznych, a ($\mathrm{Mech_\Pi}$) i ($\mathrm{mech_\Pi}$) – do zakresu zdań postaci Π_1; dopiero od stosunkowo niedawna (dzięki twierdzeniu Matiasiewicza, p. I.B.10) wiemy, że są to zakresy równoważne.

Jest oczywiste, że zachodzą następujące implikacje (w tym schemacie wszystko, co poniżej, wynika z tego, co wyżej, a na danym poziomie to, co z prawej – z tego, co z lewej):

(Mech)	$\rightarrow$	(mech)
$\downarrow$		$\downarrow$
$\mathrm{(Mech_U)}$	$\rightarrow$	$\mathrm{(mech_U)}$

$\downarrow$		$\downarrow$
$\mathrm{(Mech_M)}$	$\rightarrow$	$\mathrm{(mech_M)}$
$\downarrow$		$\downarrow$



(mech$_U$) = Słabsza teza AI.

(Mech$_U$) = Mocniejsza teza AI,

Przedostatnia teza (mech$_A$) oznacza, że wszelkie twierdzenia arytmetyczne, dostępne umysłowi prawdy o liczbach, można też otrzymać w wyniku działania pewnej (jednej ustalonej) maszyny. Ostatnia teza (mech$_D$) jest najsłabsza i najwęższa. Jest najsłabsza z wymienionych powyżej tez, bo wynika z wszystkich pozostałych, a więc jej obalenie automatycznie obala wszystkie powyższe tezy. Otóż wedle argumentu Lucasa teza (mech$_A$) nie może być prawdziwa, bo zdanie Gödla dotyczy arytmetyki: żadna niesprzeczna maszyna nie może dochodzić do tych samych prawd arytmetycznych, bo zdanie Gödla odnoszące się do niej jest dla niej nierozstrzygalne, a my widzimy, że jest prawdziwe. W gruncie rzeczy, jak wiemy, mówiąc o zdaniu Gödla, możemy mieć na myśli tylko problemy diofantyczne. A zatem gdyby argument Lucasa osiągał cel, obalałby wszystkie wymienione wyżej tezy mechanicyzmu i sztucznej inteligencji.

C. Argument Lucasa

Wspomniane zostało powyżej, że argument Lucasa da się streścić w jednym zdaniu. Formułując je nieco inaczej niż poprzednio, można rzec tak: żadna maszyna nie może być równoważna umysłowi, bo umysł rozpoznaje prawdziwość zdania Gödla dla niej, a sama maszyna – na mocy twierdzenia Gödla – nie może, chyba że jest sprzeczna, ale wtedy na pewno nie jest równoważna umysłowi ludzkiemu. To wystarcza, by obalić tezę mechanicyzmu. Jest to wywód nad wyraz atrakcyjny. Wydaje się, że ściśle, niemal matematycznie, udowodniliśmy coś głębokiego na temat natury człowieka. Nic dziwnego, iż Lucas kończy swój tekst stwierdzeniem, że „jeśli ten dowód fałszywości mechanicyzmu jest słuszny, to ma to olbrzymie konsekwencje dla całej filozofii" (Lucas [1961], 126). Nie będzie już presji wynikającej z rozwoju wiedzy przyrodniczej, by „negować wolność w imię nauki." Bo „żadne badania naukowe nie mogą wyczerpać nieskończonej różnorodności umysłu ludzkiego" (*ibidem*, 127).

Z powyższych cytatów widać, że za wywodem o charakterze w zasadzie logicznym kryje się potrzeba filozoficzna – chęć uzasadnienia przyjętej z góry tezy o niewystarczalności materializmu czy scjentyzmu. W gruncie rzeczy widać też potrzebę religijną – chęć matematycznego potwierdzenia istnienia duszy i wolnej woli.[27] Takie lub podobne pragnienie jest obecne również u wielu innych osób, które używają argumentu w stylu Lucasa.[28] Nie u wszystkich jednak odwoływanie się do takiego argumentu oznacza postawę

[27] Książka Lucasa [1970], w której obszerniej powtarza tę samą argumentację, ma tytuł „Wolność woli".

[28] Trudno ocenić, jak wiele jest takich osób. Dennett pisze w [1972] o „znaczącej [considerable] bezkrytycznej akceptacji" argumentu Lucasa wśród matematyków i filozofów. Natomiast Putnam w [1995] pisze coś wręcz odwrotnego: Penrose jest „jedynym" współczesnym myślicielem, który uznaje argument Lucasa. Sądzę, że Dennett jest bliższy prawdy, choć oczywiście ocena zależy też od decyzji, kto jest „myślicielem".

anyscjentystyczną. Na przykład najgłośniejsze w ostatnich latach takie próby, opisane w książkach Penrose'a ([1989], [1994]) są powodowane jego doświadczeniami matematycznego wglądu i przekonaniem, że wyrażają one zjawiska takie jak świadomość, które są niealgorytmiczne. Autor spekuluje, że są one zakorzenione na poziomie kwantowym. Autorom takim jak Penrose chodziłoby więc nadal o opis naukowy, który nie wykracza poza zjawiska fizyczne, więc mieści się w ramach scjentyzmu; zakładają oni jednak, że musi on sięgać do innych poziomów budowy materii niż te, które uwzględniają normalne opisy maszyn.

Oczywiście krytykują Lucasa ci, którzy wierzą, że obecnie opracowywane maszyny i ich sieci mogą osiągnąć wszystkie poziomy dostępne nam ludziom.[29] Lucas – podobnie jak wielu innych autorów, spośród których najbardziej znany jest Hubert Dreyfus (począwszy od książki [1972] o znaczącym tytule „Czego komputery nie potrafią?") – sądzi wręcz przeciwnie. Jednak zastrzeżenia wobec argumentów w stylu Lucasa nie muszą wynikać z innej wizji metafizycznej: niektórzy krytycy Lucasa, owszem, wierzą w mocną tezę AI, ale inni podzielają tezę o jakościowej różnicy między człowiekiem a maszyną. Do tych ostatnich należy sam Gödel, który był zdecydowanym przeciwnikiem materializmu i scjentyzmu (por. III.C). On, jak wielu innych, chciałby okazać, że umysł przewyższa maszyny, ale stwierdza, że samo jego twierdzenie o niezupełności do tego nie wystarcza. Krytyka wywodów w stylu Lucasa nie musi więc wynikać z innego filozoficznego punktu wyjścia: może mieć charakter merytoryczny, logiczny. Dodam, że *musi* mieć taki charakter, żeby miała istotną wartość.

Zanim rozpoczniemy krytyczną analizę argumentu Lucasa, należy wymienić najważniejsze punkty tego wywodu. Wymaga to rekonstrukcji, która, oczywiście, powinna być możliwie życzliwa. Poniżej przedstawiona wersja, pomimo prostoty, oddaje wiernie omawiane rozumowanie. Można je rozbić na cztery kroki (L1) – (L4). Dzięki temu łatwiej będzie uporządkować dyskusję najróżniejszych zastrzeżeń omawianych w literaturze przedmiotu. Mówimy o maszynach, które mogłyby być równoważne ludzkiemu umysłowi, choćby w słabszym sensie, tzn. w zakresie osiąganych wyników, a niekoniecznie jeśli chodzi o sposoby dochodzenia do nich. Celem argumentu jest „wygödlowanie"[30] maszyn.

(L1) Zauważamy, że maszyny, zwane przez Lucasa „maszynami cybernetycznymi", są z konieczności równoważne systemom formalnym. Każda maszyna M ma określoną skończoną liczbę stanów i instrukcji, a więc odpowiada konkretnemu systemowi formalnemu S w sensie ustalonym przez logikę: S dany jest przez wypisane w określonym języku formalnym aksjomaty i reguły wnioskowania. Obliczenie, a ogólniej – sekwencja operacji dokonywanych przez maszynę M, odpowiada dowodowi formalnemu w systemie S.

(L2) Jeśli maszyna M ma być modelem umysłu, to musi zawierać „mechanizm, który może oznajmiać prawdy arytmetyczne" (Lucas [1961], 115). Wyrażenia, które maszyna M może „przedłożyć jako prawdziwe" [produce as true],[31] odpowiadają twierdzeniom systemu S.

---

[29] Wśród polskich autorów przekonanie, że sztuczny mózg jest w zasięgu ręki najmocniej wyraża Buller w [1997].

[30] Ten neologizm jest użyty tu (jak i w moim [1988]) jako odpowiednik angielskiego „out-Gödel". Szumakowicz w [1989] użył słowa „przegödelizować", ale „wygödlować" wydaje mi się zgrabniejsze i trafnie kojarzące się z „wykolegować", „wykuglować".

[31] Lucas pisze wprowadzając to pojęcie: „The conclusions it is possible for the machine to produce as being true" ([1961], 115). Wyrażenie „produce as (being) true" tłumaczę jako „przedłożyć jako prawdziwe" zamiast literalnego „wyprodukować jako prawdziwe" (które jednak czasem trzeba będzie przywołać) lub bardziej

(L3) Można zatem utworzyć metodą Gödla formułę G, która nie jest dowodliwa w systemie S, czyli nie jest jego twierdzeniem. Oczywiście zakładamy tu, że S, a co najmniej jego część arytmetyczna, S_{ar} jest niesprzeczny. (W przeciwnym razie G jest dowodliwa, bo wszystko jest dowodliwe dzięki tautologii $A \wedge \neg A \rightarrow B$.) Gdyby był sprzeczny, nie nadawałby się przecież na model umysłu. Maszyna M nie może przedłożyć formuły G jako prawdziwej (na mocy twierdzenia Gödla).

(L4) *My* natomiast widzimy prawdziwość formuły G. Możemy prześledzić konstrukcję Gödla i przekonać się o jej niedowodliwości w S i o jej prawdziwości. Prawdziwość jej jest w istocie wynikiem, a nawet wyrazem, jej niedowodliwości. Nasz umysł umie więc coś, czego nie umie M. Nie da się symulować maszynowo wszystkich czynności umysłu naraz. Umysł nie może być równoważny jakiejkolwiek maszynie. „Formuła gödlowska to pięta achillesowa maszyny" (Lucas [1961], 116).

Jest to dokładne i życzliwe ujęcie wywodu Lucasa z [1961]. Argument nie uległ od tamtego czasu zasadniczym zmianom. W pracach Lucasa [1968] i [1970] a także nowszej pracy [1996], w której autor próbuje odeprzeć krytykę sformułowaną w literaturze pod adresem tego wywodu i rozumowań pokrewnych, nie pojawiają się nowe elementy, inne niż to, co zawierają powyższe punkty. Chodzi mi tu o nieobecność istotnie nowych składników *wywodu* logicznego zrekonstruowanego powyżej, bo oczywiście różne zastrzeżenia wobec niego są w kolejnych pracach dyskutowane z zamiarem odrzucenia, a pewne składniki są mocniej podkreślane (np. „dialektyczny" charakter argumentu – p. II.H). W artykule [1997] i w odczycie [1998] Lucas potwierdza swoje pierwotne stanowisko.

To samo w zasadzie rozumowanie jest podawane też przez innych autorów, w szczególności przez Penrose'a [1989]. (Potem, w [1994] oraz w [1997], podał on też zmodyfikowaną jego wersję. uwzględniająca głosy krytyczne oraz stanowisko Gödla, i sformułował odpowiedzi na krytyki – por. II.L.) Jednak każdy punkt powyższego wywodu może być kwestionowany. Punkty (L1), (L2), (L3), (L4) rozważamy poniżej po kolei. Potem zanalizujemy główną linię obrony Lucasa, a mianowicie „dialektyczną" naturę jego argumentu. Okazuje się, że pomijany początkowo przez Lucasa problem niesprzeczności jest zasadniczy. Wreszcie sformułujemy twierdzenie, które pokazuje, że groźba niesprzeczności jest zgubna nie tylko dla oryginalnego argumentu Lucasa, ale i dla każdego sposobu postępowania w tym duchu, i to nawet wtedy, gdy nie ma w nim mowy o prawdziwości.

D. Wokół (L1): Czy maszyny muszą być równoważne maszynom Turinga?

Stosunkowo najmniej wątpliwości w argumencie Lucasa budzi punkt (L1). Jednak nawet on może być podważany. Maszyny o skończonej liczbie stanów i instrukcji, które działają sekwencyjnie, tzn. krok po kroku, są w zasadzie równoważne maszynom Turinga. Dokładniej mówiąc, maszyny Turinga są ich matematycznymi idealizacjami, bo nie zważamy na ograniczenia praktyczne: dopuszczamy dowolnie wiele stanów, dowolnie wiele instrukcji, ew. dowolnie duże dane wejściowe (nawet gdy liczba stanów, instrukcji, czy rozmiar danych przewyższa liczbę cząstek elementarnych we wszechświecie, ustalaną przez dominujące obecnie teorie fizyczne i kosmologiczne). Oprócz tego dokonujemy innej podstawowej idealizacji: zakładamy, że taśma (pamięć) maszyny jest (potencjalnie) nieskończona. Produkcja wyjściowa każdej takiej maszyny rzeczywiście może być opisana jako ogół tez

---

antropomorficznego „uznać za prawdziwe". Jest to nieobojętny sposób wyrażenia się, który poddał krytyce Slezak [1982] (por. niżej: II.E.1).

dowodliwych pewnego systemu formalnego. Wystarczy zauważyć, że ten produkt jest zbiorem rekurencyjnie przeliczalnym; jeśli ograniczymy się do tezy ($mech_A$), to byłby to rekurencyjnie przeliczalny zbiór zdań arytmetyki, a taki zbiór jest aksjomatyzowalny w zwykłym rachunku logicznym.[32] Jeżeli więc argument Lucasa (tzn. pozostałe jego punkty) jest poprawny, to dowodzi, że umysł nie jest równoważny takiej maszynie idealnej, ale ją w pewnym punkcie przewyższa, a więc tym bardziej góruje nad każdą maszyną realną. Jednak czy nie ma maszyn o innej naturze, których nie da się przedstawić jako maszyn Turinga?

Jest to osobny i ciekawy problem. Dotyczy on z jednej strony tego, jakie maszyny są możliwe, a z drugiej tego, co (kogo?!) mamy prawo nazywać maszyną. Przed chwilą stwierdziłem, że wystarczy „zauważyć", że produkt maszyny jest zbiorem r.e. Jak jednak można to zauważyć? Czy przypadkiem nie chodzi tu o to, że należy to *założyć*? Te pytania wprowadzają nas w centrum problematyki dotyczącej Tezy Churcha.

1. Teza Churcha

Sprawa nie jest tak jednoznaczna, jak mogłoby się wydawać z perspektywy dzisiejszych matematyków i informatyków. Otóż Turing analizował obliczanie przez człowieka, a nie przez maszynę.[33] W centrum uwagi była ludzka działalność „mechaniczna", tzn. wedle ustalonych reguł, nietwórcza. Było tak począwszy od Fregego, który opracował pojęcie systemu formalnego, czyli takiego, w którym „wnioskowanie jest jak obliczanie", aż do Churcha, który w [1936] podaje, że każdy system logiki musi mieć reguły „efektywnie obliczalne", a zbiór reguł i aksjomatów musi być „efektywnie wypisywalny [enumerable]." Te wymagania Church *interpretuje* (po odpowiedniej arytmetyzacji) jako rekurencyjność (reguł) i rekurencyjną przeliczalność (zbioru reguł i aksjomatów). Nie jest to uzasadniane, ale – jak się wydaje – nie ma innego wyjścia, jeśli chcemy mieć ścisłe pojęcie systemu formalnego. To właśnie tę interpretację, Sieg formułuje jako „Church's Central Thesis": „Kroki dowolnej procedury efektywnej (rządzącej wywodami w logice symbolicznej) muszą być rekurencyjne." (Sieg [1994], 87). Wtedy można wykazać, że funkcje definiowalne w logice (powiedzielibyśmy teraz raczej – reprezentowalne") są rekurencyjne.[34] Jednak powyższe rozumowanie nie gwarantuje, że nie może pojawić się jakiś nowy rodzaj reguły dowodzenia, który nie byłby rekurencyjny. Dopiero Turing pokazał, dlaczego jest rzeczą racjonalną, by tego się nie obawiać. Turing rozważył problem, jakie zachowania są możliwe przy obliczaniu, bezmyślnym stosowaniu reguły wnioskowania, postępowaniu mechanicznym. Mechanicznym – czyli, rzec można za Turingiem, w zasadzie mogłaby je wykonywać maszyna. W gruncie rzeczy[35] Turing formułuje ograniczenia co do liczby obiektów, które może brać pod uwagę rachmistrz, czyli wedle określenia Gandy'ego[36] „komputor" (w przeciwieństwie do „komputera"). Oczywiście chodzi o abstrakcyjnego, czyli wyidealizowanego rachmistrza, który się nie myli i ma potencjalnie nieograniczenie wiele czasu i papieru (pamięci). Otrzymujemy w wyniku wniosek, który Gandy podsumowuje następująco: „Obliczanie odbywa się przy pomocy osobnych (dyskretnych) kroków i wytwarza zapis złożony ze skończonej (ale nieograniczonej) liczby komórek, z których każda

[32] Mówi o tym twierdzenie Craiga z pracy [1953].

[33] Podkreśla to Gandy [1988] i Sieg [1994]. Wittgenstein zauważył, że maszyna Turinga to „człowiek, który oblicza".

[34] Tę argumentację Churcha z [1936] Gandy w [1988] nazywa „step-by-step argument."

[35] Jak podkreśla uczeń Turinga Gandy w [1988], a za nim Sieg w [1994].

[36] Gandy [1988], 75, przyp. 24.

jest pusta lub zawiera symbol ze skończonego alfabetu. Na każdym kroku działanie jest lokalne i lokalnie wyznaczone zgodnie ze skończonym spisem instrukcji." (Gandy [1988], 75). Uzasadnieniem jest to, że „ludzka pamięć jest z konieczności ograniczona."[37] Z tego opisu da się dowieść, że funkcje obliczalne przez (wyidealizowanego) rachmistrza są rekurencyjne. Turing i prawie wszyscy późniejsi badacze wierzą, że i na odwrót. (Należy dodać, iż niektórzy wymagali, by dało się z góry ograniczyć liczbę kroków obliczenia, a to daje węższą klasę funkcji[38]). Ta równoważność – to teza Churcha w wersji Turinga.

Jest rzeczą uderzającą, że nie ma tu mowy o maszynach. Mówi się o maszynach Turinga, ale jest to punkt dojścia. Punktem wyjścia jest rachmistrz. Tym nie mniej od razu wydaje się oczywiste, że każda maszyna musi spełniać te same warunki. Tak sformułował to sam Church w recenzji z podstawowej pracy Turinga, [1937], pisząc, iż Turing zanalizował obliczalność maszynową. „Zdefiniować efektywność jako obliczalność przez dowolną maszynę, spełniającą warunki skończoności wydaje się adekwatną reprezentacją zwykłego pojęcia [efektywności], a gdy się tak uczyni, znika potrzeba hipotezy roboczej" (Church [1937]). Teza Churcha jest wtedy nie tyle hipotezą, co tezą, która staje się „natychmiast oczywista". Tak jest to zwykle ujmowane i dziś. Jednak dopiero analiza Gandy'ego w [1980] dała bardziej konkretne powody, by tak czynić, niż sama „intuicyjna oczywistość". Gandy kopiuje rozważania Turinga, ale stosowane wprost do „dyskretnych, deterministycznych urządzeń mechanicznych", które mogą też mieć paralelnie działające procesory. Ograniczenia na liczbę używanych elementów wywodzi się nie tyle z rozważań psychologicznych, jak u Turinga, ale fizycznych. Okazuje się, że faktycznie dowolne maszyny podległe powyższym ograniczeniom, wykonywujące dowolne operacje i ich iteracje, dają funkcje obliczalne przez maszyny Turinga, czyli rekurencyjne. Czyli „natychmiastowa oczywistość" nie zawiodła!

Mimo to nadal możemy sobie wyobrazić jakieś ogólniejsze pojęcie maszyny, które, być może, prowadzi poza maszyny Turinga. Jedno możliwe podejście polega na tym, że wyobrażamy sobie, iż maszyna może losować, co czynić w następnym kroku. Co wtedy? Otóż używanie elementu losowego, na przykład generatora liczb losowych, da się ująć w naszym modelu przez dopuszczenie niedeterministycznych maszyn Turinga. Zrobił to sam Turing. I dał odpowiedź, obecnie powszechnie znaną, że te nowe maszyny nie mogą uczynić więcej niż zwykłe (deterministyczne) maszyny Turinga: można bowiem opisać strukturę wszystkich możliwości i je kolejno przeglądać. Niedeterministyczne maszyny osiągają cel prędzej, ale w niniejszym kontekście prędkość działania nas nie interesuje.[39]

2. Turing a inne maszyny

Według Hodgesa Turing na początku swej kariery nie wykluczał związku „między umysłową 'intuicją', a nieobliczalnością, zmienił zaś swe poglądy dopiero około 1941

---

[37] Turing [1937], za: Davis [1965], 117.

[38] Tak uważał najprawdopodobniej Herbrand, ale nie Gödel (por. Sieg [1994], 82 i 103); Gandy w [1988], 60 i 77, mówi o takim wymaganiu jako o „złej monecie". W późniejszych czasach rozwinęła się cała teoria funkcji dowodliwie rekurencyjnych w różnych określonych teoriach formalnych. Są to podklasy właściwe klasy wszystkich funkcji rekurencyjnych.

[39] Lucas słusznie odpiera zarzut oparty na dopuszczeniu elementu losowego, ale czyni to w sposób niezbyt jasny: „zamiast rozważać, co całkowicie zdeterminowana maszyna *musi* czynić, rozważymy, co ona mogłaby być w stanie uczynić, gdyby miała urządzenie losujące, które by działało zawsze, gdy byłyby możliwe dwie operacje lub więcej, a żadna z nich nie prowadziła do sprzeczności" (Lucas [1961], 114).

roku."[40] Choć Turing zrazu nie wierzył w moc maszyn, potem stał się jednym z największych wizjonerów idei sztucznej inteligencji, a także kognitywistyki.[41] Burzliwa choć krótka historia AI obejmuje wczesny okres heroiczny, gdy pionierzy (jak Marvin Minsky czy Hans Moravec) uważali, że niedługo powstaną prawdziwie myślące maszyny[42], potem okres frustracji na przełomie lat siedemdziesiątych i osiemdziesiątych, wreszcie rozwijającą się od tego czasu koncepcję stopniowego postępu poprzez stosowanie przetwarzania równoległego, oraz algorytmów uczących się, w szczególności sieci neuropodobnych. Jednak refleksja teoretyczna nie musiała ulec istotnej zmianie: sieci neuropodobne mogą być symulowane przez zwykłe komputery. Zresztą w gruncie rzeczy najczęściej są![43] Choć w praktyce dają wielki postęp w implementacji pewnych funkcji, nie wydaje się, by prowadziły poza zakres funkcji rekurencyjnych. Zarówno maszyny niedeterministyczne jak i przetwarzanie równoległe nie przynoszą zasadniczego postępu, nie rozszerzają klasy funkcji obliczalnych. Chodzi – podkreślmy znowu – o wyidealizowaną obliczalność, bo gdy zaczniemy rozpatrywać praktycznie osiągalną obliczalność, poszerzanie klasy maszyn daje istotny postęp. W argumencie Lucasa chodzi o obliczalność w zasadzie, a nie w praktyce.

Mimo wszystko możemy uważać, że jakiś nowy typ maszyn nie da się sprowadzić do maszyn Turinga. Chodzi o obiekty, które dla nas będą niewątpliwie maszynami, a będą miały zasadniczo większą moc. Może chodzić o jakieś maszyny indukcyjne w odróżnieniu od dedukcyjnych.[44] Jednak póki co, pomimo zachodzącego obecnie rozwoju algorytmów uczących się, są to spekulacje. Musimy widzieć takie propozycje jako jakieś zastosowania wizji homunkulusa – ukrytego w maszynie człowieczka, który ma umysł taki, jak my.

Osobnym problemem jest to, *jak* powstaje umysł. Znamy tylko naturalnie powstałe umysły, ale czy to znaczy, że po przekroczeniu jakiegoś progu komplikacji maszyna nie może uzyskać umysłu? Nawet Lucas tego nie wyklucza. Wtedy jednak – twierdzi – ta maszyna „przestałaby być maszyną" (Lucas [1961], 126). Przy takim ujęciu część kontrowersji dotyczącej mechanicyzmu byłaby sporem o słowa. Żeby zostać przy realnym problemie, uznajmy, że bycie maszyną polega na pewnego typu funkcjonowaniu, a mianowicie, że jest ono powiązane z procedurami działania opisanymi przez Turinga. Musimy jednak uważać, by nie popaść w błędne koło. Jeśli założymy, że nasz umysł, który jest samoświadomy, nie może działać wedle tych procedur, to po prostu zakładamy to, czego mieliśmy dowieść w argumencie Lucasa i cała zabawa z twierdzeniem Gödla jest niepotrzebna.[45] Nie zakładajmy

---

[40] Hodges [2002], 8. Ta opinia jest zawarta w przedmowie do polskiego wydania, bo w samej książce (z 1983 roku) przedstawiony jest nieco inny obraz. Por. też Hodges [1998], 47.

[41] P. np. Hodges [2002], 315, i wspomniany już artykuł Copeland i Proudfoot [1999].

[42] Np. Minsky twierdził w roku 1970, że „w ciągu 3-8 lat zbudujemy maszynę dorównującą człowiekowi pod względem ogólnej inteligencji" (za Coveney i Highfield [1997], 170). Jeśli chodzi o argument antymechanicystyczny, to w 1961 napisał, że Rosenbloom w [1950] mówiąc wyższości umysłu (czy też mózgu) opiera się „na błędnej interpretacji sensu 'twierdzenia o nierozstrzygalności' Gödela" (p. Feigenbaum i Feldman [1972], 420).

[43] W praktyce sieci neuronowe często symuluje się na zwykłych komputerach, a nie na fizycznych układach neuropodobnych. Zwraca na to uwagę Searle w artykule [1990], który zawiera ulepszoną wersję argumentu Chińskiego Pokoju i polemikę z Churchlandami, którzy wyrażają nadzieję – choć przyznają, iż bez gwarancji – że dostatecznie rozbudowana „chińska sala gimnastyczna" będzie rozumiała chiński.

[44] George [1962]; może tu chodzić uczenie się lub „samo-programowanie" maszyn, czyli obiektów, które możemy „efektywnie skonstruować". Chari w [1963] przypomina o braku regularności w procesie twórczym, gdy z góry nawet nie wiadomo, jaki jest zestaw możliwych hipotez.

[45] Webb w [1968], 158, zarzuca Lucasowi, że zbliża się do takiego błędnego koła.

więc nic szczególnego na temat natury działania umysłu – poza tym, co na pewno wiemy z introspekcji.

Wszystkie te rozważania nie prowadzą więc do unieważnienia kroku (L1). Rozumowanie Lucasa stosuje się przynajmniej do tej obszernej klasy maszyn, które (jako struktury wyidealizowane) są równoważne maszynom Turinga, czyli – mówiąc matematycznie – urządzeniom, których globalny produkt, czyli ogół wyrażeń otrzymywanych na wyjściu, jest rekurencyjnie przeliczalny. Rekurencyjnie przeliczalny ma być zbiór wyrażeń, które pojawiają się na wyjściu, gdy pomijamy ograniczenia czasowe i gdy zakładamy, że nie ma nic na wejściu, albo, że na wejściu jest coś ustalonego, albo nawet, że na wejściu jest też coś rekurencyjnie przeliczalnego. Nie możemy dopuścić *dowolnych* manipulacji na wejściu, bo wtedy już tam dałoby się zawrzeć coś, co nie jest rekurencyjnie przeliczalne, a więc oczywiście i na wyjściu dałoby się wykroczyć poza rekurencyjną przeliczalność.[46] Z punktu widzenia matematyki ustalone wejście można by zapisać jako część (programu) maszyny, więc równie ogólne jest rozważanie maszyn bez żadnego wejścia. Nie możemy jednak zupełnie pominąć pojawiania się czegoś na wejściu, bo może to być potrzebne przy rozważaniu „dialektycznego" charakteru argumentu Lucasa (p. niżej II.H).

E. Wokół (L2): Czym jest prawdziwość dla maszyny?

„Każdy mechaniczny model umysłu musi zawierać mechanizm, który może oznajmiać prawdy arytmetyczne, bo to jest coś, co umysły umieją czynić", pisze Lucas.[47] Trudno temu odmówić racji. Maszyna musi pewne wyrażenia, które pojawiają się na jej wyjściu określać jako „prawdziwe". Mówimy, że „przedkłada je jako prawdziwe", choć w oryginale jest „produkuje [produces] jako prawdziwe." Wydaje się to na pierwszy rzut oka niezbyt zgrabne, ale niewinne. Jednak Benacerraf w [1967] i – nieco obszerniej – Slezak w [1982] (a również nawiasowo Wang[48]) twierdzą, że mamy tu do czynienia z ekwiwokacją.

1. Ekwiwokacja

Rzecz w tym, że używa się *jednocześnie* wyrażenia odpowiedniego dla maszyny („produkuje") i wyrażenia odpowiedniego dla ludzi („prawdziwe"). Użyte pojęcie musi ujmować to, co umysł może uczynić, a maszyna nie jest w stanie, musi więc jednocześnie pasować do trybu maszynowego – zimne „produkuje", „generuje", „drukuje", czy choćby rzeczowe „przedkłada na wyjściu" – oraz, z drugie strony, odpowiadać ludzkiemu sposobowi ujmowania – wyrażające rozumienie i akceptację „uznaje za prawdziwe". Okazuje się, że ekwiwokacja nie jest skutkiem niestaranności, ale leży u podstaw argumentu, który ma traktować o maszynach i ludziach jednocześnie, a zarazem nie dopuszczać do ich utożsamienia.

---

[46] To, że przy omawianiu argumentu Lucasa chodzi właśnie o to, zapisał w szczególności D. Lewis: „(…)to be a machine is (…) to be something whose output, for any fixed input, is recursively enumerable" ([1979], 375).

[47] „(…) enunciate arithmetical truths (…)" (Lucas [1961], 115). Penrose pisze też: „ascertain truths".

[48] O nieprecyzyjności tego wyrażenia pisał też przedtem Chihara w [1972], a Wang w [1974], pisząc o najwcześniejszej uwadze Posta („my możemy dowieść twierdzenia, którego [maszyna] dowieść nie potrafi" – p. wyżej II.A.1), zarzuca jej i „późniejszym argumentom" właśnie „ekwiwokację pomiędzy prawdziwością a dowodliwością" (Wang [1974], 327, przypis 6).

Jeśli bowiem mówimy o maszynach jako o odpowiednikach systemów formalnych, to wystarczy mówić o (formalnej) wywodliwości. Pojęcie prawdziwości nie jest niezbędne do wypowiedzenia twierdzenia Gödla; wystarczy powiedzieć, że niesprzeczny system jest niezupełny (syntaktycznie), czyli jest taka formuła A, że ani formuła A, ani formuła $\neg$A nie jest wywodliwa. Wiemy, że taką formułą jest właśnie formuła Gödla G. Warto podkreślić ten moment: wynik Gödla ma sens w płaszczyźnie czysto syntaktycznej. Aby operować tym twierdzeniem w odniesieniu do jakiejś teorii T, nie trzeba mieć nawet pojęcia prawdziwości dla formuł tej teorii! Możemy mieć na przykład teorie mnogości, dla których pojęcie prawdziwości jest niejasne. Nadal jednak wiemy, że niesprzeczność zapewnia niezupełność. Oczywiście w przypadku arytmetyki, tzn. gdy teoria T jest jakąś teorią liczb naturalnych, pojęcie prawdziwości nie wydaje się wątpliwe. Podobnie jest, gdy teoria T ma bogatszy język, ale interesują nas tylko zawarte w niej (może nawet nie bezpośrednio) zdania arytmetyczne; w takiej sytuacji nawet dziwna teoria mnogości nie będzie się wyróżniać, bo i tak będziemy zważać jedynie na zdania o liczbach naturalnych.

Niezupełność syntaktyczna faktycznie oznacza, że niesprzeczny system nie jest kompletny, tzn. nie zawiera wszystkich arytmetycznych zdań prawdziwych. Chodzi o prawdziwość w sensie potocznym, pierwotnym, metasystemowym. Można ją precyzować, można dowodzić prawdziwości wyrażeń przy ustalonej interpretacji. Jednak pojęcie prawdziwości leży, *oczywiście*, poza rozważanym systemem formalnym. (To tylko przed twierdzeniem Gödla można było mieć nadzieję, iż prawdziwość da się zredukować do dowodliwości w ustalonym systemie.) Używając określenia „produkuje jako prawdziwe", Lucas może mówić jednocześnie o ograniczeniach maszyn i o braku ograniczeń dla możliwości umysłu. „Lucas może ustanowić swoją tezę tylko stosując ekwiwokację pomiędzy dowodliwością a prawdziwością" (Slezak [1982], 452).

Potwierdzeniem tego zarzutu jest fakt, że omawiana dwuznaczność zawarta jest w sformułowaniu, którego użył Lucas niedawno, po co najmniej 35 latach dyskusji, w artykule [1996]: „...umysł może je wygödlować, produkując zdania gödlowskie dla nowych wersji tej maszyny i widząc, że są one prawdziwe, czego maszyna uczynić nie może."[49]

Należy w tym miejscu zapytać, czy nie da się uniknąć sposobu mówienia, który naraża na zarzut ekwiwokacji. Lucas zupełnie się nim nie przejął. Ale nawet przyznając, że ma miejsce niepokojąca ekwiwokacja, nie możemy na tym poprzestać. Samo przyznanie trudno uznać za obalenie argumentu Lucasa. Zresztą może da się go przeformułować? Benacerraf jako pierwszy napisał, że nie można bezrefleksyjnie raz odwoływać się do reguł (dowodzenia) maszyny a w innym miejscu – do reguł Lucasa, bo jeśli możliwe są dowody nieformalne dla człowieka, to może i dla maszyny? Zarazem sam Benacerraf twierdził, iż można sformułować argument bez popadania w ekwiwokację. Chodziło mu jednak o to, że byłoby to przypisanie Lucasowi dowodliwości w którymś z poprawnych (czyli mających tylko prawdziwe twierdzenia) systemów formalnych, ale „suma wszystkich systemów formalnych, które on produkuje, nie jest systemem formalnym."[50] Czyli u podstaw argumentu Lucasa byłoby założenie, że Lucas nie jest mechaniczny, a to nic nie wnosi. Wracając do kwestii ekwiwokacji jako używania naraz prawdziwości i dowodliwości, możemy dostrzec dwie możliwe drogi przezwyciężenia zarzutu. Może pojęcie prawdziwości da się z sensem odnieść do maszyn? A może w kontekście argumentu Lucasa da się w ogóle zrezygnować z mówienia o prawdziwości? Rozważymy po kolei obie te możliwości.

---

[49] Lucas [1996], 110, oraz [1997], 4.

[50] Benacerraf [1967], 21. Por. zaczęte przez Gödla rozważania o hierarchii teorii w IV.A.4.b.

2. Prawdziwość a maszyny

Czy pojęcie prawdziwości może odnosić się do maszyn? Z punktu widzenia krytyka mocnej AI maszyna nie może odnosić się do prawdziwości, bo prawdziwość to pojęcie semantyczne. Choć definicja prawdziwości jest trudna do zadowalającego sformułowania, wiemy, co oznacza to pojęcie i wiemy – pomimo całej tradycji krytycznej wobec klasycznego pojęcia prawdy – że odnosi się jakoś do porównywania stwierdzenia z faktami. Na przykład okazanie prawdziwości zdania Gödla polega właśnie na tym: interpretowane jako zdanie o (numerach) formuł i innych obiektów języka, mówi ono o pewnej liczbie, wspomnianej w sposób nieco pośredni, że jest ona numerem formuły niedowodliwej (w rozważanym systemie). Ta formuła – to akurat ona sama, więc jest ona prawdziwa, bo rzeczy się mają tak, jak ona stwierdza. Jest widoczne, że stosujemy normalne pojęcie prawdziwości i w zasadzie mówimy dwukrotnie to samo innymi słowami: raz, że zdanie Gödla nie jest dowodliwe, a potem, że jest prawdziwe, czyli jest, jak mówi, czyli ... nie jest dowodliwe. Mamy tu niemiłe poczucie ciągłego mieszania poziomów, ale jakoś możemy z tego wybrnąć. Może maszyna też by mogła? Przecież pozostajemy w zakresie zdań arytmetycznych i metaarytmetycznych, można to wszystko ściśle opisać, więc również – zmechanizować. (Dokładniejsza analiza jest poniżej w II.G.3.)

Jednak filozof może uznać, że maszyna może manipulować obiektami językowymi, ale tylko na poziomie syntaktycznym. Prawdziwość jako taka jest poza ich zasięgiem, bo jest cechą ludzką.[51] Możemy odmówić zdolności do rozumienia prawdziwości maszynom, nie tylko tym aktualnym, ale i wszelkim możliwym. Jednak w kontekście analizy argumentu Lucasa musimy uważać, by nie strywializować rozważań. Jeśli założymy, że „prawdziwa" prawdziwość nie jest dostępna maszynom, jest natomiast dostępna ludziom, to argument Lucasa nie jest potrzebny, bo po prostu *zakładamy* naszą wyższość nad maszynami, czyli to, czego mieliśmy dowieść.

Nie można wykluczyć tego, że maszyna może operować pojęciem prawdziwości i innymi pojęciami semantycznymi. Choć wysiłki w dziedzinie sztucznej inteligencji, aby to uczynić i osiągnąć „semantyczne przetwarzanie informacji" napotkały na głębokie trudności, to kto wie, czy postęp techniki nie będzie stopniowo prowadził do coraz większego wyrafinowania komputerów w dziedzinie, która dla nas jest sferą znaczenia: rozumienia sytuacji, sensu zdań, porównywania zdania z sytuacją. Nie brak ludzi, którzy w to wierzą. Gandy uznaje za możliwe, że za kilkadziesiąt lat komputery którejś kolejnej generacji, badając problemy matematyczne, będą wykazywały inteligentne zachowanie, czyli *będą* kolegami matematyków. Nawet jeśli na najwyższym poziomie będzie nimi zarządzał „master program", to będziemy o nich mówić używając normalnych terminów i pojęć stosowanych do mówienia o inteligencji i racjonalności. Gdzieś pomiędzy chipami a najwyższym programem będzie miejsce na heurystykę, a nawet „mechaniczne olśnienia." „Na wyrafinowanym poziomie nie jest ani rzeczą praktyczną, ani pożyteczną, ani rozsądną omawiać zachowania inteligentne używając wyłącznie pojęcia algorytmów i programów maszynowych" (Gandy [1996], 136). Jest to optymistyczny, choć umiarkowany głos w dyskusji nad AI. Oczywiście nie możemy być pewni, że trafnie przewiduje on przyszłość, a co ważniejsze, że prawdziwość może naprawdę być cechą maszyn.

Nie byłoby jednak dobrze, gdyby obalenie argumentu Lucasa polegało po prostu na *założeniu*, że maszyny *są* w stanie rozumieć, bo jeśli je dostatecznie rozwiniemy,

[51] Warto zauważyć, że Searle twierdzi, że program komputerowy nie może rozumieć, ale nie wyklucza, że „możliwe byłoby zbudowanie myślącej maszyny z innego materiału" (Searle [1991], 10).

automatycznie pojawi się cała sfera semantyczna. Innymi słowy, że „chińska sala gimnastyczna" dzięki swym rozmiarom przezwycięży ograniczenia „chińskiego pokoju". Takie założenie byłoby bowiem bliskie przyjęciu, że umysł jest rodzajem maszyny, czyli po prostu zignorowania argumentu Lucasa. Wtedy prawdziwość byłaby dowodliwością w ramach nakreślonych przez tą maszynę. *Tego* nie muszą zaś zakładać nawet zwolennicy mocnej AI, tzn. nie muszą utożsamiać prawdziwości z dowodliwością w danym systemie.[52]

Nie możemy *dowieść*, że maszyna nie ma możliwości ujęcia semantyki, albo czegoś równoważnego naszej semantyce. Czujemy jednak różnicę, więc możemy *założyć*, że jest ona nieusuwalna. I to właśnie czyni Lucas i ludzie podobnie myślący. Można to, moim zdaniem, uczynić w sposób niesprzeczny. (Por. IV.A.3 na temat rozumienia pojęcia liczby.) Nie sądzę więc, by dało się takie przekonanie obalić. Nie oznacza to jednak, że da się je udowodnić, a w szczególności udowodnić przez użycie twierdzenia Gödla.

3. Obycie się bez prawdziwości

Nie powinniśmy zakładać żadnego rozstrzygnięcia kwestii stosowności mocnej AI, czyli stosowalności kategorii prawdziwości do stosunku między maszynami (i przyszłymi maszynami) a obiektami językowymi. Argument przeciw Lucasowi powinien być w stosunku do tego neutralny. Najlepiej gdyby w ogóle nie mówić o prawdziwości.

Czy da się obyć bez pojęcia prawdziwości? Zupełnie niezadowalającym wyjściem byłoby przyjęcie radykalnego formalizmu, tzn. rezygnacja z pojęcia prawdziwości, przynajmniej w odniesieniu do matematyki (czy tylko do arytmetyki). Nie czynił tak nawet Hilbert. Gödel odrzucał to jawnie, podkreślając, że aby uprawiać matematykę, musimy mieć dostęp do „matematyki właściwej"[53], czyli pewnych prawd matematycznych. Jednak w naszym kontekście, by uczynić argument Lucasa jak najlżej strawnym (po to, by go potem nieodwołalnie obalić), możemy uznać, że maszyna albo ma dostęp do prawdziwości, albo tylko *udaje*.

Problem osiągnięcia przez maszyny możliwości ujęcia semantyki lub czegoś jej równoważnego jest otwarty, i to zarówno jako zagadnienie praktyczne, jak i jako kwestia czysto teoretyczna (czy *w zasadzie* może jakaś maszyna osiągnąć takie możliwości?). Możemy jednak uznać, że maszyna ma zielone światełko, które zapala tylko wtedy gdy na wyjściu pojawia się wyrażenie, które ona „przedkłada jako prawdziwe". Zamiast prawdziwości mamy światełko, udawanie, że chodzi o prawdziwość. Oczywiście zamiast sugestywnego światełka może być tak, że przy tych wyrażeniach pojawia się specjalny symbol oznaczający „prawdziwość".[54] Możemy tu ograniczyć się do wyrażeń w języku arytmetyki. Cokolwiek oznacza ich prawdziwość dla nas, cokolwiek może „prawdziwość" „znaczyć" dla maszyny, jeśli maszyna działa według ustalonego programu, zbiór tych wyróżnionych (światełkiem lub specjalnym symbolem) wyrażeń, które mogą być „przedłożone" przez maszynę wydaje się dobrze określony. Jeśli maszyna jest ustalona, ma ustalony program, według którego „przedkłada" twierdzenia arytmetyczne, tzn. generuje je na

---

[52] Lucas pisze: „Mechanicysta, uznając człowieka za mniej niż ludzi, bo za maszynę, uznaje jego koncepcję prawdy za coś mniej, a mianowicie, że jest dowodliwością w danym systemie" (Lucas [1968], 148). Slezak pisze, że widać tu, iż Lucas polemizuje z mechanicystą „słomianym", czyli kukłą, tzn. przeciwnikiem spreparowanym, ustawionym do polemiki (Slezak [1982], 45).

[53] Por. Gödel [1951] i omówienie tego w II.M.1.

[54] Tak czyni Penrose w [1994], choć na początku mówi o tym, że maszyna „ascertains truths". Symbolem używanym przez komputery jako znak „imprimatur" jest tam gwiazdka ([2000], 205).

wyjściu zapalając zielone światełko, czyli gdy jest równoważna (w zasadzie, w teorii) maszynie Turinga, ten zbiór twierdzeń będzie rekurencyjnie przeliczalny, a więc tworzy teorię (rekurencyjnie) aksjomatyzowalną.[55] Pozostaje problem, czy twierdzenie Gödla wyklucza możliwość, że jakaś maszyna produkuje dokładnie te wyrażenia arytmetyczne, które umysł ludzki jest w stanie uznać za prawdziwe. Tylko tego może ewentualnie dowieść argument w stylu Lucasa.

Sam Lucas stosuje pojęcie prawdziwości w istotny sposób – w krokach (L2), (L3) i (L4). Przed chwilą pokazaliśmy, że mógłby go nie używać w (L2). Potem (w II.K) zobaczymy, że możemy mu pomóc i w ogóle nie używać pojęcia prawdziwości – ale że to i tak nie uratuje argumentu.

Nasuwa się też komentarz historyczny. Gödel sformułował swoje twierdzenie tak, by można było zupełnie pominąć pojęcie prawdziwości, a użyć tylko niesprzeczności i ω-niesprzeczności (p. I.C.0 oraz III.B..2.g i III.D.3). My staramy się przeformułować pewne rozumowania tak, by nie trzeba było mówić o prawdziwości. To *déja vu* nie jest chyba przypadkowe. Pojęcie prawdziwości okazuje się podejrzane przy rozpatrywaniu maszyn. Pojęcie prawdziwości było też podejrzane przy rozpatrywaniu teorii matematycznych w szkole Hilberta. Wówczas chciano mechanizować matematykę (na potrzeby badań metamatematycznych), obecnie mechanizuje się matematykę na potrzeby badań w ramach sztucznej inteligencji. Wtedy nie przyniosło to oczekiwanego efektu (znalezienia rozstrzygalnego systemu obejmującego całą matematykę) z powodu twierdzenia Gödla. Argument Lucasa jest próbą pokazania, że i teraz nie będzie oczekiwanego efektu (stworzenia sztucznej inteligencji równej człowieczej) – również z powodu twierdzenia Gödla. Nic dziwnego, że próby w stylu Lucasa odczuwamy jako przedsięwzięcie ważne i atrakcyjne. Co nie znaczy, że mogą się powieść. A to z kolei nie oznacza, że program zbudowania prawdziwie sztucznej inteligencji może się powieść. Wiemy, że jest to o wiele trudniejsze zamierzenie niż się to wydawało jego pionierom (por. III.E.3).

F. Wokół (L3): Niesprzeczność maszyny i człowieka

Utworzenie formuły Gödla dla odpowiedniej teorii jest kluczowym punktem rozumowania Lucasa i każdego innego argumentu w tym stylu. Albo bowiem chodzi o bezpośrednie „wygödlowanie" tak jak w (L3), albo o odwołanie się do GII, albo też do innej formy twierdzenia o niezupełności, w szczególności twierdzenia Turinga o nierozstrzygalności problemu stopu, co z upodobaniem stosuje Penrose. Jednak są to metody w zasadzie równoważne (por. I.B.10).

Niedługa refleksja (niezbyt jednak obecna w pierwotnej pracy Lucasa [1961]) pozwala stwierdzić, że nośność filozoficzna argumentu w stylu Lucasa może być podważona przez co najmniej dwa zasadnicze fakty: po pierwsze, metoda tworzenia zdania Gödla jest całkowicie mechaniczna, po drugie – jej stosowanie opiera się na założeniu niesprzeczności teorii, do której stosujemy twierdzenie Gödla. Zostawiając kwestię pierwszą, czyli algorytmiczny charakter wygödlowywania, na potem (p. II.J), rozważmy teraz drugą, tzn. fundamentalną ważność założenia niesprzeczności.

[55] Jak już było wspomniane w II.D, to, że zbiór konsekwencji rekurencyjnie przeliczalnego zbioru formuł jest aksjomatyzowalny (w logice pierwszego rzędu) przez rekurencyjny zbiór formuł, jest treścią twierdzenia Craiga.

Otóż zdanie Gödla jest niezależne tylko wtedy, gdy teoria, dla której jest skonstruowane, jest niesprzeczna. Jeśli jest sprzeczna, to oczywiście i ono, i jego zaprzeczenie, jest dowodliwe. Rozumowanie prowadzone w punkcie (L3) można rozbić na dwa przypadki:

Przypadek I: Teoria S jest niesprzeczna. Wtedy używamy zdania Gödla, by wygödlować maszynę M.

Przypadek II: Teoria S jest sprzeczna. Wtedy Maszyna M jest zdyskwalifikowana.

Otóż zasadniczy problem polega, jak zobaczymy, na trudności w rozróżnieniu Przypadku I od Przypadku II (p. II.H). Jednak załatwienie samego Przypadku II nie jest wcale bezdyskusyjne. Zaczniemy więc od niego.

1. Sprzeczna teoria S się nie nadaje?

System S, który miałby być równoważny umysłowi, musi być niesprzeczny, stwierdza Lucas. Dlaczego? Jest to wyraz wiary w naszą racjonalność. Pomimo tego, że popełniamy błędy, racjonalność oznacza logiczność, a więc brak sprzeczności. Gdybyśmy wierzyli w dwa zdania sprzeczne, to zgodnie z zasadami logiki moglibyśmy dowieść dowolnego zdania. W sformułowaniu kroku (L3) wspomniana jest tautologia $A\wedge\neg A\rightarrow B$, ale zauważmy, że wystarczy mieć do dyspozycji dwie bardzo proste i podstawowe reguły wnioskowania, a mianowicie dołączanie i opuszczanie alternatywy: $A/A\vee B$ oraz $A\vee B,\neg A/B$.[56] Te reguły, jak się wydaje, oddają sposób, w jaki przeprowadzamy rozumowania.[57] Nawet mało wyrobiony logicznie człowiek wnioskuje w taki sposób.

Można jednak żywić wątpliwości. Przecież nie wyciągamy jako wniosku dowolnego stwierdzenia, a jest niewątpliwe, że ciągle popadamy w sprzeczności: zmieniamy zdanie, mówimy jednocześnie „tak" i „nie", inni wytykają nam, że właśnie powiedzieliśmy coś wręcz przeciwnego niż poprzednio. Co więcej, umysły mamy podobne, ale niekoniecznie prowadzi to do takich samych poglądów. Ludzie o podobnym stopniu racjonalności i podobnej wiedzy wyznają nieraz przeciwstawne tezy. U nas, czyli w naszych umysłach, w przeciwieństwie do systemów logiki klasycznej, sprzeczność nie prowadzi do przepełnienia (tzn. uznania dowolnego stwierdzenia).

a) Naprawiamy własne błędy

Lucas załatwia sprawę dwutorowo. Po pierwsze anegdotycznie: Czyż nie jest faktem, że ludzie są sprzeczni? „Z pewnością takie są kobiety i politycy" (Lucas [1961], 120). Zapamiętajmy mu tę opinię. Po drugie zaś podkreśla, że nasze sprzeczności są chwilowe, bo po wykryciu są poprawiane. „Odpowiadają chwilowym błędom w działaniu maszyny" (*ibidem*,121), a nie rzeczywistej sprzeczności. Jesteśmy omylni, ale sami naprawiamy omyłki. Maszyna o tych samych cechach nadal podlega twierdzeniu Gödla, jest bowiem w ostatecznym rachunku niesprzeczna.

Teza o samonaprawialności brzmi przekonywująco, ale to nie zamyka sprawy. Jest wątpliwe, czy Lucasowskie tłumaczenie jego własnej niesprzeczności, wynikającej z

---

[56] Wtedy, mając jak założenia zarówno A jak i $\neg A$, wnioskujemy najpierw z A formułę $A\vee B$, a następnie z niej i $\neg A$ wnioskujemy B.

[57] Należy wspomnieć, że istnieją parakonsystentne systemy logiki, w których nie ma tej drugiej reguły, a (niektóre) sprzeczności są dopuszczone. Pierwsi takie idee mieli Stanislaw Jaśkowski (w 1948) i Newton da Costa (w 1963). Obecnie jest wiele takich logik i szereg prób ich stosowania.

poprawiania się itd., jest sformalizowaneprzez *Cons*. Jakieś inne ujęcie, bardziej adekwatne, może natomiast z powodzeniem być dowodliwe – zauważa Webb.[58] Można jednak przyjąć, że Lucas się uprze, iż *Cons* odpowiada akurat jego poczuciu swojej niesprzeczności.

Niektórzy dyskutanci uważają, że nie ma przejścia od niesprzeczności maszyny M do niesprzeczności odpowiadającej jej teorii S.[59] Jest oczywiste, że maszyna może produkować najróżniejsze formuły. Należy rozróżniać pomiędzy niesprzecznością maszyny a niesprzecznością systemu S, a dokładnie arytmetycznej jego części. Oznaczmy przez 'S_{ar}' ogół twierdzeń arytmetycznych dowodzonych w S, tzn. przez maszynę M. Otóż nawet jeśli maszyna M jest niesprzeczna w swoim funkcjonowaniu, to może ona produkować dwa zdania sprzeczne, albo wprost zdanie „0=1". Tak jest, ale to nie szkodzi: gdy mówimy o niesprzeczności M czy S w kontekście argumentu Lucasa, chodzi nam o niesprzeczność części arytmetycznej, S_{ar}, teorii odpowiadającej naszemu umysłowi, która składa się nie tyle z jakichkolwiek zdań wypisywanych czy produkowanych przez maszynę, ale ze zdań przedkładanych „jako prawdziwe". Znów wraca problem prawdziwości. Przezwyciężony został przez wprowadzenie „zielonej lampki" i tego się możemy trzymać. Niesprzeczna, czy w ostatecznym rachunku niesprzeczna, ma być teoria S_{ar}, złożona ze zdań arytmetycznych produkowanych na wyjściu wraz z zapaleniem „zielonej lampki". Nie należy mylić poziomów.[60]

Zarzut oparty na obserwacji, że niesprzeczna, tzn. poprawnie funkcjonująca maszyna, może z łatwością produkować sprzeczność arytmetyczną, należy więc uchylić. Natomiast można postawić zarzut znacznie głębszy. Wedle prowokacyjnego sformułowania Putnama: może jesteśmy maszynami sprzecznymi?[61] Jak to być może?

b) Ukryta sprzeczność?

Jest niewątpliwe, że poprawiamy nasze błędy. Nie wynika jednak z tego, że jesteśmy fundamentalnie niesprzeczni. Być może, w zasobach naszych umysłów kryją się zasady myślenia, które w niesprzyjających okolicznościach prowadzą do sprzeczności? Nie widać, jak można by wykluczyć taką możliwość. Istnieje wiele przykładów sprzeczności, w jakie popadli wybitni myśliciele. Poza matematyką można i należy to naprawiać poprzez rozróżnienia pojęciowe. W gruncie rzeczy tak bywa nawet w matematyce, co pokazują prace Lakatosa.[62] Jednak mamy poczucie, że w matematyce można dojść do definitywnych rozróżnień. Choć Gödel uważał, że tak samo jest w każdym racjonalnym dyskursie (por. III.C), to jest możliwe, że w filozofii tak być nie musi: wskazywanie paradoksalnych konsekwencji założeń filozoficznych jest uprawnionym sposobem nie tylko analizy, ale i dążenia do zrozumienia systemów idei filozoficznych i społecznych; być może nie ma sposobu ostatecznego uporania się z tym problemem. Dobrym przykładem tego rozróżnienia jest okoliczność podkreślana przez Gödla: antynomie teorii mnogości są przezwyciężone i nie stanowią problemu dla matematyki; natomiast antynomie logiczne (semantyczne) są nadal wielkim problemem dla logiki. Tym bardziej tak być może w innych działach filozofii.

---

[58] Webb [1980], 194. Por. dyskusję różnych sposobów formalnego wyrażania niesprzeczności w IV.A.1.

[59] Np. Gandy w [1996], 134.

[60] Jasno widział to Turing. Np. fascynowało go, że komputer „działa doskonale deterministycznie na jednym poziomie, podczas gdy na innym produkuje coś pozornie losowego – dobry model do pogodzenia determinizmu i wolnej woli" (Hodges [2002], 314). Hofstadter w [1979] z rozróżniania poziomów uczynił główną metodę opisu skomplikowanych struktur.

[61] Jako wypowiedź ustna Putnama jest to cytowane przez Lucasa w [1961], 120.

[62] Szczególnie jego słynne studium o wzorze Eulera i pojęciu wielościanu (Lakatos [1976]).

Zakładamy tylko, iż utrzymywanie się paradoksów nie prowadzi do sprzeczności arytmetycznych, bo w matematyce paradoksalne konsekwencje są sygnałem katastrofy.

Pomińmy systemy filozoficzne, gdzie paradoksy są ważne i może nieuniknione. Ponieważ interesuje nas S_{ar}, spójrzmy tylko na matematyków. Nawet najwięksi mylili się i produkowali sprzeczności. Podstawy rachunku nieskończonościowego były sprzeczne, co słusznie wytykał bp Berkeley: pewne wielkości traktowano jako niezerowe, a za chwilę jako równe zeru. Upłynęły pokolenia zanim sprawa została zadowalająco wyjaśniona.[63] Euler, jeden z największych matematyków w historii, proponował wyjaśnienia, które brzmią dla nas żenująco naiwnie. Na przykład, że poprawność rachunku różniczkowego bierze się z kompensacji błędów, które znoszą się wzajemnie. Albo: „Nie przeczę, że stosując rachunek różniczkowy i całkowy popełnia się jakiś błąd; utrzymuję jednak, że ostatecznie błąd ten staje się nieskończenie mały i całkowicie niczym."[64] Logicy, którzy są szczególnie wyczuleni na sprzeczność, nie są wcale mniej na nią podatni. Powszechnie znany wśród filozofów jest przykład Fregego, którego system logiki okazał się sprzeczny. Oczywiście sprzeczności grożą nadal. Pamiętam kursującą wśród matematyków opowieść, jak dwie poważne grupy topologów pracując nad tym samym problemem doszły do przeciwstawnych wniosków. Znalezienie błędu było niezwykle trudne. Matematyka robi się coraz bardziej skomplikowana. Czy naprawdę możemy mieć pewność, iż każdy przypadek sprzeczności będzie usunięty? Nagle może pojawić się sprzeczność równie fundamentalna jak paradoksy teorii mnogości, ale tak uwikłana w zaawansowane teorie, że jej przezwyciężenie będzie dla nas niewykonalne. Na dodatek rosnąć będzie rola matematyki wywodzonej z empirii, szczególnie z doświadczeń komputerowych. Można sobie wyobrazić, iż sprzeczność gdzieś się pojawi, a mimo to w normalnych zakresach matematyka będzie funkcjonować po staremu i bez specjalnych stresów. W gruncie rzeczy tak już było, gdy pojawiły się paradoksy teoriomnogościowe.

Trudno absolutnie wykluczyć czarny scenariusz: pojawia się sprzeczność i nie wiadomo, jak ją wyeliminować. Jednak nie możemy poprzestać na takiej konstatacji. Bez wątpienia musimy przyjąć, że matematyka nie może zrezygnować z dążenia do niesprzeczności. Byłoby to jej śmiercią. Normalne teorie wydają się *niewątpliwie* niesprzeczne. Doświadczenie matematyczne wskazuje, że dostatecznie głębokie poznanie problemu i wprowadzenie odpowiednich rozróżnień eliminuje sprzeczność. Historia rachunku różniczkowego i całkowego jest doskonałym tego przykładem. Naszą naturalną i spontaniczną reakcją na zarzut naszej (matematycznej) sprzeczności jest gwałtowny sprzeciw. Możemy się mylić, ale nie jesteśmy sprzeczni. Będziemy szukać błędu, a nawet jeśli nie uda się go znaleźć, będziemy nadal przekonani, że musi tam gdzieś tkwić. Niesprzeczność, choć nie możemy jej być absolutnie pewni, jest dla matematyki czymś w rodzaju idei regulatywnej w sensie Kanta.

Twórcy rachunku nieskończonościowego wyczuwali, a może po prostu intuicyjnie wiedzieli, że sprzeczność, która się zawiera w ich sformułowaniach, jest pozorna: odnosi się do naszego sposobu wyrażania się, a nie do rzeczy samej. Niesprzeczność jako swego rodzaju idea regulatywna przyświeca zapewne wszelkiej naszej działalności intelektualnej, podlegającej rygorom logiki. W niektórych dziedzinach można dążyć do poskramiania sprzeczności przez podkreślanie metaforyczności ujęcia (jestem sobą i nie jestem sobą) lub

---

[63] Por. Boyer [1964], A. Robinson [1974]. Abraham Robinson stworzył nowoczesną ścisłą teorię nieskończenie małych, „analizę niestandardową", która według Gödla, „w tej wersji lub innej będzie analizą przyszłości" (Robinson [1974], ix, oraz [CW2], 311). Teoria Robinsona jest przystępnie streszczona też w: Krajewski [1976].

[64] Za Grattan-Guiness [1970], 8.

przez wskazanie na nieadekwatność wyrazu słownego w stosunku do omawianej materii (stwierdzenia, że jeden byt jest identyczny z trzema różnymi bytami, nie skłaniają nikogo do wysuwania arytmetycznej tezy „1=3”[65]). W zakresie liczb naturalnych sprzeczność odarłaby z sensu wszystko, co czynimy.

Bądźmy jednak uważni: można twierdzić, iż sprzeczność to śmierć, tylko w odniesieniu do sprzeczności jawnej. Sprzeczność głęboko ukryta może być mało groźna. Może być tak, że zdanie A ujawnia się w innych kontekstach niż zdanie $\neg$A.[66] Mamy obecnie doświadczenia, których pozbawione były poprzednie pokolenia. Wiemy, że duże programy komputerowe zawierają błędy („bugs”). Doświadczenie programistów wskazuje, że po pierwsze, w dużych systemach nie da się w praktyce wyeliminować błędów,[67] a po drugie, że nie przeszkadza to w codziennym funkcjonowaniu programu. Tylko w niecodziennych okolicznościach błąd się ujawnia i powoduje „sprzeczność” – zawieszenie programu, wykonanie nieprzewidzianych czynności, itp. Program jest więc zły, ale w praktyce jest wystarczająco dobry. Logicy nie są przyzwyczajeni do takich sytuacji[68]. Gdy w teorii jest sprzeczność, teoria jest od razu do niczego. Jednak faktycznie używane teorie matematyczne, podobnie jak programy, mogą być może zawierać ukryte, niejawne sprzeczności bez szkody dla ich funkcjonowania w normalnych warunkach. Można powiedzieć, że te metody są poprawne, niesprzeczne w *normalnym* zakresie ich stosowania. My natomiast jako istoty dążące do fundamentalnej niesprzeczności, usuwamy wszelkie wyłaniające się sprzeczności.

2. „Ja jestem niesprzeczny”

Slezak proponuje następującą analogię[69]: jeśli jakaś osoba nie jest w stanie schylić się i dotknąć swoich stóp, a Lucas mówi, że jest w stanie to uczynić, i na dowód tego dotknie stóp tej osoby, to będzie to tylko żart. Chodziło o to, czy Lucas może dotknąć *swoich* palców. Podobnie, gdy Lucas twierdzi, że może dowieść zdania Gödla dla jakiejś maszyny M, to nie ma w tym nic szczególnego, bo on jest inny niż M; chodzi o to, czy jest czymś jakościowo innym od maszyn. Problem więc polega na tym, czy jest w stanie dowieść zdania Gödla dla siebie samego. To jest użyte w argumencie Lucasa, który jest rozumowaniem nie wprost: gdybym był maszyną, dowiódłbym naszego wspólnego zdania Gödla, a ona tego nie może, więc nie mogę być (tą) maszyną. Jeśli pominąć kontekst argumentu Lucasa, to oczywiście jest w najwyższym stopniu niejasne, czym jest owo zdanie Gödla „dla mnie”. Możemy jednak spróbować pewnych analogii.

a) Uprawniony wniosek: nasza niesprzeczność nie jest dowodliwa

Rozważymy pewien wniosek z GII, który ma sens niezależnie od argumentu Lucasa. Jak wiemy, GII mówi, że w żadnej dostatecznie silnej i niesprzecznej teorii nie da się

---

[65] Jeśli stosować argument arytmetyczny, to, jak trafnie zauważył Życiński, można przywołać inny wzór. Sprzeczny z rozsądkiem wydaje się bowiem wzór $\aleph_0+\aleph_0+\aleph_0=\aleph_0$; podobnie – dodaje ten autor – jak teza, że „Bóg jest jeden, ale w trzech osobach” (Życiński [1985], 200).

[66] W. Robinson zauważa w [1992], 133, że jest do pomyślenia, iż okoliczności, w których stwierdza się jakieś zdanie, mogą nigdy nie wystąpić jednocześnie z okolicznościami, w których stwierdza się jego zaprzeczenie.

[67] De Millo et al. w [1979] pierwsi argumentowali, że w praktyce nie do przezwyciężenia są zasadnicze różnice między sprawdzaniem poprawności dużych programów a dowodzeniem rozpatrywanym w logice.

[68] Są, co prawda, wyjątki. Np. logiki parakonsystentne dopuszczają pewne typy sprzeczności (por. drugi przypis do II.F.1), a Graham Priest wprowadził pojęcie transkonsystencji.

[69] Slezak [1982], 50

udowodnić jej własnej niesprzeczności. Zastosujmy to do nas. Zakładamy, iż nasza niesprzeczność jest faktem. Wtedy twierdzenie wydaje się dowodzić, że nie możemy dowieść naszej własnej niesprzeczności.[70] Oczywiście, takie użycie twierdzenia matematycznego jest dziwaczne. Na pierwszy rzut oka takie przejście zdaje się *zakładać*, że człowiek jest maszyną, bo tylko wtedy można ewentualnie zastosować twierdzenie Gödla. Konkluzja nie miałaby więc uzasadnienia dla tych, którzy nie wierzą w to, że jesteśmy maszynami (czy raczej, że ogół twierdzeń, które my możemy dowieść, można wyprodukować przy pomocy jednej maszyny).[71] Jednak konkluzja o niemożności stwierdzenia naszej niesprzeczności może być dowiedziona subtelniej. Gdyby bowiem dało się udowodnić naszą niesprzeczność jakkolwiek, ale ściśle i nieodparcie, *more geometrico*, to ten dowód dałoby się sformalizować, czyli symulować na maszynie, która by zawierała odpowiednią, potrzebną do jego przeprowadzenia, część naszych mocy matematycznych. Taka maszyna tym bardziej dowodziłaby swojej niesprzeczności (bo zawierałaby tylko część tego, co dostępne dla naszych umysłów). Na mocy GII byłaby więc sprzeczna sama ta maszyna, a raczej odpowiadający jej system formalny. Tym bardziej – system obszerniejszy, odpowiadający naszemu umysłowi. Zakładając dowodliwość naszej niesprzeczności, dowiedlibyśmy naszej sprzeczności. Oznacza to, że takiego założenia uczynić nie można. *Nawet, gdy faktycznie jesteśmy niesprzeczni, nie da się tego matematycznie dowieść!* Jest to wniosek filozoficzny.

Są dwa nieoczywiste założenia użyte powyżej. Jednym jest to, że ścisły dowód naszej niesprzeczności daje się sformalizować, a drugim – że da się „naszą niesprzeczność" wyrazić. To pierwsze jest wnioskiem z ogólniejszej zasady: zupełna ścisłość oznacza możliwość formalizacji. Jeżeli ta ścisłość jest matematyczna, „geometryczna", to oznacza właśnie formalizowalność. Ta zasada wydaje się trudna do podważenia.[72] W takim razie w rozumowaniu powyższym przyczepić się można do tej drugiej kwestii: nie jest mianowicie jasne, jak *wyrazić* „naszą niesprzeczność". Są możliwe dwa przypadki: albo (i) jest to zdroworozsądkowe stwierdzenie „Jestem niesprzeczny", albo (ii) jego formalny odpowiednik.

Jeśli chodzi o przypadek (i), czyli o stwierdzenie zdroworozsądkowe, to jego związek z jakimikolwiek rozważaniami formalnymi jest niepewny. Hao Wang rozpatruje takie właśnie zdanie „Ja jestem niesprzeczny".[73] To nieformalne zdanie – uważa Wang – raczej nie jest dowodliwe. Ma tak być niezależnie od przeprowadzonego powyżej wywodu, że niesprzeczność, gdy ma miejsce, nie jest dowodliwa. Powód jest ogólniejszy i dość przekonywujący: nie wiemy, jak dokonywać formalnych wywodów, które by do takiego zdania „o nas" prowadziły. Jeśli by więc było dowodliwe, to w jakimś innym sensie. Ale w takim razie już sama możliwość przeprowadzania tych nieformalnych dowodów oznaczałaby, że (nawet w zakresie rozumowań) nie jesteśmy maszynami. Można w to wierzyć, ale jedyną podstawą jest ogólne poczucie.

Natomiast w przypadku (ii) mamy do czynienia z formalnym odpowiednikiem zdania zdroworozsądkowego, np. postaci „S_{ar} jest niesprzeczna", przy założeniu, że teoria S odpowiada moim mocom matematycznym. Wtedy omawiany wywód pokazuje, że nie da się tego dowieść, o ile faktycznie jestem niesprzeczny. Jest jednak wątpliwe, czy to akurat zdanie

---

[70] Wspomina o tym w swym pierwszym artykule Lucas ([1961], 120), powołując się na Hartleya Rogersa. Wydaje się oczywiste, że idea ta pochodzi od samego Gödla, choć teza ta nie jest wprost wypowiedziana w [1951]. (Por. jednak fragment w [CW3], 309, a także napisane po rozmowach z Gödlem uwagi Wanga w [1974], 319). Jest sformułowana u Wanga w [1974], 324, oraz potem, np. [1993], 119.

[71] To właśnie stwierdza Lucas w swym pierwotnym artykule ([1961], 124).

[72] Ma to związek z Tezą Churcha. Por. też II.J.1.

[73] Wang [1974], 317-320. Chodzi o zdanie „I am consistent", oznaczane jako „A".

formalne typu $Cons_S$ rzeczywiście oddaje treść zdania „ja jestem niesprzeczny.” Przecież chodzi nam o dążenie do unikania sprzeczności, o spójność naszej wizji świata, o co najwyżej – jeśli chcemy pójść w kierunku ujęcia formalnego – o zwykłe zdanie, że metody używane przez matematyków są niesprzeczne. Jakiekolwiek zdanie typu *Cons* jest od tego odległe.

Można przedstawić formalną wersję powyższego argumentu na rzecz tezy, że nie da się wykazać naszej niesprzeczności. Nie powinno budzić zdziwienia, że jest ona powtórzeniem rozumowania dowodzącego GII. Załóżmy, że nasze zdolności do ścisłego dowodzenia są oznaczone przez ‘$\vdash$’, a przekonanie o bezspornej prawdziwości zdania arytmetycznego przez ‘B’. Wtedy możemy się zgodzić, że są spełnione warunki Löba: $\vdash A \Rightarrow \vdash B(\ulcorner A \urcorner)$, $\vdash (B(\ulcorner A \urcorner) \wedge B(\ulcorner A \rightarrow C \urcorner) \rightarrow B(\ulcorner C \urcorner))$, $\vdash (B(\ulcorner A \urcorner) \rightarrow B(\ulcorner B(\ulcorner A \urcorner) \urcorner))$. Można zatem powtórzyć abstrakcyjną wersję dowodu GII (por. I.C.1.c i I.C.5), co pokazuje, że non $\vdash (\neg B(\ulcorner 0=1 \urcorner))$, czyli nie da się dowieść przekonania o naszej niesprzeczności. Inaczej mówiąc: jeżeli mogę *dowieść*, że jakikolwiek jawny fałsz f nie jest przedmiotem mego nieodpartego przekonania, czyli $\vdash \neg B(f)$, to popadam w sprzeczność.

Z takiego rozumowania Chalmers wyciąga wniosek, iż należy wycofać założenie, że „system wie, że jest adekwatny [sound]” (Chalmers [1995], 3.12). Lepiej byłoby rzec: nie można zakładać, iż „system dowodzi, że zna swą niesprzeczność”.

b) Sprzeczność a świadomość

A co się dzieje, jeśli *nie* jesteśmy niesprzeczni? Gdybyśmy byli sprzeczni, wszystko powinno być dowodliwe! Ale takie stwierdzenie – pisze np. Wang[74] – nie wydaje się przekonywujące. My nie działamy jak maszyna Turinga, nawet jeśli gdzieś u podstaw naszego działania kryje się coś w rodzaju maszyny Turinga. W ten sposób wracamy do problemu nieujawniającej się sprzeczności, który był już omówiony powyżej (w II.F.1). Sprzeczność może być gdzieś głęboko ukryta i nie wywoływać zgubnych konsekwencji w zwykłym życiu, tak jak poważne programy funkcjonują udatnie pomimo nieuniknionej obecności błędów. Może więc jesteśmy sprzeczni? Może jesteśmy sprzecznymi maszynami?!

Trudno zaprzeczyć, że wyobrażenie sobie siebie jako maszyny sprzecznej, budzi bunt. Nawet jeśli niesprzeczność znaczy co innego w odniesieniu do nas i w odniesieniu do maszyn, możemy wyobrazić sobie maszynę tak zaprogramowaną, by naśladować nasze myślenie. Ujawnienie sprzeczności nie powoduje wyprowadzenia dowolnego wniosku, ale zamiast tego powoduje poszukiwanie przyczyn sprzeczności i rewizję założeń, które do niej prowadziły. W tym sensie rachunek zdań nie jest częścią bezpośredniego mechanizmu rządzącego naszym myśleniem. Taktyka rewidowania założeń da się najprawdopodobniej odtworzyć w komputerach.[75] Konkluzja, że możemy być sprzeczni, choć nie da się jej wykluczyć, jest jednak mało przekonywująca – dla mnie i dla bardzo wielu osób. Lucas ma rację, że jakiekolwiek rozsądne modelowanie naszego myślenia musi jakoś zawierać rachunek zdań i elementarną arytmetykę, a zatem również przekonanie o niesprzeczności elementarnej arytmetyki.[76] Muszę dodać, że zgadzam się z obiekcją Lucasa, iż poważne przyjęcie tezy o sprzeczności, zawartej nieusuwalnie w centrum naszego rozumu, jest wyrazem irracjonalizmu. Wtedy racjonalna polemika z mechanicyzmem nie jest możliwa.[77]

---

[74] Wang [1974], 319.

[75] Wspominał o tym (jako pierwszy?) Wang [1974], s. 320, a potem np. Hofstadter [1979], 578, a później np. Gandy [1996] – por. jego uwagi o maszynach jako kolegach-matematykach (powyżej w II.F.1).

[76] Lucas [1996], 121.

[77] Lucas [1996], 121-2.

Nasza świadomość mówi nam, że nie jesteśmy sprzeczni – przynajmniej w zakresie działalności matematycznej lub matematyzowalnej. Istnieje cały sposób argumentacji, który polega na przeciwstawieniu umysłu teoriom formalnym i maszynom na podstawie tego, że umysły – w przeciwieństwie do tamtych obiektów – mają szczególną własność: umiejętność introspekcji. Mianowicie odpowiadają na pytania o siebie, nie ulegając zmianie. Niektóre uwagi Lucasa można rozumieć jako stosowanie, bądź nawiązanie do takiej wizji. Zdanie Gödla jako samozwrotne jest nierozstrzygalne wewnątrz teorii, bo teorie są właśnie ograniczone, nie mają możliwości introspekcji. Jednak ani teorie formalne nie są takie słabe, ani my – tacy mocni. Jak wiadomo[78], liczne teorie świetnie sobie radzą z niektórymi zdaniami samozwrotnymi, np. z formalnym odpowiednikiem zdania „ja jestem formułą dowodliwą" (mówi o tym twierdzenie Löba - por. I.C.4). Poza tym, *my* też nie potrafimy odpowiedzieć na wszelkie pytania dotyczące naszego umysłu. Zdanie o niesprzeczności ma status szczególny: rzeczywiście wydaje się, iż mamy na nie pozytywną odpowiedź tylko na podstawie wczucia się w nasz umysł. Jednak jest niewątpliwe, że w tej kwestii możemy się mylić. Jak widzieliśmy, zdarza się to najtęższym umysłom. Wtedy stosowanie wygödlowywania prowadzi do następnej sprzeczności. Zresztą w II.K jest pokazane, że wszelka procedura w tym stylu prowadzi, niezależnie od wszystkiego, do sprzeczności.

Może ktoś powiedzieć, że nie chodzi o żadne formuły i dowodliwość, ale o właśnie o to wczucie się, czyli świadomość. Otóż świadomość możemy tak rozumieć, że maszyny z definicji są jej pozbawione. Wtedy teza mechanicyzmu jest obalona, ale nie potrzeba do tego twierdzenia Gödla. Po prostu zakładamy to, co chcemy dowieść.

Niejasny jest związek inteligencji ze świadomością. Penrose pisze, że „zagadnienie inteligencji jest uzupełniające w stosunku do świadomości" (Penrose [1995], 447). Świadomość, w sensie samoświadomości, jest, być może, jeszcze trudniejsza do zrozumienia niż inteligencja. Można sobie wyobrazić byty (osobniki?) inteligentne, ale nie samoświadome. Może da się takie skonstruować? Gdyby tak było, to Penrose uznałby, że „zagadnienie inteligencji [go] w istocie nie interesuje" (*ibidem*). Ta uwaga dobrze ilustruje nasz obecny brak rozumienia fenomenu umysłu. To, czy maszyny mogą mieć świadomość, jest nadal w dużej mierze decyzją arbitralną. Turing i inni próbowali wyrazić tę kwestię w kategoriach bardziej uchwytnych.

Nieraz wydaje się, że Lucas zakłada to, co chciał dowieść. Np. wtedy, gdy zdaje się mówić: mogę przedłożyć zdanie Gödla jako prawdziwe, *nie dowodząc*. Dlaczego? Bo odpowiednia teoria (maszyna) jest niesprzeczna. Czemu? Bo „niesprzeczność to takie nieszczęście, z którym się nie pogodzimy" (Lucas [1968], 157). Nie sposób się z tym nie zgodzić, ale czy można widzieć w tym przekonywujący argument, że S nie jest sprzeczna?

Należy zaznaczyć, że dla Gödla nie ulegało wątpliwości, iż jesteśmy niesprzeczni. Jest to konsekwencją założenia zasadniczej racjonalności człowieka. W praktyce wszyscy ludzi skłonni do analizy argumentu Lucasa wierzą w swoją fundamentalną niesprzeczność. Jeśli mamy rację, to – jak wynika z wywodu podanego przed chwilą (w II.F.2.a) – nie dałoby się tego dowieść w sposób nieodparty. Wydaje się jednak, że możemy nadal zakładać naszą niesprzeczność i hipotetyczną równoważność maszynie (tzn. rekurencyjną przeliczalność S). W tej sytuacji argument Lucasa może być przeprowadzany. Należy więc kontynuować jego analizę.

Podsumowując, możemy więc uznać, że gdy jesteśmy niesprzeczni, to albo (a) nie jest to wyrażalne w sposób formalny, albo (b) jest, ale nie da się tego dowieść (chyba że

[78] Np. Wang [1974], 320.

stosowalibyśmy metody innego rodzaju niż te, które poddają się formalizacji). W przypadku (a) zakładamy, że umysł nie jest maszyną, w przypadku (b) dopuszczamy, że jest. Jeżeli (a), to cel argumentów w stylu Lucasa byłby osiągnięty (człowiek jest czymś więcej niż maszyną), ale w wyniku błędnego koła, bo niewiele jest dodane do wyjściowego przekonania intuicyjnego, że *oczywiście* nie jesteśmy maszynami. Jeżeli (b), to należy kontynuować rozbiór argumentu Lucasa.[79]

G. Wokół (L4): Skąd znamy prawdziwość zdania Gödla?

Zasadniczym krokiem w argumencie Lucasa jest stwierdzenie, że *my* widzimy prawdziwość formuły G. Lucas pisze w swym pierwotnym artykule – i podobnie mówią liczni zwolennicy metafizycznego zastosowania twierdzeń Gödla – że zdanie Gödla jest niedowodliwe w omawianym systemie, dla którego jest ono skonstruowane, ale „my, stojąc na zewnątrz systemu, potrafimy zobaczyć jego prawdziwość" (Lucas [1961], 113). Niektórzy myślą, że miałoby chodzić o prawdziwość w szczególnym sensie. Pozostawanie na zewnątrz systemu formalnego urastałoby do roli jakiegoś rewelacyjnego faktu, który w tajemniczy sposób daje nam możliwość uchwycenia niezwykłych prawd. Bo one muszą być niezwykłe, skoro nawet w bardzo mocnym systemie S nie da się ich dowieść. Nasza moc „widzenia prawdy" nabiera niemal mistycznego charakteru. Sądzę, że w takim postrzeganiu problemu leży źródło, może nawet główne źródło, atrakcyjności argumentu Lucasa w jego różnych wersjach. Tymczasem samo stanie na zewnątrz systemu nie daje nam jakichś niezwykłych przewag. Prawdziwość formuły Gödla nie jest jakąś specjalną, specyficzną prawdziwością, ale zwyczajną prawdziwością matematyczną, nie inną niż prawdziwość wyrażenia, że nie ma rozwiązania danego równania. Z kolei „bycie poza" nie jest niczym nadzwyczajnym: Slezak przypomina oczywisty fakt, że inna maszyna też jest na zewnątrz wyjściowej maszyny.

1. Twierdzenie Gödla a zdanie Gödla

Z jakiego powodu mówimy o tym, że z zewnątrz (systemu) lepiej widać? Otóż dowodząc, że formuła G jest prawdziwa, rozumujemy następująco: „W trakcie dowodu twierdzenia Gödla okazaliśmy, że G nie jest dowodliwa w S (o ile S jest teorią niesprzeczną). Ale G stwierdza, że ona sama, czyli G, nie jest dowodliwa w S. A zatem jest tak jak G mówi, czyli G jest prawdziwa." Porównując treść G z faktem o G, a mianowicie z niedowodliwością G w S, stajemy na zewnątrz S, bo mówimy o dowodzeniu w S, czyli o całej teorii S, traktowanej jako obiekt rozważań. Można jednak popatrzeć na sytuację inaczej.

a) Zwyczajna prawdziwość

Fakty, potrzebne do wyciągnięcia wniosku o G, dają się zarówno wyrazić jak i udowodnić wewnątrz S: $S \vdash (G \leftrightarrow \neg Pr_S(\ulcorner G \urcorner))$. Stawanie na zewnątrz systemu nie jest niezbędne. Co więcej, jak wspomniane było w I.B.9, można dowieść więcej niż równoważności. Formuła G została tak sprytnie skonstruowana, że sprawdzenie zachodzenia powyższej równoważności jest sprawą elementarną, tzn. teoretycznie prostą, nie odwołującą

[79] Rozróżnienie (a) od (b) wyjaśnia, dlaczego według Putnama, Lucas myli dwa zdania: „zwykłe zdanie, że metody używane przez matematyków są niesprzeczne" oraz „skomplikowane matematycznie zdanie, które powstaje, gdy zastosować metody Gödla do hipotetycznej formalizacji tych metod" (Putnam [1995], 370-371).

się do mocniejszych środków logicznych niż prosta arytmetyka. Powyższa równoważność jest więc nawet dowodliwa w Ar, która może być znacznie słabsza niż S.

Można również podobnie potraktować fakt, że niezależność zdania G jest warunkowa: jeśli S jest niesprzeczna, to S non $\vdash$ G_S. Po zarytmetyzowaniu, jak wiemy (p. I.C.1.b), dostajemy: $Cons_S \rightarrow G_S$. Jest rzeczą niezmiernie ważną, iż implikacja ta nie wymaga do dowodu mocnych środków logicznych. Wręcz przeciwnie – jest dowodliwa w słabej teorii:

$$Ar \vdash (Cons_S \rightarrow G_S).$$

Zarówno ta, jak i poprzednia zależność, czyli

$$Ar \vdash (G_S \leftrightarrow \neg\, Pr_S(\ulcorner G_S \urcorner)),$$

dowodliwa jest tym bardziej we wszystkich teoriach zawierających Ar. Ponadto, ponieważ nie wątpimy, że aksjomaty teorii Ar są prawdziwe, to wiemy, że zarówno powyższa implikacja jak i równoważność są prawdziwe. Jest to zwykła prawdziwość, chociaż jej dostrzeżenie przez nas nie jest zwykłym postępowaniem matematycznym.

Prawdziwość G jest zwykłą prawdziwością formuł arytmetycznych. Jakiś arytmetyczny supermózg mógłby postrzegać tę prawdziwość podobnie jak w przypadku innych skomplikowanych formuł. Chodzi bowiem o to, czy istnieje rozwiązanie pewnego konkretnego równania diofantycznego. Jednak to, że *my* widzimy tę prawdziwość, wymaga rozumienia z poziomu metateorii, bo zależy od naszej wiedzy (logicznej) dotyczącej teorii S, a mianowicie wiedzy o jej niesprzeczności. Prawdziwość G jest bowiem warunkowa: zachodzi, *o ile* teoria S jest niesprzeczna. Sama formuła G stwierdza niedowodliwość pewnej formuły. Twierdzenie Gödla – jeśli je formułować mówiąc o prawdziwości – nie brzmi: „G jest prawdziwa", ale „G jest prawdziwa, *o ile* S jest niesprzeczna." Istotą rzeczy jest więc nie samo „bycie na zewnątrz", ale wiedza, że S jest niesprzeczna. Jeżeli więc Lucas chce wygödlować, to musi wiedzieć czy widzieć, że system S jest niesprzeczny. Skąd można to wiedzieć? Nie jest to matematycznie proste. Wymaga wiedzy o naszej niesprzeczności, a to cofa nas do sytuacji rozważanej poprzednio w II.F.

b) Argument Putnama

Putnam był pierwszym autorem, który napisał,[80] iż do „wygödlowywania" trzeba wiedzieć, że jeśli S jest niesprzeczna, to o tym wiemy, czy to „widzimy", natomiast nie wystarczy, że wiemy, „widzimy", iż jeśli S jest niesprzeczna, to prawdziwe jest G_S. Innym słowy z tego, że

wiadomo, iż $(Cons_S \rightarrow G_S)$,

nie wynika, że

$Cons_S \Rightarrow$ wiadomo, iż G_S.[81]

Powód jest następujący: brakująca przesłanka to właśnie wiedza dotycząca niesprzeczności odpowiedniej teorii, czyli:

wiadomo, że $Cons_S$.

---

[80] Putnam [1960], 77.

[81] Putnam w [1960] pisze o elementarnej arytmetyce, a nie o dowolnej teorii S, ale w ogólniejszej sytuacji jest tak samo. (Tego uogólnienia nie dokonał, jak się wydaje, Yu w [1990], 148, i dlatego problem niesprzeczności elementarnej arytmetyki Peana przedstawia jako bardzo poważne zagadnienie, choć – przynajmniej w kontekście argumentu Lucasa – tak nie jest, bo praktycznie nikt nie wątpi w niesprzeczność akurat tej teorii.)

Ten argument można zresztą uprościć. Rolę G_S może bowiem, jak wiemy, odgrywać $Cons_S$. Wtedy jest jasne, iż choć

wiadomo, że ($Cons_S \rightarrow Cons_S$),

to niekoniecznie zachodzi:

$Cons_S \Rightarrow$ wiadomo, że $Cons_S$.

Problemem jest więc ustalenie prawdziwości tej przesłanki, czyli udowodnienie $Cons_S$. Nawet gdy teoria S jest niesprzeczna, możemy nie mieć dostatecznych podstaw, by to wiedzieć! Ustalenie niesprzeczności jakiejś teorii może być bardzo trudne. Weźmy dla przykładu teorię mnogości NF Quine'a.[82] Otóż nie wiemy, czy jest ona niesprzeczna, czyli nie potrafimy rozstrzygnąć, czy prawdziwe jest arytmetyczne zdanie $Cons_{NF}$. Nie pomoże tu ani stanie „na zewnątrz", ani prześledzenie dowodu Gödla, ani głębia myślenia na różnych poziomach na raz. Nie powinna nas niepokoić okoliczność, że zdanie Gödla jest zdaniem arytmetycznym, a więc wydawałoby się, że powinno być łatwe do rozstrzygnięcia. Nie jest łatwe właśnie dlatego, że jest w nim zakodowany złożony fakt dotyczący teorii S. Na czym polega więc prawdziwość zdania Gödla?

c) Prawdziwość G

Najprościej jest, gdy znowu zamiast formuły G weźmiemy formułę $Cons_T$. Otóż z jednej strony jej prawdziwość oznacza, że teoria T jest niesprzeczna. Od początku wiemy, że tak właśnie formuła ta została skonstruowana. Przyjrzyjmy się jednak budowie tej formuły. Otóż ma ona postać

($\forall x$) (x nie jest dowodem formuły „0=1" w teorii T).

Zresztą oryginalna pierwotna formuła Gödla, G, ma ten sam kształt, ale zamiast formuły sprzecznej „0=1" widnieje opis formuły, która okazuje się formułą G. Stosują się do niej wszystkie wnioski wynikające z kształtu formuły. Powyższa formuła da się zapisać jako formuła arytmetyczna, dzięki zastosowaniu formuł reprezentujących użyte tam pojęcia: formuła, dowód w teorii T. Otrzymujemy formułę postaci

($\forall x$)$D(x)$,

gdzie $D(x)$ (oznaczenie wprowadzone na użytek niniejszego rozumowania; czyt. „liczba x jest dobra") jest formułą ograniczoną (wszystkie kwantyfikatory są ograniczone do x). Zresztą D(x) może mieć postać p(x)=0, gdzie p jest pewnym wielomianem, a 'x' traktujemy teraz jako ciąg zmiennych. Co oznacza jej prawdziwość? Zgodnie z ogólną zasadą, prawdziwość formuły z kwantyfikatorem ogólnym to tyle co prawdziwość każdego podstawienia. Prawdziwość formuły $Cons_T$ oznacza więc

prawdziwość $D(0)$, $D(1)$, $D(2)$, ….

---

[82] Zaproponowana w Quine [1937] jako „nowe podstawy (New Foundations - stąd nazwa NF) dla logiki matematycznej", jest aksjomatyczną teorią mnogości, w której aksjomaty istnienia zbiorów definiowalnych przez daną formułę są ograniczone do formuł stratyfikowalnych, tzn. takich, w których da się w taki sposób przypisać liczby zmiennym występującym w formule, że w każdej podformule postaci $x \in y$ liczba przypisana x jest o jeden mniejsza niż liczba przypisana y. Mimo podobieństwa do teorii typów nie udało się dotąd dowieść niesprzeczności NF, nawet przy założeniu niesprzeczności mocnych teorii mnogości w stylu ZF. Częściowy wynik uzyskał Jensen: lekka modyfikacja teorii przez dopuszczenie „urelementów" pozwala na dowód jej niesprzeczności przy założeniu niesprzeczności teorii typów, co wynika z niesprzeczności fragmentu teorii ZF. (Nieco więcej jest w moim haśle w Marciszewski [1987], 117-119; nowsza literatura np. w Holmes [2001].)

Co to oznacza? To, że ani zero, ani jeden, ani dwa, ani żadna inna liczba nie jest numerem dowodu sprzeczności w teorii T. Innymi słowy, że nie ma dowodu sprzeczności w T, czyli, że T jest niesprzeczna. Zakładając niesprzeczność T, otrzymujemy prawdziwość formuły $Cons_T$. I na odwrót: z prawdziwości tej formuły wynika niesprzeczność tej teorii. Prawdziwość formuły $Cons_T$ jest wynikiem naszego założenia, a nie jakiegoś szczególnego wglądu.

Takie same uwagi można zastosować do zdania Gödla w oryginalnej postaci. Widzimy więc, że prawdziwość gödlowskiego zdania niezależnego można przedstawić jako naturalne przejście od prawdziwości formuł $D(0)$, $D(1)$, $D(2)$,… do prawdziwości formuły $(\forall x)D(x)$, czyli od prawdziwości dla każdej z liczb do prawdziwości dla wszystkich liczb. Problem prawdziwości formuły Gödla (w przeciwieństwie do prawdziwości *twierdzenia* Gödla, która nie budzi wątpliwości) sprowadza się do kwestii, czy wiemy, że rozważana teoria T jest niesprzeczna.

2. Zastrzeżenia co do prawdziwości zdania Gödla

Zgłoszone zostały pewne zastrzeżenia dotyczące sensowności stwierdzenia, że formuła Gödla jest prawdziwa, ale niedowodliwa. Jak wiemy, twierdzenie Gödla można wyrazić tak: jeśli T jest niesprzeczna, to T non $\vdash Cons_T$. Stwierdziliśmy, że założenie niesprzeczności oznacza prawdziwość formuł $D(n)$, a ponieważ są one logicznie stosunkowo proste (mają tylko kwantyfikatory ograniczone), to są dowodliwe w słabej formalnej arytmetyce, np. w Ar: Ar $\vdash D(0)$, Ar $\vdash D(1)$, Ar $\vdash D(2)$, … . Zarazem T non $\vdash (\forall x)D(x)$. Formuła $(\forall x)D(x)$ jest więc niedowodliwa, ale prawdziwa w tym sensie, że każde jej podstawienie (tzn. podstawienie (nazwy) liczby naturalnej za zmienną „x”) jest prawdziwe. Według Goodsteina, taki sposób mówienia jest „mylący” (Goodstein [1963], 218). Formuła $(\forall x)D(x)$ *nie* wyraża, wedle niego, pojęcia „$D(m)$ dla każdej liczby m” wewnątrz systemu formalnego. Powodem jest to, że system formalny może być interpretowany w modelach niestandardowych, gdzie nie wszystkie podstawienia formuły $D(x)$ zawierają się w (standardowym) ciągu: $D(0)$, $D(1)$, $D(2)$, … . Tak więc Goodstein wydaje się twierdzić, że formuła ogólna w systemie formalnym odnosi się do wszystkich jego interpretacji (realizacji). Jej prawdziwość oznaczałaby prawdziwość w każdej interpretacji, czyli – jak wiadomo z twierdzenia o pełności – formalną dowodliwość w systemie.

Pochodną powyższej interpretacji kwantyfikatora ogólnego jest podważanie rozumienia formuły Gödla jako formuły odnoszącej się do samej siebie. Ma ona bowiem kształt $(\forall x)C(x)$, przy czym o formule $C(x)$ można powiedzieć to samo, co zostało właśnie stwierdzone o formule $D(x)$. Jak wiemy, $C(x)$ odpowiada stwierdzeniu „x nie jest dowodem formuły G w teorii T”, a zarazem G jest formułą $(\forall x)C(x)$. Otóż dla kolejnych liczb naturalnych m zachodzi (tzn. jest prawdziwe i dowodliwe) $C(m)$, czyli „m nie jest dowodem formuły G w teorii T”, ale to nie znaczy, że możemy stąd wnioskować na temat prawdziwości formuły ogólnej $(\forall x)C(x)$. Ona „nic nie mówi” (Goodstein [1963], 219) na temat G. Utrzymywanie, że mówi, jest przezeń określone jako „inny popularny błąd”. Wydaje mi się, że bardziej konsekwentne niż stwierdzenie, że „nic” nie mówi, byłoby powiedzenie, że stwierdza ona nieistnienie jakiegokolwiek dowodu G w teorii T. Jakiegokolwiek – czyli w dowolnej realizacji: przecież kwantyfikator $(\forall x)$, użyty wewnątrz systemu, odnosi się do dowolnych możliwych interpretacji. Wtedy jest to ten sam problem (wszystko jedno, czy widzimy w nim błąd, czy nie), co poprzednio: istnienie modeli niestandardowych powoduje, że z punktu widzenia systemu formalnego czym innym jest $(\forall x)C(x)$, a czym innym koniunkcja formuł C(m) dla m=0,1,2,… . Przy takim rozumieniu jest to formuła fałszywa, bo

w pewnych modelach są (niestandardowe) dowody formuły G. (Gdyby ich nie było, dowodliwa byłaby formuła ¬G, co nie zachodzi przy założeniu ω-niesprzeczności lub innych tego typu założeniach adekwatności, a jeśli użyć formuły Rossera – bez żadnych dodatkowych założeń poza niesprzecznością).

Co sądzić o krytyce Goodsteina? Otóż ma ona sens, o ile odrzucamy istnienie wyróżnionego modelu standardowego. Wtedy wszystkie modele są równie dobre, a raczej równie złe. Prawdziwość w każdym modelu oznacza dowodliwość wewnątrz teorii. Dowodzenie w ramach teorii jest wtedy jedynym sposobem na bycie prawdziwym. Goodstein wyrażałby więc taką tezą swoje przywiązanie do formalizmu w filozofii matematyki. Jest to jednak nie do pogodzenia z normalną praktyką matematyczną.[83]

Oprócz tego z konsekwentnie formalistycznej perspektywy można utrzymywać, że formuła w ogóle nic nie mówi (tzn. że nie tylko nie mówi o sobie, ale w ogóle o niczym), a sens pojawia się dopiero na innym poziomie językowym. Wielość poziomów to inny częsty temat komentarzy.

3. Formuła G na różnych poziomach języka

Problem samozwrotności formuły G można rozpatrywać rozróżniając poziomy, na których używamy formuł, zdań, twierdzeń. Jest niewątpliwe, że przy rozważaniu formuły Gödla operujemy na różnych poziomach. Jest to po pierwsze poziom F, poziom systemu formalnego T, w którym są niezinterpretowane formuły. Następnie jest poziom M, poziom metamatematyczny, na którym mówimy o formułach, dowodach formalnych, niesprzeczności i innych własnościach systemu T. Wreszcie jest poziom A, poziom zwykłej matematyki, w szczególności arytmetyki, na którym mówimy o liczbach, podzielności, ciągach liczb. Mamy więc formułę Gödla G, postaci $(\forall x)C(x)$, która ma numer gödlowski g. Mamy jej odpowiednik G^M na poziomie M:

G^M: „żaden obiekt nie jest dowodem formuły G”,

czyli „G jest niedowodliwa w T” oraz G^A, odpowiednik G na poziomie A:

G^A: „żadna liczba nie jest kodem ciągu, który … i kończy się liczbą g.”

(Trzy kropki symbolizują opis teorioliczbowych własności ciągów liczb, które kodują dowody formalne teorii T; na przykład to, że liczba g jest końcowym wyrazem ciągu, to stwierdzenie, że g+1 jest wykładnikiem najwyższej potęgi liczby pierwszej, przez którą dzieli się liczba kodująca ten ciąg.) Formuła G^A bezpośrednio mówi o podzielności i zależnościach arytmetycznych pomiędzy liczbami, ale wiemy, że używając numeracji gödlowskiej dostajemy taką treść G^A, która odpowiada zdaniu G^M:

„żadna liczba nie jest numerem dowodu formalnego (w T) formuły (o numerze) g.”

Choć więc G^A nie mówi bezpośrednio o formułach i dowodach, to dzięki numeracji gödlowskiej jest równoważna formule G^M. Można to nazwać „ekstensjonalną równoważnością.”[84] Liczba g nie jest w G dana wprost, ale jako wartość pewnej funkcji pierwotnie rekurencyjnej, której opis jest zawarty w G. Można powiedzieć, że G^A jest zamierzoną interpretacją G w dziedzinie liczb naturalnych. Dokonawszy tych rozróżnień stwierdzamy, że G^M nie jest samozwrotna, bo mówi o G, a nie o sobie. Również G^A nie jest

[83] Inną próbę podważenia standardowego rozumienia formuły Gödla podjął Detlefsen w [1979]. Jego teza jest streszczona poniżej w IV.A.1.

[84] Termin użyty w Lacey i Joseph [1968].

samozwrotna, bo mówi o liczbie g, a nie o sobie. Sama formuła G nie jest samozwrotna, o ile uznać, że jest niezinterpretowaną formułą języka formalnego, która z założenia nie mówi o niczym.

Jeśli natomiast rozpatrywać system T jako zinterpretowany, to odpowiednie zmienne przebiegają liczby naturalne, a formuła G mówi o liczbie g (jako wartości odpowiedniej funkcji, której nazwa lub opis jest zawarty w G). Liczba g jest numerem Gödla formuły G, więc G mówi w ten sposób o sobie. Mówiąc dokładniej, G interpretujemy jako G^A, a dzięki ekstensjonalnej równoważności mamy też interpretację G jako G^M, która mówi o formule G.[85]

Niektórzy autorzy, np. Goodstein[86], uważają, że rozróżniając poziomy, nie możemy mówić o samozwrotności formuły G. Rzecz znowu sprowadza się do tego, czy jako zamierzoną interpretację formalnej arytmetyki, w której konstruujemy G, bierzemy pod uwagę zwykłe liczby i różne możliwe ich użycia – też jako kodów pojęć z poziomu metateorii, czy nie. Jeśli nie, to faktycznie G nie mówi o sobie, bo nie mówi o żadnych formułach, a co najwyżej o liczbach; to, że – jak się okazuje – odpowiadają one obiektom metamatematycznym, jest dodatkowym sensem, wniesionym z zewnątrz do arytmetyki. Samo zdanie nie odnosi się do swojego sensu, podobnie jak zdanie „to jest napisane kredą" (Goodstein [1966], 219). Jeśli natomiast włączamy całą interpretację, to G mówi o sobie, bo po to Gödel, a za nim my wszyscy, robił arytmetyzację metamatematyki, by liczby oznaczały też formuły, a niektóre związki między liczbami można było interpretować jako relacje metamatematyczne.

To drugie podejście, a nie „konfuzja" jest powodem, dla którego np. Andrzej Mostowski pisze: „dla dowolnego zdania Z teorii T istnieje arytmetyczne zdanie Z' (…), które mówi, że Z jest niedowodliwe. Nie ma nic paradoksalnego w tym, że dla odpowiednio dobranego Z zdanie Z' – by tak rzec – przypadkowo okazuje się identyczne z Z."[87] Termin „przypadkowo" jest użyty nieprzypadkowo, bo występuje w tym kontekście już w oryginalnej pracy Gödla.[88]

Jeszcze dalej idące wnioski z rozróżniania poziomów wyprowadza Wittgenstein. Jest to dla niego tylko przykład ogólniejszego zjawiska: tak samo zapisana formuła, ujmowana w ramach różnych teorii, nie jest tą samą formułą (p. IV.B.3).

4. Przykład Whiteleya i zawarte w nim nieporozumienie

Nierozróżnianie między naturą prawdziwości zdania Gödla a sposobem ustalania jego prawdziwości przez nas jest źródłem nieporozumień. Prawdziwość G jest zwykłą arytmetyczną prawdziwością. Jednak postrzeganie jej prawdziwości przez *nas* jest związane z ujęciem sytuacji z poziomu metateorii. Jest to wynikiem arytmetyzacji, jest więc zrelatywizowane do konkretnej numeracji gödlowskiej.[89] Gdy ktoś uznaje, że prawdziwość zdania Gödla to szczególna, samozwrotna prawdziwość, widzi rzecz w sposób podobny do

---

[85] Fitzpatrick w [1966] przedstawia wywód Gödla używając na różnych poziomach różnych języków: łaciny (do wyrażeń formalnych), francuskiego (do zdań z metateorii) i angielskiego (do stwierdzeń o liczbach).

[86] Goodstein [1966], a za nim Webb [1968], 166; p. też Lacey i Joseph [1968].

[87] Mostowski [1952], 9. W oryginale mowa jest o zdaniu S i teorii (S), ale te oznaczenia zmieniam dla przejrzystości.

[88] W [1931], w przypisie 15: „gewissermaßen zuffällig" ([CW1], 150).

[89] Tę relatywność, odrębną od wewnętrznych własności formuły arytmetycznej podkreśla Webb [1968], 163.

znanych antynomii semantycznych. To wydaje się leżeć u podstaw przykładu Whiteleya[90], który został podany w celu podważenia argumentu Lucasa, a mianowicie pokazania, iż Lucas – i oczywiście każdy z nas – podlega podobnym ograniczeniom, co maszyny. Można bowiem zrobić podobne do gödlowskiego zdanie, które może stwierdzić każdy poza Lucasem. Zdanie to, które nazwijmy W, jest następujące:

W: „To zdanie nie może być uznane przez Lucasa bez sprzeczności."

Otóż Lucas nie może uznać [assert], czyli przedłożyć jako prawdziwego, tego zdania bez sprzeczności, bo gdyby to zrobił to by popadł w sprzeczność. Nie może, ale – czy raczej: a więc – to zdanie jest prawdziwe. Pewne prawdziwe zdanie umyka więc możliwościom Lucasa. Każdy inny człowiek (a może i dostatecznie zaawansowana maszyna?) może bez sprzeczności uznać, przedłożyć prawdziwe zdanie W

Ciekawe, że Lucas nie komentuje tego ataku. Zapewne nie wie, jak się z tym uporać. Proponowane zdanie jest podobne do antynomii kłamcy. Mamy bowiem:

$W \Leftrightarrow$ (Uznanie W przez Lucasa $\Rightarrow 0=1$)

Antynomię kłamcy można ująć tak: $K \Leftrightarrow$ (K jest prawdziwe $\Rightarrow 0=1$).

Wyjściem z sytuacji, sugeruje Whiteley Lucasowi, jest stwierdzenie zdania W bez wspominania go, np. przez powiedzenie: „Nie mogę uznać bez sprzeczności zdania Whiteleya". Wedle Whiteleya moglibyśmy spowodować, by maszyna uznała, czy raczej – używając naszej terminologii – „przedłożyła jako prawdziwe" zdanie Gödla zrobione dla niej bez użycia tego zdania.[91] To powinno być możliwe. Wszystko jest sprowadzone do semantycznego triku.

Tymczasem pośrednie uznanie, stwierdzenie, czy przedłożenie formuły Gödla nic nie pomoże. Ona jest wyrażeniem pewnego faktu arytmetycznego. Użyte w niej predykaty są zupełne normalne, a nie semantyczne, a samozwrotność jest pośrednia i nie jest cechą definicyjną zdania, a konsekwencją – „niejako przypadkową" (by jeszcze raz przypomnieć sformułowanie Gödla) – jego konstrukcji. Prawdziwość zdania G nie ma nic wspólnego z tym, kto i jak je wypowie. W tym jest ono zasadniczo odmienne od zdań antynomialnych sformułowanych w języku potocznym.

5. Czy każdy może wygödlować maszynę?

Jest jeszcze jeden niepokojący element w argumencie Lucasa. Mianowicie stosuje się on tylko do ludzi logicznie wykształconych. Skoro kluczowym krokiem jest stwierdzenie prawdziwości zdania Gödla, to ile osób wchodzi w grę? To nie jest takie łatwe![92] Czyżby pozostali nie podlegali zbawczej sile argumentu i nie było pewne, czy oni również są lepsi od maszyn?

Jest kilka sposobów obrony omawianego argumentu. Po pierwsze można by rzec, że chodzi o wszystkich, którzy byliby w stanie zrozumieć rozumowanie Gödla, gdyby się tym zajęli. Ale to nie jest zadowalające. Czyżby ci, co nie są w stanie, mogliby być maszynami, a

---

[90] Whiteley [1962]. Cytuje go bez dystansu – jako jeden z argumentów, że obalanie mechanicyzmu się nie powiedzie – Hofstadter w [1979], 477.

[91] „we must provide it with some way of assering what the Gödel formula asserts without using the formula" (Whiteley [1962], 61).

[92] Takie zastrzeżenie wysunął Coder w [1969].

w każdym razie nie byli dowodliwe niemechaniczni?! Można z kolei powiedzieć, że użyjemy argumentu pośrednio: ludzie są do siebie dostatecznie podobni, więc udowodnienie czegoś o umyśle Lucasa pośrednio świadczy o umyśle każdego. Jednak ten kij ma dwa końce, bo to, że niektórzy ludzie nie są w stanie zrozumieć twierdzenia Gödla, pokazuje właśnie różnice pomiędzy naszymi umysłami. Można następnie twierdzić, że wymyślenie twierdzenia było bardzo trudne (czy gdyby nie Gödel, mielibyśmy je dziś do dyspozycji?), ale samo jego stosowanie nie wymaga głębszego zrozumienia. Wystarczy powołać się na to, że Gödel zrobił to, co trzeba. Jednak i ten kij ma zaskakująco nieprzyjemny drugi koniec: jeśli nie trzeba rozumieć, to czy nie może posłużyć się nim również maszyna? Przeciw innej maszynie, albo i przeciw nam… (Ta kwestia jest podjęta poniżej w II.J i II.K.)

Ci, którzy stosują argument w stylu Lucasa, myślą o sobie; z reguły są w stanie lub wydaje im się, że są w stanie zrozumieć Gödla. Prawdziwi zwolennicy mechanicyzmu i tak się tym nie przejmą. Mogą zresztą wskazać, że skonstruowanie zdania Gödla jest w praktyce bardzo trudne. Mając do dyspozycji tylko program działania maszyny, nawet najmędrsi spośród nas mogą nie być w stanie wygenerować odpowiedniego zdania. Jednak przeciwnik mechanicyzmu może tę sytuację wykorzystać na swoją korzyść. Mianowicie zdanie Gödla istnieje, czyli mamy je w teorii, choć nie w praktyce. A zatem spełnia ono swą rolę zarówno w odniesieniu do logików, jak i przedszkolaków.

Nie wiem, czy istnieje naprawdę przekonująca obrona uniwersalnej stosowalności argumentu Lucasa. Mam zarazem poczucie, że jest to mało istotne. Oczywiście, dzieło obalania argumentu Lucasa należy kontynuować nawet wtedy, gdy się je stosuje tylko w odniesieniu do logików, czy wręcz tylko do Lucasa.

Jest dość jasne, że argument Lucasa nikogo nie nawraca. Ci, którzy i tak wierzą w niemechaniczność umysłu, chętnie widzą taki matematyczny dowód swojej słuszności, ale raczej nie będą gotowi rozróżniać ludzi i ich umysłów według tego, czy rozumieją twierdzenie Gödla, czy nie. Ci, którzy wierzą, że maszyna może być równoważna naszemu umysłowi, nie przejmą się zanadto argumentem w stylu Lucasa.

Podsumujmy wyniki dotychczasowych rozważań o kolejnych krokach w argumencie Lucasa. Jeśli chodzi o (L1), to mimo niejasności związanej z zakresem pojęcia maszyny, pozostaje faktem, że rozumowanie Lucasa stosuje się przynajmniej do tej nadzwyczaj obszernej klasy maszyn, które (jako struktury wyidealizowane) są równoważne maszynom Turinga. Jeśli chodzi o (L2), to mamy do czynienia z ekwiwokacją, bo używa się jednocześnie wyrażenia odpowiedniego dla maszyny („produkuje”) i wyrażenia odpowiedniego dla ludzi („prawdziwe”). Spór filozoficzny o to, czy maszyna może operować pojęciem prawdziwości i innymi pojęciami semantycznymi, nie jest rozwiązany, a może nie jest rozwiązywalny. Jeśli założymy, że „prawdziwa” prawdziwość nie jest dostępna maszynom, jest natomiast dostępna ludziom, to argument Lucasa nie jest potrzebny, bo po prostu zakładamy naszą wyższość nad maszynami, czyli to, czego mieliśmy dowieść. Z drugiej strony, gdyby obalenie argumentu Lucasa polegało po prostu na założeniu, że dostatecznie rozwinięte maszyny są w stanie rozumieć, to byłoby to zignorowaniem argumentu. To zagadnienie można obejść, uznając, że maszyna ma zielone światełko, które zapala tylko wtedy, gdy na wyjściu pojawia się wyrażenie, które ona „przedkłada jako prawdziwe”. Przechodząc do (L3), rozróżniamy Przypadek I (teoria S jest niesprzeczna), gdy używamy zdania Gödla, by wygödlować maszynę M, oraz Przypadek II (S jest sprzeczna), a wtedy maszyna M jest zdyskwalifikowana. Jeśli chodzi o Przypadek II, to można zauważyć, że na ogół łatwo popadamy w sprzeczności. Pomińmy życie codzienne i filozofię, bo interesuje nas przede wszystkim ogół uznawalnych przez nas stwierdzeń arytmetycznych S_{ar}. Nawet ograniczając się do arytmetyki, nie możemy być pewni niesprzeczności, bo i matematycy popadają w sprzeczności. Jednak wtedy szukamy błędu. Niesprzeczność, choć

nie możemy jej być absolutnie pewni, jest dla matematyki niezbędną ideą regulatywną. Tezy „jestem niesprzeczny" zapewne nie dałoby się dowieść w sposób nieodparty. Mimo to sensowne jest, by zakładać naszą niesprzeczność i hipotetyczną równoważność maszynie. W związku z (L4), zauważamy, iż prawdziwość zdania G dla S_{ar} jest wynikiem naszego założenia, a nie jakiegoś szczególnego wglądu. Problem prawdziwości zdania Gödla (w przeciwieństwie do prawdziwości *twierdzenia* Gödla, która nie budzi wątpliwości) sprowadza się do kwestii, czy wiemy, że rozważana teoria jest niesprzeczna. Musimy *wiedzieć*, że omawiana maszyna M, czyli teoria S, jest niesprzeczna. Nawet gdy jest, zauważa Putnam, możemy tego nie wiedzieć.

Przejdziemy teraz do innej zasadniczej krytyki argumentu Lucasa: nie da się w ogólności odróżniać Przypadku I od Przypadku II. A następnie podejmiemy drugi najważniejszy wątek: każdy trik w stylu Lucasa może wykonać jakaś maszyna.

H. „Dialektyczny" charakter procedury „wygödlowywania"

Aby wygödlować, musimy być pewni niesprzeczności teorii S. Nawet gdy teoria S jest niesprzeczna, możemy nie mieć dostatecznych podstaw, by to wiedzieć. Jak widzieliśmy na przykładzie NF, ustalenie niesprzeczności teorii może być bardzo trudne. Na tym opiera się zarzut Putnama.

1. Gra ze zwolennikiem mechanicyzmu

W jednej z ostatnich prac Lucas odpiera te obiekcje, tzn. wskazanie konieczności poznania niesprzeczności odpowiedniej teorii w celu poznania prawdziwości zdania Gödla, twierdząc, że zarzuty Putnama są chybione, bo nie biorą pod uwagę „dialektycznej natury argumentu Gödlowskiego" (Lucas [1996], 117). Jest to jego ulubiony chwyt, obecny już, choć w mało dobitny sposób, w pierwotnym artykule [1961], gdzie mówi się, że mamy do czynienia z grą,[93] a podkreślony jako punkt centralny w [1968], który jest odpowiedzią na krytyki, szczególnie na artykuł Benacerrafa [1967]. Otóż chodzi o to, że argument Lucasa nie jest zwykłym argumentem dowodzącym jakiejś tezy, w naszym przypadku tezy, że umysł nie jest maszyną, ale jest to argument dialektyczny, czyli warunkowy: *jeśli* ktoś twierdzi, że jakaś maszyna jest równoważna umysłowi, *to* w odpowiedzi pokazuje się, że popada on w sprzeczność. Argument jest więc „schematem obalania jakiejkolwiek poszczególnej wersji mechanicyzmu. (…) Zależy od tego, czy mechanicysta uczyni pierwszy krok i przedłoży swoją tezę do sprawdzenia" (Lucas [1968], 146). Mówiąc jeszcze inaczej, jest to „argument między dwiema osobami, a nie ciąg dowodowy skonstruowany przez jednego człowieka" (*ibidem*, 154). Zgódźmy się, że można mówić o dialektyczności argumentu w sensie przywołanym przez Lucasa. Nasze punkty (L1) – (L4) są zgodne z taką jego interpretacją. Dlaczego jednak miałoby to przezwyciężać zarzut Putnama?

Otóż mamy grę, w której ktoś, kogo możemy nazwać Mechanicystą, wskazuje jakąś maszynę (por. (L2)), twierdząc, że jest ona równoważna umysłowi (w zakresie arytmetyki), a Lucas mu odpowiada, wskazując zdanie Gödla (por. (L3) i (L4)). W tej grze niesprzeczność proponowanej maszyny powinna być punktem wyjścia (por. (L3)). „Niesprzeczność maszyny jest ustalana nie przez matematyczne umiejętności umysłu, ale na podstawie słowa Mechanicysty" (Lucas [1996], 117). Mamy prawo żądać od niego, aby w grze prezentował

[93] Lucas [1961], 118.

tylko niesprzeczne maszyny M (tzn. maszyny, dla których teoria S jest niesprzeczna). Czy rzeczywiście?

Natychmiastowym zastrzeżeniem wobec takiego wymagania jest zwrócenie uwagi na to, że nie da się mechanicznie rozstrzygać, które maszyny są niesprzeczne. Jest to nie tylko trudne w praktyce, ale jest niemożliwe teoretycznie, o ile chcemy mieć algorytm zawsze dający poprawną odpowiedź, czy dana maszyna jest niesprzeczna, tzn. czy ogół zdań przez nią produkowanych nie pociąga sprzeczności. Ściślej mówiąc, jeżeli efektywnie utworzony ciąg

$M_1, M_2, \ldots, M_n, \ldots$

zawiera wszystkie maszyny Turinga, to nie jest rekurencyjny (a nawet nie jest rekurencyjnie przeliczalny) zbiór C, złożony z numerów maszyn niesprzecznych (tzn. takich, że zbiór zdań arytmetycznych przez nie przedkładanych jest niesprzeczny):

$C = \{n: M_n \text{ jest niesprzeczna}\}$.

Fakt: C nie jest rekurencyjnie przeliczalny.

Dowód tego faktu można oprzeć na twierdzeniu Gödla. Gdyby mianowicie C był r.e., to byłby taki również zbiór numerów zdań Gödla dla niesprzecznych teorii $T(M_n)$ wyznaczonych przez maszyny M_n, czyli zbiór $D=\{G_n: n \in C\}$. Wtedy dla pewnego k byłoby: $D=T(M_k)$. Ponieważ D jest niesprzeczny (składa się ze zdań prawdziwych), czyli $T(M_k)$ jest niesprzeczna, więc $k \in C$. Zdanie G_k jest więc w D, czyli w $T(M_k)$, wbrew twierdzeniu Gödla.[94]

Argument odwołujący się do nierekurencyjności zbioru C został po raz pierwszy użyty przez Wanga w [1974].[95] Tak więc wymaganie, by Mechanicysta proponował tylko niesprzeczne maszyny jest zakładaniem, że ma on możliwości „nadludzkie”[96], a w każdym razie niemechaniczne. Byłoby to więc błędne koło: dowodząc niemechaniczności umysłu, zakładamy niemechaniczność oponenta, czyli niemechaniczność pewnego ludzkiego umysłu!

Lucas próbuje się bronić twierdząc, że chodzi nie o pełną siłę sprawdzania niesprzeczności, a jedynie o możliwość uczynienia tego w pewnych sytuacjach, a mianowicie w tych, gdy ktoś chce poważnie zaproponować jakąś maszynę jako równoważnik umysłu. Takie maszyny muszą mieć odpowiednie rekomendacje, a ich częścią musi być zaświadczenie o niesprzeczności. Nie jest to argument przekonujący z tych samych powodów co poprzednio: zakłada niemechaniczne możliwości u oponenta, który musiałby mieć dostęp do takich instancji rekomendujących maszyny, które są w stanie stwierdzać (prawdziwie!) niesprzeczność. Jesteśmy znów wplątani w błędne koło: jeśli w procedurze wygödlowywania założymy, że my, ludzie, możemy rozstrzygać niesprzeczność, która jest własnością nierekurencyjną, to nic dziwnego, że dochodzimy do konkluzji, że my, ludzie, jesteśmy pod pewnym względem lepsi niż maszyny.

2. Złośliwe pytanie

---

[94] Jak zauważył Bowie w [1982], ten dowód pokazuje, że zbiór C jest „produktywny”, czyli niejako efektywnie nie r.e.; innymi słowy uzupełnienie C jest zbiorem „twórczym” – (por. Rogers [1967]): dla dowolnego jego r.e. podzbioru znajdujemy liczbę, która jest w C, ale poza tym podzbiorem.

[95] Potem wzmocnił go Bowie w [1982], a następnie ja (rozwinięcie tego jest zawarte poniżej w II.K).

[96] Wang [1974], 317.

Dodatkowym trikiem, zaproponowanym przez Lucasa,[97] jest zadanie złośliwego pytania oponentowi, czyli Mechanicyście. Czy mianowicie maszyna, którą proponuje, uznałaby za prawdziwe swoje (czyli odpowiadające systemowi S dla niej) zdanie Gödla? Jeśli powie, że tak, to znaczy, że maszyna jest sprzeczna, więc nie może być równoważna umysłowi, natomiast jeśli nie, to znaczy, że jest niesprzeczna, a zatem umysł może ją wygödlować. Jest to o tyle ciekawe, że odpowiedź pozytywna pociąga sprzeczność w mocnym sensie, bo wiemy, w którym punkcie pojawia się sprzeczność. Ta sprzeczność nie może pozostać ukryta. Mianowicie jeśli T dowodzi G_T, które mówi, że nie ma dowodu w T dla pewnej formuły, to niech liczba m_0 będzie numerem dowodu G_T w T. W normalnej teorii T (a szczególnie takiej, która ma modelować umysł) prawdziwe zdania ograniczone (z kwantyfikatorami ograniczonymi) będą dowodliwe. A zatem $T \vdash Prf(\mathbf{m}_0, \ulcorner G_T \urcorner)$, a to jest jawna sprzeczność z dowodliwością G_T, czyli tym, że $T \vdash \neg(\exists x)Prf(x, \ulcorner G_T \urcorner)$.

Oparty na powyższym triku argument jest jednak wadliwy – i to aż z trzech powodów. Po pierwsze dlatego, że znowu musimy założyć, że Mechanicysta wie, czy maszyna naprawdę dowodzi odpowiedniego zdania Gödla, czyli czy jest niesprzeczna, czy nie. To sprowadza nas z powrotem do poprzedniego zarzutu – do błędnego koła, polegającego na założeniu niemechaniczności oponenta. Po drugie, dokładnie takie samo pytanie można zadać Lucasowi. Czy dowiedzie on zdania Gödla dla siebie, czy stwierdzenia o swojej własnej niesprzeczności? W ten sposób wracamy do problemu omówionego wyżej w II.F.2. Może nie dowiedzie, ale czy z tego ma wynikać coś o nim?

Wreszcie po trzecie – i tu przechodzimy do kwestii fundamentalnej – ten trik może być równie dobrze wykonany przez jakąś maszynę. Przecież zadanie odpowiedniego pytania (Czy w teorii S odpowiadającej maszynie M da się dowieść zdanie G_S?) i zareagowanie jak wyżej (jeśli tak, to S jest sprzeczna, jeśli nie, to G_S jest niedowodliwa lecz prawdziwa) jest zupełnie mechaniczne! Nie wymaga żadnych szczególnych możliwości, może to uczynić jakaś konkretna maszyna. Gdyby ten argument naprawdę działał, to pokazywałby, że owa maszyna jest różna od wszystkich maszyn!

Ta obserwacja, możliwość maszynowego wykonania czynności, które składają się na postępowanie proponowane przez Lucasa, jest jednym z najważniejszych zarzutów, które można formułować pod adresem każdej wersji argumentu Lucasa.

J. Algorytmiczność argumentu Lucasa

Produkowanie zdania Gödla nie wymaga wglądu w naturę teorii, jest natomiast zrealizowaniem pewnego algorytmu. Niewiele przesadzając, można sprowadzić rzecz do następującej karykatury: mogę „dowieść", że jestem lepszy niż ty, bo jakąkolwiek podasz liczbę, ja zawsze podam liczbę większą. Podobnie pokażę, że jestem inny od dowolnej maszyny M. Niech maszyna wskaże pewną liczbę. Ja wtedy wskażę liczbę większą. Jestem więc inny od dowolnie zadanej maszyny M. Dialektycznie okazuję, że nie jestem maszyną. Problem w tym, że jakaś maszyna może odegrać tę samą rolę w grze we wskazywanie liczb. Jasne jest, że ten żart dowodzi, a raczej wyraża, tylko tyle, że dla każdej liczby istnieje większa. Podobnie jest w grze we wskazywanie formuły gödlowskiej. Dlatego Slezak twierdzi, że argument Lucasa korzysta tylko z tego, że Lucas jest innym obiektem niż

[97] Lucas [1996], s. 118. Por. też Lucas [1968].

maszyna M, ale to nie znaczy, że nie jest on maszyną, czy czymś równoważnym maszynie, może nawet rozpatrywanej właśnie maszynie M.[98]

1. Maszyna argumentuje

a) Maszynowe wygödlowywanie

Argument Lucasa może równie dobrze przedłożyć maszyna. Wskazywanie liczby większej od danej jest łatwe (w każdym razie przy założeniu, że liczby przedstawia się w zapisie dziesiętnym), natomiast wskazywanie formuły gödlowskiej jest trudniejsze. Ale nie ma fundamentalnej różnicy, jeśli chodzi o algorytmiczność konstrukcji. Dialektyczna natura argumentu nic nie pomoże. Zauważono to wcześnie, a niewątpliwie było to od razu jasne dla Gödla. Wedle Webba, to mechanizacja diagonalizacji jest „istotą pracy Gödla" (Webb [1980], 151). Nie wiem, kto pierwszy o tym pisał w kontekście argumentu Lucasa. Może Good, który zauważył, że dowodzenie wyższości człowiek nad maszyną metodą wygödlowywania, da się „obalić przez obserwację, że konstrukcja Gödla sama może być przeprowadzona przez inną maszynę (deterministyczną)" (Good [1967], 144) oraz Benacerraf, który sformułował podobną uwagę: być może, da się znaleźć wynikającą z twierdzenia Gödla wadę w każdej maszynie, ale „jest do pomyślenia, że jakaś maszyna może to także uczynić" (Benacerraf [1967], 22).

Najpełniej wykorzystał tę obserwację Webb, w którego książce [1980] jest ona jednym z filarów rozumowania wspierającego idee mechanicyzmu. Sprawa jest ogólniejsza. „Jest to zasadniczy dylemat, wobec którego staje antymechanicyzm: gdy konstrukcje, których używa w swojej argumentacji, stają się dostatecznie efektywne, by można je było traktować jako pewne," to wtedy na podstawie faktu, że procedury efektywne, obliczalne mogą być symulowane przez maszyny Turinga, wnioskujemy, że konstrukcje te może symulować maszyna.[99] Ten zasadniczy fakt zaobserwował już Post ok. roku 1924, a więc przed pracami Gödla. „Zasadą redukowalności dla operacji skończonych" nazwał obserwację, że to, czego możemy stać się „całkowicie świadomi", da się zmechanizować.[100] Powyższa zasada Posta to wczesna wersja tezy Churcha, wspominanej powyżej (w II.D). Z tego powodu nazywa się ją zresztą również tezą Posta-Churcha, albo Posta-Churcha-Turinga.[101] Widać, że przyjęcie tej zasady jest sprawą arbitralną. Jak już zauważaliśmy, stosunek do argumentu Lucasa zależy od przyjętego podstawowego założenia, a mianowicie: czy się da, czy się nie da, maszynowo naśladować rozumowań tworzonych przez umysł.

Nawiasem mówiąc, istnieje argument na rzecz możliwości zmechanizowania umysłu, który opiera się na błędzie logicznym. Mianowicie zakładając, że każda całkowicie określona, ściśle i w pełni opisana czynność umysłowa da się sformalizować, a więc zmechanizować, można by wnioskować – może ktoś powiedzieć – że cały umysł jest mechanizowalny, a więc wysiłek w stylu Lucasa jest z góry skazany na niepowodzenie. Otóż nie! Nawet przyjmując przesłankę, że każdy ścisły fragment myślenia jest mechanizowalny, a czyni tak również

---

[98] Slezak [1982], 48.

[99] Webb [1980], 232. P. też. IV.A.4.c.

[100] „Axiom of Reducibility for Finite Operations", p. tekst Posta w antologii Davisa [1965], 424.

[101] Dobre wprowadzenie daje Epstein-Carnielli [1989]. Nie wchodząc w problem (obszernej) literatury na ten temat, warto wspomnieć tu książkę Hofstatdtera [1979], gdzie są podane najróżniejsze sformułowania, nie tylko zupełnie poważne.

Lucas,[102] nie musimy wnioskować, że *całość* działania umysłu jest mechanizowalna. Nie ma tu przejścia logicznego.[103]

Zastrzeżenie wobec tezy o algorytmiczności procedury użytej przez Lucasa może być wysunięte przez tych, którzy utrzymują, że nasze rozumienie prawdziwości zdania Gödla nie może być udziałem maszyny, bo jak dotąd nie daje się zaprogramować funkcji semantycznych, a może nie uda się to nigdy. Jednak, jak wskazują poprzednie rozważania (w II.G), prawdziwość zdania Gödla nie wynika z jakichś nadzwyczajnych umiejętności wglądu z metapoziomu. Jest to zwykła prawdziwość, chociaż jej dostrzeżenie przez nas wymaga operowania z poziomu metateorii. Poniżej pokażemy, że prawdziwość w ogóle nie musi być przywoływana, by wpędzić Lucasa w kłopoty (p. II.K).

b) Pozaskończona iteracja

Należy wspomnieć inną kwestię związaną z algorytmicznością odwoływania się do zdania Gödla, a mianowicie nieregularności pojawiające się w sposób nieunikniony przy pozaskończonej iteracji tego postępowania. Mechanicysta może próbować dołączać zdanie Gödla do maszyny, a Lucas zastosuje swoją procedurę do maszyny wzmocnionej. Można dodać do maszyny od razu wszystkie kolejne takie zdania Gödla (indeksowane liczbami naturalnymi), a Lucas wtedy doda zdanie Gödla dla maszyny wzmocnionej wszystkimi poprzednimi zdaniami. I tak dalej. Wygödlowywanie można kolejno powtarzać. Sytuację analizował Turing, a potem Feferman.[104] Okazuje się, że choć rozstrzyga się w ten sposób wszystkie zdania Π_1, to jednak zależy to od sposobu przedstawiania liczb porządkowych. Dla Gooda w [1969] jest to dowód, że chodzi nie tyle o Gödla, co o pozaskończone liczenie. Potem podjął ten argument Hofstadter w [1979]. Problem dla procedury Lucasa ma wynikać stąd, że nie ma rekurencyjnego sposobu opisania pozaskończonych konstruktywnych liczb porządkowych (odpowiadających rekurencyjnym dobrym porządkom). Mówi o tym twierdzenie Churcha-Kleene'go. Więc „nie ma algorytmicznej metody, która mówi jak zastosować metodę Gödla do wszystkich możliwych rodzajów systemów formalnych.(…) każdy człowiek po prostu osiągnie w pewnym momencie granice swoich możliwości gödelizowania" (Hofstadter [1979], 476).

Powyższy argument Hofstatdtera jest jednak słaby. Lucas ma rację, że tu dialektyczna natura argumentu załatwia sprawę: cokolwiek zaproponuje Mechanicysta, wymyślając być może szczególne metody opisu liczb porządkowych, można go wygödlować jednym ruchem.[105] Tzn. można, o ile jest niesprzeczny – czego Lucas już nie dodaje. Główny szkopuł – czyli kwestia stwierdzania niesprzeczności maszyny – był już (zaczątkowo) obecny u Wanga w [1974], ale Hofstadter tego nie cytuje. Najwyraźniej nie znał ani tej pracy, ani drugiej najważniejszej z dostępnych wówczas krytyk Lucasa, a mianowicie artykułu Putnama [1960]. Mimo niebywałej swady, z jaką napisana jest porywająca książka Hofdtatdtera,

---

[102] O obaleniu (w swoim mniemaniu) mechanicyzmu mówi: „This is not to say that we cannot build a machine to simulate *any* desired piece of mind-like behaviour" (Lucas [1961], 115).

[103] Oczywiście nie jest tezą logiki formuła $(\forall x)(\exists y)\mathrm{R}(x,y) \rightarrow (\exists y)(\forall x)\mathrm{R}(x,y)$. Na lekcjach logiki podaje się kontrprzykład: dla każdej liczby naturalnej jest liczba większa, ale nie ma liczby większej od wszystkich liczb. Lucas wspomina ten przykład (Lucas [1961], 116), mając najwyraźniej na myśli powyższą formułę rachunku kwantyfikatorów.

[104] Turing [1939], Feferman [1962]. Dobry przegląd zawiera Fefermana [1988]. Por. też IV.A.2.

[105] Lucas [1996], 112. Lucas zarzuca też Hofstadterowi błędne koło, a mianowicie przyjęcie założenia, że umysł nie może rozpoznawać liczb porządkowych w niealgorytmiczny sposób. Bo nie można wykluczyć, że jest to możliwe; wierzył w to np. Gödel. Widać tu znowu, że pierwotny ogólny pogląd na naturę umysłu jest istotny.

brakuje mu trochę do pełnej znajomości tematu. Jego dyskusja argumentu Lucasa jest niekonkluzywna, bo nie porusza najistotniejszej sprawy – problemu sprzeczności.[106] Można do tego dodać, że nie jest przekonujący również jego drugi argument[107] przeciw Lucasowi: należy rozróżniać poziomy; na wyższym może być inteligencja, a na niższym formalne reguły; dlatego na wyższym niekoniecznie od razu A i $\neg$A daje sprzeczność. Jest to podobne do rozważań w II.F powyżej, ale pozostaje faktem, że w argumencie Lucasa chodzi o S czy S_{ar}, czyli o poziom konsekwencji arytmetycznych. Wszystko jedno, czy to jest ten wyższy poziom, czy niższy, pytamy o niesprzeczność S, a to ma zawsze sens.

Powyższa dyskusja nie zmienia faktu, że wszystko, co się czyni przy „wygödlowywaniu", jest czynione według prostego algorytmu, a więc *jest* mechaniczne. Nie zależy to od naszych postaw wobec różnych wariantów Tezy Churcha. Trzeba tylko znać maszynę, tzn. jej kod, albo – co na jedno wychodzi – jej numer w przeliczeniu wszystkich maszyn Turinga. (Normalne, efektywne przeliczenia uzależniają numer danej maszyny w przeliczeniu od jej specyfikacji, czyli opisu całego zestawu jej instrukcji.) Można ten algorytm przedstawić w sposób bardziej techniczny.[108] Funkcja rekurencyjna, która znajduje „pięty achillesowe" funkcji rekurencyjnych, może być bez kłopotu zastosowana do samej siebie, tzn. swojego numeru, by dać swoją „piętę achillesową".

Oczywiście praktyczna wykonalność odpowiedniego algorytmu – to inna sprawa. Przez cały czas idealizujemy umysł, dopuszczając dowolnie długi czas działania, dowolnie dużą pamięć i komplikację. (Zresztą wystarczy czas i pamięć, bo komplikacja może być ustalona dzięki istnieniu uniwersalnych maszyn Turinga.) Niektóre łatwe w zasadzie obliczenia są w praktyce dla nas nieosiągalne. Przykładem, który lubią filozofowie matematyki co najmniej od czasów Brouwera, jest obliczanie rozwinięcia liczby π. Czy da się kiedykolwiek naprawdę obliczyć n-ty element tego rozwinięcia dla wielkiego n, np. exp(9, exp(9, (exp(9, exp(9,9)))))? (Tu exp(a,b) oznacza a^b.)

2. Poważne traktowanie „dialektyczności"

a) Próba odpowiedzi Lucasa

Algorytmiczność argumentu Lucasa bardzo osłabia sens jego stosowania przeciw mechanicyzmowi. Odpowiedź Lucasa, w jednej z nowszych prac, polega na rozróżnieniu pomiędzy dwoma sensami wygödlowywania: w sensie ścisłym, gdy znamy specyfikację maszyny i w sensie luźnym, gdy chodzi o „pewien styl argumentowania, podobny do oryginalnego argumentu gödlowskiego w inspiracji, ale niezupełnie lub niedokładnie określony" (Lucas [1996], 113). Najżyczliwsza interpretacja tego budzącego poważne wątpliwości sformułowania polegałaby chyba na porównaniu do systematycznej wieloznaczności w teorii typów: poszczególne konstrukcje mnogościowe są możliwe tylko dla każdego typu z osobna, ale my widzimy, że jest to zawsze taka sama konstrukcja. Jednak świadczy to raczej o ograniczeniach teorii typów niż o przewadze umysłu. Nie wydaje mi się, by poza odwołaniem się do zdania Gödla, co jest określonym krokiem matematycznym, krył się jakiś nieformalny argument. Zresztą jeśli owo niedokładne określenie argumentu gödlowskiego ma być nieałgorytmiczne, to popadamy w błędne koło, bo zakładamy

---

[106] Nasuwa mi się refleksja, że niedocenianie kwestii niesprzeczności ma związek z tym, że Hofstadter jest fizykiem – jak Penrose.

[107] Prezentowany w rozdz. XVII książki Hofstadtera [1979].

[108] Czyni tak Webb w [1980], 230.

nierekurencyjną moc człowieka, czyli to co mieliśmy dowieść. Jeśli zaś jest to algorytmiczne, to nic nam nie pomoże, jak zobaczymy za chwilę w głównym twierdzeniu z II.K. W istocie przeciwstawienie ścisłego i luźnego sensu procedury wygödlowywania jest odrzucone poniżej w II.K, bo przedstawione tam twierdzenie odnosi się do obu sensów, o ile tylko sens luźny nie zawiera błędnego koła z powodu założenia czyichś nierekurencyjnych możliwości.

Lucas przyznaje, że „pozostaje aura paradoksu" (Lucas [1996], 114). Przekonywujący, ale nieformalizowalny argument? Nie – mówi Lucas – nie chodzi tu o istnienie argumentów „absolutnie nieformalizowalnych." Jednak coś musi pozostać niesformalizowane – choćby użycie reguł wnioskowania.[109] To prawda, ale na to jest prosta odpowiedź. Tak samo jest z maszynami: w komputerach pewne reguły są po prostu zawarte w procesorach! Po drugie – kontynuuje Lucas[110] – niesformalizowane pozostaje pole możliwych zastosowań argumentu. Lucas tego nie rozwija, ale w rozważanej przez nas sytuacji jest to uwaga chybiona. Gdy rozpatrujemy możliwe maszyny Turinga, to właśnie możemy je *wszystkie* przedstawić w postaci rekurencyjnego ciągu. Procedura tworzenia zdania Gödla zależy tylko od numerka w przeliczeniu, czyli od specyfikacji maszyny.

Aby zastosować argument Lucasa, trzeba znać kod maszyny, jej specyfikację. A czy zawsze mając maszynę znamy jej kod? Wydaje się to praktycznie bardzo wątpliwe, nawet gdy przyjąć idealizację. Lucas uznaje, że jest to czepianie się, bo w zasadzie możemy poznać ten kod. Możemy to przyjąć, ale rozumowanie zawarte poniżej w II.K, w którym właśnie zakładamy, że znamy kod prezentowanej maszyny, i tak przekreśli argument Lucasa.

Natomiast zasadnicza różnica pomiędzy wygödlowywaniem a grą w pokazywanie większej liczby leży gdzie indziej. Wskazujemy zdanie Gödla tylko wtedy, gdy teoria jest niesprzeczna. Jest to trochę podobne do gry, w której mamy pokazać liczbę pierwszą większą od zadanej liczby pierwszej. Gdy ktoś zada liczbę złożoną, nie musimy odpowiadać. Jednak podobieństwo tych gier jest niepełne i nie dotyka istoty sprawy. Bycie liczbą pierwszą jest własnością rekurencyjną, a to trywializuje grę. (W teorii! W praktyce często jest zbyt trudno stwierdzić, czy dana liczba jest złożona, czy nie. Szyfrowanie opiera się właśnie na tym fakcie.) Sens zastosowania procedury Lucasa zawsze zależy od tego, czy rozważana maszyna (odpowiadająca jej teoria arytmetyczna) jest niesprzeczna. W tym jest zawarty zasadniczy problem. Jak wspomniane było powyżej, zbiór numerów maszyn niesprzecznych jest nierekurencyjny. Fakt ten jest w całej pełni wykorzystany w następnym podrozdziale (II.K).

b) Zmieniające się maszyny

Poważne potraktowanie dialektyczności argumentu Lucasa prowadzi, jak się wydaje, do maszyn zmieniających się wraz z odpowiedzią. Chihara rozważa wiedzę, czy raczej ogół tego, co się przedkłada jako prawdziwe, zmieniającą się w czasie: S_t w chwili t. Zauważa jednak, że można wziąć sumę wszystkich S_t, czyli to co się da *kiedykolwiek* przedłożyć jako prawdziwe.[111] Wtedy można wrócić do statycznej sytuacji, analizowanej dotychczas. Jednak potem dodaje inny element: mogą być zewnętrzne dane [inputs], które wpływają na pracę maszyny.

Ale i to nie pomoże – stwierdza Chihara. To, co my robimy, by pokazać naszą niemechaniczność, może tez robić maszyna. Jego przykład[112] to maszyna Turinga, która, gdy

---

[109] To dowiódł Żółw Achillesowi, jak zauważył Lewis Caroll. Przedrukował to Hofstadter w [1979].

[110] Lucas [1996], s. 117.

[111] Chihara [1972], 522-3.

[112] Chihara [1972], 525-6.

dać jej na wejściu jej własny kod, zmienia się w inną maszynę. Może więc pochwalić się innym maszynom: nie jestem maszyną, bo jakikolwiek program mi pokażą, okazuję się inną maszyną! Maszyna nie byłaby maszyną. Skąd ta sprzeczność? Rzecz w tym, że pokazanie programu na wejściu musi się łączyć z ustalonym sposobem postępowania; inne możliwe ujęcia tego programu (nie mówiąc o innych jego kodach) przez maszynę Chihary (np. przetrzymywanie w specjalnym miejscu) nie liczą się jako podanie na wejściu kodu maszyny.[113]

Putnam uważa, że aby naśladować ludzi, którzy zmieniają zdania, potrzeba „programu, który mógłby zmieniać zdanie". Są programy takiego typu, twierdzi, „ale twierdzenie Gödla się do nich nie stosuje" (Putnam [1995], 373).

David Lewis inaczej podejmuje wątek zmieniającej się maszyny. Odwołuje się też do faktu, że nawet pojedyncza pomyłka co do niesprzeczności maszyny powoduje sprzeczność Lucasa.[114] Mianowicie wszystko zależy od tego, twierdzi, które zdanie typu *Cons* przyjmiemy jako odpowiedź Lucasa na okazanie maszyny, która ma być mu równoważna.

Lewis rozumuje następująco: obok S(M) i S(L) (ogół zdań, które uznać może Lucas), rozpatrzmy $S^N(L)$, czyli ogół tych zdań, gdy Lucas jest oskarżony o bycie maszyną N (czy raczej równoważnym maszynie N w zakresie arytmetyki). Podobnie czynimy z maszyną, tzn. rozpatrujemy $S^N(M)$. Przyjmujemy (łaskawie dla Lucasa), że $Ar \subseteq S(L)$ i że S(L) składa się ze zdań prawdziwych. Zgodnie z sensem oryginalnego wygödlowywania, $S^M(L) = S(L) + Cons_{S(M)}$.

Przypuśćmy teraz (w celu przeprowadzenia Lucasowskiego dowodu nie wprost), że Lucas jest maszyną M. Innymi słowy, ma tę samą moc arytmetyczną i tak samo reaguje (to ostatnie nie jest wprost powiedziane przez Lewisa). Zatem S(L) = S(M) i oznaczmy ten zbiór przez 'S', oraz $S^M(L) = S^M(M)$ i możemy ten zbiór oznaczyć 'S^M'. Niech teraz $T = Cn(S^M)$, czyli $T = Cn(S + Cons_S)$. Ta teoria jest niesprzeczna, a nawet prawdziwa (założyliśmy prawdziwość S(L), a więc i prawdziwość $Cons_{S(L)}$).

Lewis wydaje się zakładać, że Lucas, co do mocy arytmetycznych, *staje* się $S + Cons_S$.

Lewis twierdzi, że pozór sprzeczności (która kończyłaby dowód nie wprost) wynika z tego, że się nie rozróżnia, jak należy, między $Cons_S$ (wtedy nic dziwnego, że $T \vdash Cons_S$) i $Cons_T$ (wtedy oczywiście $T \text{ non} \vdash Cons_T$, chyba że T jest sprzeczna, a wtedy $Cons_T$ jest fałszywe). Konkluzja Lewisa jest taka, że wszystko jest możliwe: może Lucas nie jest maszyną, ale może jest, a wtedy: albo reaguje na przedłożenie równej mu maszyny M (której – a więc i Lucasa – moc arytmetyczna jest wyrażona przez zbiór S) wskazaniem zdania $Cons_S$, albo wskazaniem zdania $Cons_T$, gdzie $T = Cn(S + Cons_S)$. Wedle Lewisa, zdanie $Cons_S$ jest prawdziwe, $Cons_T$ – fałszywe. Gdy Lucas wskazuje to pierwsze, jest maszyną niesprzeczną, gdy to drugie – sprzeczną.

Mam wrażenie, że rozróżnianie między $Cons_S$ czy $Cons_T$ nie jest szczególnie pouczające. Przecież założyliśmy prawdziwość S! Przyjęcie, że S jest semantycznie adekwatna powoduje, że dowodliwie prawdziwe są i $Cons_S$, i $Cons_T$. Jest tak już przy słabej adekwatności: $S+1\text{-}\underline{Cons}_S \vdash Cons S$, $Cons(S + Cons_S)$, $Cons(S + Cons(S + Cons_S))$, itd. (por. I.C.4.) Jeśli zatem Lucas z założenia jest 1-niesprzeczny, to dowodzi i $Cons_S$, i $Cons_T$.

---

[113] Coś podobnego, ale w luźnym ujęciu, jest też u Dennetta [1972], 530. Mianowicie pisze on, iż trzeba ustalić, co w zachowaniu jest istotne dla obliczania, w szczególności, co ma być interpretowane w znaczeniu „przedkłada jako prawdziwe", a co jest szumem.

[114] Lewis nie pisze tak, ale wynika to z mojego ujęcia jego propozycji.

Na miejscu wydaje się uwaga Pudláka: „argumentując, że nowy [rozszerzony] system jest niesprzeczny, używamy nieświadomie mocniejszych założeń [niż sama niesprzeczność wyjściowego systemu]" (Pudlák [1999], 337)[115]. Warto dodać, że przyjmuje on podejście w duchu matematycznego formalizmu i odrzuca mówienie o prawdzie i poprawności [soundness], a zamiast tego proponuje mówienie o jedynym sposobie, „jaki widzi, by to uściślić", a mianowicie – zasadach refleksji. Jest to posunięcie podobne do tego, co pierwotnie uczynił Gödel, wprowadzając ω-niesprzeczność. Faktycznie, zasada refleksji formalizuje założenie poprawności, czyli stwierdzenie, że „to, co dowodliwe, jest prawdziwe", bo ma postać: $Pr_T(\ulcorner\varphi\urcorner) \rightarrow \varphi$. Zarazem, jak wiemy, Gödel używał pojęcia absolutnej prawdziwości i nie miał wątpliwości, że to, co dowodzimy w sposób absolutny, jest prawdziwe. Pudlák kilka linijek po powtórnym stwierdzeniu, że poprawność teorii trzeba rozumieć jako zasadę refleksji, jakby nieświadom ironii sytuacji, używa intuicyjnego pojęcia prawdziwości: jeśli chodzi o najprostsze zdania, to „bez wątpliwości wierzymy, że są prawdziwe" (Pudlák [1999], 341). Mimo to ma on niewątpliwie rację, zauważając, że matematycy nabierają wiary w słabsze aksjomaty (lub schematy aksjomatów), gdy okazuje się, że mimo wielu prób nie da się wyprowadzić sprzeczności z mocniejszych aksjomatów. Główne przykłady pochodzą z teorii mnogości, w której są coraz mocniejsze aksjomaty nieskończoności. Na to właśnie zwracał uwagę Gödel. Natomiast przykłady z dziedziny arytmetyki, choć też zapoczątkowane przez Gödla, są mniej oczywiste – i Pudlák je uwypuklił.

c) Rozumowanie Benacerrafa

Najpierw zobaczmy, w jaki sposób Benacerraf analizuje uściśloną wersję argumentu Lucasa po to, by pokazać, jak wynika z niego to, że nie da się wykluczyć, że jesteśmy maszyną, ale nie wiemy jaką. Jego wywód jest tu nieco uproszczony.

Załóżmy, że B jest złożone z tych zdań, które umysł (Benacerrafa, czy ogólniej nasz) może dowieść, i dodatkowo z wszystkich zdań, które z tamtych wynikają logiczne. (Chodzi nie tylko o twierdzenia arytmetyczne. W [1967] jest to oznaczone przez 'S*'.) Zakładamy, że dowodzimy tylko zdań prawdziwych, a więc tym bardziej, że jesteśmy niesprzeczni i o tym wiemy, czyli (nieformalnie) możemy to dowieść:

(0) NsprzB $\in$ B. ('Nsprz' oznacza nieformalną niesprzeczność.)

(Oczywiście już to się wydaje bliskie sprzeczności, ale jest nieformalne, więc jeszcze poczekajmy.) Założymy też oczywistą własność B:

(1) Jeśli '$p \wedge q \rightarrow r$' $\in$ B, p $\in$ B, q $\in$ B, to r $\in$ B.

Zakładamy teraz, że jest dany rekurencyjnie przeliczalny zbiór W, który jest postaci W_j, czyli mamy dany *explicite* jego indeks (kod odpowiedniej, generującej go, maszyny – por. I.B.6.c), który spełnia następujące warunki:

(a) '$Q \subseteq W$' $\in$ B, (b) '$W \subseteq B$' $\in$ B, (c) B = W.

Warunek (a) oznaczy tylko to, że odpowiednia porcja arytmetyki jest osiągalna i że to wiemy. Warunek (c) (u Benacerrafa tylko B $\subseteq$ W, ale to jest równoważne z uwagi na (b)) oznacza, że Benacerraf jest maszyną – i ponieważ W = W_j, to wiadomo jaką. Zauważmy, że kod W musi być dany *explicite*, by móc skonstruować $Cons_W$. Warunek (b) jest mocniejszy

[115] Pudlák dowodzi kilku zależności postaci „X(T) $\rightarrow$ X(T+Y(T))", gdzie X to zasada mocniejsza a Y – słabsza. Dla przykładu 1-Cons(T) $\vdash$ 1-Cons(T+*Cons*(T)) i podobnie dla schematu ω-Cons. (Pisze o tym też np. Feferman w [1988], wskazując, że pierwsze tego typu wyniki uzyskał Turing w [1939].)

niż zawieranie; mówi, że jest ono nie tylko faktem, ale że my to *wiemy*. (Warunki (a) i (b) są mocniejsze niż założenia *explicite* przyjęte u Lucasa, które można oddać po prostu jako: $Q \subseteq B$ i $B = W$; wg Benacerrafa to wzmocnienie ma być niezbędne; w każdym razie jest użyte w wywodzie, który następuje.)

Pokażemy, jak z tych trzech warunków wynika sprzeczność. Jeśli pokażemy, że $Cons_W \in W$, to otrzymamy sprzeczność z twierdzeniem Gödla. W celu pokazania $Cons_W \in B$, co na mocy (c) wystarczy, zauważmy, że twierdzenie Gödla stwierdza, iż jeśli $Q \subseteq W$ i NsprzW, to (jedno ze zdań Gödla) $Cons_W$ jest prawdziwe. Można to ująć jako: '$Q \subseteq W \wedge$ NsprzW $\rightarrow Cons_W$' $\in B$. I faktycznie $Cons_W \in B$, bo (1), (a) oraz NsprzW $\in B$, dzięki (0) i (b).

Powyższe rozumowanie ma prowadzić do wniosku, że zakwestionować należy (b) lub (c) (bo raczej nie będziemy podważać (a)). A zatem albo nie jestem maszyną ($\neg$(c)), albo jestem, ale tego nie wiem ($\neg$(b)).[116]

Potwierdza to opinię Gödla, ale ogólny argument anty-lucasowski z II.K sięga głębiej.

Dotychczasowe rozważania pozwalają więc stwierdzić, że twierdzenie Gödla nie wyklucza, iż nasz umysł jest maszyną, ale nie wiemy jaką. Jest to pierwsza z dwu metod atakowania argumentu Lucasa, wymienionych przez Burgessa (p. II.A.2). Dokładnie taką możliwość wspominał Gödel w [1951] (p. II.M), co nie znaczy oczywiście, że tak uważał. Analiza Benacerrafa wydaje się komentarzem do tej uwagi Gödla.[117]

Druga metoda ataku wspomniana przez Burgessa – to możliwość, że jesteśmy maszynami sprzecznymi. O tym wspomina nie tylko Putnam, ale i Benacerraf, a pierwsza wzmianka znajduje się u Gödla w [1951]. Otóż sprzeczny okazuje się przede wszystkim Lucas.

K. Sprzeczność Lucasa

Jakie założenia czyni Lucas, a ogólniej Antymechanicysta, w skrócie p. A (co czytamy „pan A" lub „pani A", w zależności od nastroju), aby wygödlować oponenta, Mechanicystę? Spróbujmy sformułować jak najogólniejsze, tzn. możliwie najsłabsze, warunki, które muszą być spełnione, by stosować jakiś wariant procedury Lucasa. Okaże się, że już to wystarczy, by pognębić każdego, kto ją stosuje, a Lucasa w szczególności.

1. Warunki, które muszą być spełniane przez procedury w stylu Lucasa

---

[116] Jest to sformułowanie tylko nieistotnie inne niż u Benacerrafa w [1967], 29. Należy dodać, że Benacerraf analizuje na końcu pewien paradoks, który potem rozważa Hanson w [1971] i Chihara w [1972]. Nie podejmuję tego, bo wydaje się, że rozbiór argumentu Lucasa można zrobić pomijając te trudności, w które wikłamy się rozpatrując zbiór S = {x: ja mogę dowieść x}. Ponadto dalej idącym stwierdzeniem wydaje mi się nieudowodnialność naszej niesprzeczności – por. II.F.2.a oraz II.L.2.d.

[117] Dla Szumakowicza polemika między Lucasem a Benacerrafem jest sporem pojęciowym: jeden uznaje, że kod danej maszyny musi być nam dostępny, drugi – że niekoniecznie. „Czy bardziej adekwatny jest efektywny mechanicyzm Lucasa, czy nieefektywny mechanicyzm Bencerrafa?" (Szumakowicz [1989], 371). Tymczasem chodzi o to, że „efektywny mechanicyzm" prowadzi (przy pewnych założeniach) do sprzeczności (co najogólniej jest pokazane poniżej w II.K), a Lucas w ogóle nie zauważa możliwości „nieefektywnego mechanicyzmu". (I dalej tego nie widzi, sądząc po późniejszych pracach.) Potem Penrose próbuje to uwzględnić.

Wyobraźmy sobie, że stosujemy najwygodniejszą dla p. A procedurę „dialektyczną” (por. II.H), czyli reagujemy na każdą maszynę, jaką zechce przedłożyć oponent. Jakie to mogą być maszyny? Dowolne, ale zakładamy, żeby ułatwić życie p. A, iż nikt nie będzie wymyślał maszyn, które nie sprowadzają się do maszyn Turinga. (Jest to zgodne z konkluzją dyskusji w II.D.) Pewne wymagania wobec świata są jednak nieuniknione. Przedłożenie maszyny musi oznaczać możliwość znajomości kodu maszyny, a co najmniej numeru równoważnej jej (choćby w zakresie arytmetyki) maszyny Turinga w ustalonym przeliczeniu takich maszyn. Jest to ograniczenie, bo spotkając maszynę w postaci wielkiej szafy, albo grubego tomu zawierającego jej program, nie mamy bezpośrednio dostępu do jej numeru. To jest zarzut „może jesteś maszyną, ale nie wiesz jaką”. Aby nie sparaliżować p. A, zakładamy więc

(W1) Każda wchodząca w grę maszyna jest równoważna maszynie Turinga i jesteśmy w stanie wskazać jedną z takich maszyn Turinga.

Przyjmujemy, że każda maszyna, która wchodzi w grę, „dowodzi” jakichś zdań w języku arytmetyki liczb naturalnych. Natura tego dowodzenia jest nieistotna, nie przesądzamy, czy jest to wynik rozumienia, czy tylko bezmyślnych operacji, czy ma coś wspólnego z prawdziwym dowodzeniem, czy nie. Ma polegać na czymś w rodzaju zapalania zielonego światełka przy drukowaniu niektórych zdań arytmetycznych. (Jest to zgodne z rezultatem dyskusji z II.E.) Nie możemy z góry ograniczyć zakresu maszyn, które wchodzą w grę. Możemy natomiast założyć, że p. A musi reagować na maszyny niesprzeczne, tzn. o niesprzecznym zbiorze „dowiedzionych” zdań. Maszyny sprzeczne, albo takie, które w ogólne nie produkują zdań arytmetycznych, możemy pomijać, albo i nie. Jeśli na nie „reagujemy”, to może to się odbyć w dowolny sposób; nie musimy niczego wykazywać, bo ich sprzeczność je dyskwalifikuje. (Innymi słowy wymagamy reakcji w Przypadku I rozważonym w trakcie dyskusji w II.F; w Przypadku II pozwalamy na pełną dowolność.) Zakładamy więc

(W2) Trzeba reagować na każdą maszynę niesprzeczną (w zakresie arytmetyki).

Na czym ma polegać reakcja p. A na supozycję, że jest równoważnikiem (i to nawet jedynie w zakresie arytmetyki) jakiejś maszyny? Musi to być przedłożenie zdania arytmetycznego, którego ta maszyna nie „dowodzi”. Przez przedłożenie rozumiemy normalnie przedłożenie zdania prawdziwego. Jednak ułatwijmy życie p. A i nie wymagajmy niczego, jeśli chodzi o prawdziwość tego zdania. Fałszywe zdania są też do przyjęcia, trzeba tylko uważać, żeby nie popaść w sprzeczność. Jest to do pomyślenia. Nie zakładamy, że dowolne prawdziwe zdanie jest przez nas dowodliwe, czy jakkolwiek inaczej nam dostępne *jako* prawdziwe.[118] Natomiast twierdzenie Gödla-Rossera mówi, że wiele zdań jest niezależnych, więc z pary zdań sprzecznych można czasem w niesprzeczny sposób wybrać którekolwiek. Jest to więc bardzo zliberalizowane wymaganie, które całkowicie ignoruje komplikacje (rozważane w II.G) związane z ekwiwokacją lub stwierdzaniem prawdziwości zdania Gödla, jak również całą problematykę przechodzenia na poziom metateorii. Dla Lucasa było istotne, że my widzimy prawdziwość G, więc my mu takiego postępowania nie zakazujemy, ale dopuszczamy też procedury w ogóle nie odwołujące się do prawdziwości. Nie wymagamy dowodliwości przedkładanego zdania w jakimkolwiek systemie. Zdanie może być jakiekolwiek, więc przy okazji zupełnie pomijamy całą sprawę, na ile realne jest utworzenie zdania Gödla z kodu maszyny i czy p. A musi być logikiem (por. II.G.5). Zakładamy więc

[118] Tak w gruncie rzeczy czyni Yu w [1992].

(W3) Reakcja na maszynę niesprzeczną polega na wskazaniu zdania, którego ona nie „dowodzi".

Jest jedno zasadnicze ograniczenie, które musimy nałożyć na p. A. Mianowicie wskazywanie zdania nie może być dowolne, ale musi się odbywać zgodnie z pewną efektywną procedurą. Efektywność jest niezbędna. Gdyby z tego zrezygnować i uznać, że p. A może działać niemechanicznie, przyznalibyśmy mu (czy też jej) niemechaniczną moc, a więc to, czego ma dowieść. Byłoby to więc jawne błędne koło. Przypomnijmy, że Lucas stosował w istocie procedurę algorytmiczną (p. II. J). Można ją dowolnie modyfikować, ale nie można pozwolić na odejście od efektywności. (Byłoby takim odejściem np. losowe wybieranie zdania, gdyby to było możliwe. W takiej sytuacji nie sposób byłoby stwierdzić efektywnie, że nie jest ono dowodliwe przez maszynę.) Stosowana procedura musi też być określona całkowicie, a nie zależeć od dodatkowych zewnętrznych okoliczności. Była już o tym mowa wyżej (w II.F): gdybyśmy mogli, na przykład, wymagać od oponenta, Mechanicysty, by przedstawiał tylko maszyny niesprzeczne – jak czynił to momentami Lucas[119] i inni[120] – założylibyśmy, że może on korzystać z niemechanicznych umiejętności. Nic dziwnego, iż dowodzilibyśmy niemechaniczności umysłu (jakiegoś) człowieka. Zatem stawiamy warunek

(W4) Reakcja na maszynę jest efektywnie wyznaczona.

To wymaganie efektywności powinno być przypisane do numeru odpowiedniej maszyny Turinga, bo trudno jest powiedzieć, co mamy do dyspozycji mając maszynę daną empirycznie. Alternatywnie możemy powiedzieć, że najpierw efektywnie wyszukujemy numer maszyny Turinga odpowiadającej zadanej maszynie, a potem dokonujemy ruchu, który w ustalony sposób zależy od tego numeru.

2. Twierdzenie o sprzeczności antymechanicysty

Przełożymy opisane warunki na język logiki matematycznej. Załóżmy więc, że mamy jakąś metodę pokazywania, że umysł ludzki nie jest maszyną. Zakładamy, że każda maszyna może być reprezentowana przez maszynę Turinga. Wszystkie maszyny Turinga są efektywnie ustawione w ciąg:

$M_1, M_2, \ldots, M_n, \ldots$

Nie zakładamy, że mamy bezpośredni dowód, ale raczej, że mamy do czynienia z dialektyczną procedurą F, która zastosowana do maszyny Turinga M_n dowodzi, iż umysł nie jest równoważny M_n. (W ten sposób spełniamy warunek (W1).) Tak więc F jest funkcją, która dla n (indeksu maszyny M_n) przyjmuje wartość F(n), dzięki której okazuje się, że nasz umysł jest różny od M_n.

Wynikiem działania naszej procedury na maszynie M_n jest formuła arytmetyczna F(n), która nie jest dowodliwa przez M_n. Innymi słowy, przyjmując, że „$S(M_n)$" oznacza zestaw twierdzeń arytmetycznych, dowodzonych przez M_n, mamy:

$S(M_n)$ non $\vdash$ F(n).

---

[119] Np. w [1996], por. uwagi w II.H.1.

[120] Takie założenie wykorzystywał Reinhardt w pewnym fragmencie pracy [1986]. Przedstawił też system logiki epistemicznej (z predykatem B oznaczającym dowodliwość), w której niesprzeczne jest zdanie formalizujące sytuację, że umysł jest maszyną. Wniosek, że taka możliwość może faktycznie zachodzić, zależy od tego, jak trafnie aksjomaty sformułowanego systemu oddają własności pojęć logicznych i epistemicznych.

Czy jednak ma sens takie żądanie, gdy teoria $S(M_n)$ jest sprzeczna? Oczywiście nie, bo w takiej teorii wszystko jest dowodliwe. Przyjmujemy więc, że niedowodliwość F(n) ma miejsce, o ile $S(M_n)$ jest niesprzeczna. (Warunek (W3).)

Pomijamy wiele okoliczności, które zachodzą, gdy stosujemy akurat formułę Gödla. Nie zakładamy nic na temat komplikacji F(n), choć wiemy, że użycie twierdzenia Gödla daje formułę klasy Π_1. Nie wymagamy w ogóle rozumienia tej formuły – na jakimkolwiek poziomie. Co więcej, nie zakładamy, że F(n) jest prawdziwa, choć prawdziwość odpowiedniej formuły jest istotna dla oryginalnego podejścia Lucasa. Przyjmujemy tylko, że F(n) nie da się dowieść w $S(M_n)$, o ile ta teoria jest niesprzeczna.

Opuszczając warunek prawdziwości, nie tylko dopuszczamy wiele nowych procedur wygödlowywania, ale rezygnujemy z wymagania, by pokazać, że zdanie Gödla jest dowodliwe w teorii mocniejszej. Dowodzimy tylko *różności* maszyny oraz Lucasa. Tylko to jest istotne. Zdanie F(n) może być fałszywe i w ogóle jakiekolwiek, byle nie było dowodliwe w $S(M_n)$.

Do jakich maszyn nasza procedura musi być stosowalna? Najprościej byłoby założyć, że do wszystkich. Przyjmijmy mniej: co najmniej do wszystkich maszyn niesprzecznych. Nie możemy z góry wykluczyć żadnej maszyny niesprzecznej, bo *a priori* nie wiadomo, która z nich może się okazać udaną symulacją umysłu. Dla maszyn sprzecznych F(n) może być czymkolwiek, np. „0=0”. W tym przypadku nie stawiamy żadnych ograniczeń. Pozostaje jednak w mocy ograniczenie wynikające z faktu, że niesprzeczność nie jest efektywna, czyli zbiór maszyn niesprzecznych jest nierekurencyjny:

$C = \{n: S(M_n)$ jest teorią niesprzeczną$\}$ jest nierekurencyjny.

Nie możemy więc założyć, że F jest określona *tylko* na C. Nie możemy bowiem założyć, że mamy możliwość bezbłędnie stwierdzać, czy n należy do C, czy nie. Oznaczałoby to bowiem, że zakładamy, iż mamy niemechaniczne możliwości, a przecież tego właśnie mamy dowieść. Nie musimy rozstrzygać z góry, jaka jest dziedzina funkcji F. Przyjmijmy więc tylko, że F jest funkcją częściową, określoną dla każdego indeksu maszyny niesprzecznej: $C \subseteq dom(F)$. (To odpowiada warunkowi (W2)).

Najpoważniejsze założenie dotyczy efektywności naszej procedury. Jako się rzekło, jeśli na starcie moglibyśmy mieć niemechaniczne umiejętności, to dowód, że je posiadamy, nie ma żadnej wartości. Aby „wygödlowywanie” w jakiejkolwiek wersji było sensowne, procedura F musi więc być efektywna. Możemy więc żądać, by F była funkcją częściowo rekurencyjną. (To czyni zadość warunkowi (W4)).

Korzystamy tu z tezy Churcha. Jeśli ją odrzucić, nie da się wykluczyć możliwości, że pewne efektywne metody nie dadzą się ująć przez funkcje (częściowo) rekurencyjne.

Ostatecznie otrzymujemy następujące założenia. Załóżmy, że funkcja F jest określona dla niektórych liczb naturalnych (traktowanych jako indeksy efektywnego przeliczenia maszyn Turinga), ma wartości będące formułami (numerami gödlowskimi formuł) języka arytmetyki, przy czym:

(i) F jest częściowo rekurencyjna,

(ii) $C \subseteq dom(F)$,

(iii) dla każdego $n \in C$: $S(M_n)$ non $\vdash F(n)$.

Te (bardzo w sumie słabe) założenia wystarczają do dowodu zaskakującego twierdzenia.

Twierdzenie (***o sprzeczności***): Przy powyższych założeniach zbiór wartości funkcji F jest sprzeczny.

Dowód: Przypuśćmy, że zbiór

$$A = \{F(n): n \in dom(F)\}$$

wartości funkcji F jest niesprzeczny. Ponieważ dzięki (i) jest on rekurencyjnie przeliczalny, więc jest produkowany przez pewną maszynę Turinga. Możemy przyjąć, że dla pewnej liczby k: $A = S(M_k)$. Ponieważ A jest niesprzeczny, więc $k \in C$, a zatem na mocy (ii) F(k) jest określona. Z (iii) wynika, że $S(M_k)$ non $\vdash F(k)$, a więc $F(k) \notin S(M_k)$, czyli $F(k) \notin A$, a to przeczy definicji A. Otrzymana sprzeczność dowodzi, że zbiór A musi być sprzeczny.

Powyższe twierdzenie jest daleko idącym wzmocnieniem obserwacji, że zbiór C jest nierekurencyjny, a zatem nie da się efektywnie rozróżniać między Przypadkiem 1 a Przypadkiem 2 w procedurze Lucasa. To było już uwzględnione przez Wanga w [1974][121]. Zbiór wszystkich formuł Gödla dla teorii $S(M_n)$ był rozpatrywany przez Webba w [1980]. Następnie G. Lee Bowie w [1982] zauważył, że to dowodzi, iż Lucas jest sprzeczny. Uogólnienie tego zostało wzmiankowane przez mnie w [1983], a ogólne warunki (i)-(iii) wspomniane w [1988] i [1993].

Uwagi

a) Powyższy dowód pokazuje, że nawet bardzo wyrafinowane modyfikacje metody „wygödlowywania", również te, które nie korzystałyby z twierdzenia Gödla, ale z innych może dotąd zupełnie nieznanych sposobów ustanawiania niezupełności, wpadają w pułapkę globalnej sprzeczności. Globalnej, bo cały zbiór A jest sprzeczny, choć nie potrafimy orzec, który ze skończonych jego podzbiorów jest sprzeczny. (Wiemy, że taki musi istnieć, bo sprzeczność ma charakter skończony, czyli jest zwarta.)

Można też dodać, że ponieważ sprzeczność globalna wynika z założeń (i), (ii), (iii), to zdania F(n) nie mogą być wszystkie prawdziwe. To, że któreś jest fałszywe, nie jest samo w sobie rozstrzygające. Tylko metoda polegająca na przedkładaniu zdania Gödla załamuje się od razu, gdy podać choćby jeden fałszywy przykład. Gdy bowiem F(n) jest po prostu zdaniem Gödla, uwzględnienie choćby jednego takiego zdania dla $n \notin C$ daje natychmiast konkretną sprzeczność (wspomnianą już w II.H.2). Zdanie Gödla jest bowiem wtedy fałszywe i dowodliwe, czyli istnieje odpowiedni dowód formalny zdania F(n), które nazywajmy teraz 'G_n', w teorii $T(M_n)$. Jeżeli m_0 jest kodem tego dowodu formalnego, to zdanie „liczba m_0 jest dowodem G_n w $T(M_n)$" jest prawdziwym zdaniem o kwantyfikatorach ograniczonych. Jest więc dowodliwe w każdej teorii zawierającej rozsądne minimum arytmetyki. A zatem $T(M_n) \vdash Prf(\mathbf{m}_0, \ulcorner G_n \urcorner)$, a to jest jawna sprzeczność z dowodliwością G_n, czyli tym, że $T(M_n) \vdash \neg(\exists x)Prf(x, \ulcorner G_n \urcorner)$.

b) Założenie (ii) nie wyklucza *a priori*, że zachodzi równość, czyli że funkcja F jest określona tylko na numerach maszyn niesprzecznych. To, że tak być nie może – przy założeniu (i), które pociąga rekurencyjną przeliczalność dziedziny F – wynika z Faktu udowodnionego w II.H.1.

c) Należy wyjaśnić, że w (W1) jest mowa o „jednej z takich maszyn", a nie wymagamy i nie oczekujemy wskazania np. najmniejszej takiej maszyny. Gdybyśmy tego oczekiwali, wpadlibyśmy w subtelną pułapkę. Funkcja $m(n) = \min \{k: S(M_k) = S(M_n)\}$ nie jest rekurencyjna. (Gdyby była, to mielibyśmy $S(M_k)=S(M_n) \Leftrightarrow m(k)=m(n)$, czyli sprawdzanie

[121] Wang [1974], 317.

równości produkcji dwu maszyn byłoby r.e. a tak nie jest, bo to dawałoby rozstrzygnięcie problemu stopu.) Gdybyśmy tego zażądali, *założylibyśmy* niemechaniczną moc p. A, co oczywiście nie jest fair.

d) Można by odrzucić założenie (ii), czyli globalność argumentu. Dialektyczność oznaczałaby jedynie reagowanie w tych kilku przypadkach, w których Mechanicysta naprawdę proponuje jakąś maszynę M. Jeśli ma to być Lucasowskie wskazywanie zdania Gödla, to różnica w porównaniu z dotychczasowymi rozważaniami mogłaby polegać, według Davida Lewisa, na przemianie Lucasa w wyniku tego aktu wygödlowywania. Zmiana polegałaby na pojawieniu się zdania typu *Cons*, którego przedtem nie było. Wedle Lewisa, tak poważne potraktowanie idei dialektyczności procedury Lucasa prowadzi również do jego „upadku" (p. powyżej II.J.2.b) Ideą Lewisa jest zmienianie arytmetycznej „produkcji" Lucasa w sytuacji, gdy jest on „oskarżony" o bycie pewną maszyną. Mianowicie dodaje zdanie Gödla dla tej maszyny. Nie rozpatrujemy więc wszystkich możliwych odpowiedzi wspólnie, jak w Twierdzeniu, ale oddzielnie.

Ogólność założeń Twierdzenia sprawia, że Lucas i każdy, kto próbuje jakiejkolwiek wersji wygödlowywania, popada z konieczności w sprzeczność. Ironią losu okazuje się okoliczność, że jeśli nawet ktoś jest skądinąd niesprzeczny (tzn. jest taki ogół możliwych do uznania przezeń stwierdzeń arytmetycznych), to w momencie, gdy zdecyduje się na jakąkolwiek procedurę w stylu Lucasa, automatycznie popada w sprzeczność. Niezależnie od tego, jak jest ze sprzecznością kobiet i polityków, zostało – jak się wydaje – dowiedzione, że klasa osób sprzecznych zawiera na pewno tych filozofów, którzy wierzą w to, że dzięki twierdzeniu Gödla dowiedli wyższości swego umysłu nad maszynami.

3. Ewolucja maszyn: umysł i roboty

Antymechanicysta nie może więc dowieść, że ma rację. Nie oznacza to oczywiście, że nie ma racji. Jakie możliwości, jeśli chodzi o związek pomiędzy umysłem ludzkim a maszyną, pozostają niewykluczone w świetle powyższych rozważań?

a) Możliwe relacje między umysłem a maszyną

Otóż jeśli umysł nie jest maszyną, jak tradycyjnie wierzyli wszyscy, a obecnie chyba też znakomita większość ludzi, w tym nie tylko Lucas, ale również Gödel, to gdy się mu przedstawi maszynę hipotetycznie mu równą, to albo nie będzie w stanie znaleźć jej kodu (Gödel, Benacerraf, Putnam), albo będzie w stanie i wtedy przedłoży zdanie Gödla. Będzie ono albo prawdziwe, co jest przykładem różnicy między nim a tą maszyną (Lucas), albo fałszywe, co może nastąpić w szczególności wtedy, gdy maszyna jest sprzeczna, ale nie jesteśmy w stanie tego stwierdzić (Putnam).

Jeśli umysł jest maszyną M, to albo 1) jest niesprzeczny, albo 2) jest sprzeczny. W przypadku 2) jesteśmy sprzeczną maszyną (dopuszcza to Putnam) i przedłożenie zdania Gödla tylko potwierdza naszą sprzeczność. W przypadku 1) jesteśmy niesprzeczną maszyną i nie potrafimy znaleźć jej kodu. Dopuszcza to jako hipotezę Gödel (por. cytat w II.A.2 i II.M), a następnie Benacerraf, Putnam i np. Kripke, który stwierdził, iż niemożność wykrycia programu takiej maszyny Turinga nie jest paradoksalna, gdy zważyć, że takie odkrycie zawierałoby rozróżnienie „tego, co mogę naprawdę (absolutnie) dowieść, od tego, co jedynie myślę, że mogę dowieść" (p. Chihara [1972], 524). Jeśli potrafimy znaleźć ten kod, to nie jesteśmy w stanie dowieść prawdziwości zdania Gödla. Musimy dopuszczać jego fałszywość. Twierdzenie Gödla wyklucza to, że jesteśmy maszyną niesprzeczną i potrafimy dowieść zdanie Gödla, arytmetycznie wyrażające tę niesprzeczność.

Czyli – mówiąc jeszcze swobodniej – albo umysł nie jest maszyną i wtedy twierdzenie Gödla go nie ogranicza, albo jest maszyną i jest sprzeczny, a wtedy też nie ma ograniczeń gödlowskich, albo jest maszyną i jest niesprzeczny, a wtedy nie może dowieść zdania Gödla dla tej maszyny, czyli dla siebie samego. To sformułowanie jest bliskie Alternatywy Gödla (p. niżej II.M.2).

b) Ewolucja: powstaje Luke

Jak można sobie wyobrazić maszynę równoważną człowiekowi? Jedna możliwość to produkcja w jakimś laboratorium, które w sposób niewyobrażalny przewyższa istniejące obecnie fabryki robotów. Druga możliwość to ewolucja maszyn. Von Neumannn pokazał, że jest możliwe, by maszyna replikowała siebie samą. Może też wytworzyć maszynę bardziej złożoną. Zaproponował więc, żeby sobie wyobrazić ich ewolucję w wyniku doboru naturalnego.[122] Mogłyby też pojawić się przypadkowe mutacje. Scriven sugeruje, by wyobrazić sobie przedstawicieli cywilizacji robotów z innej planety.[123] Rucker popuszcza wodze fantazji i opisuje cywilizację robotów na księżycu.[124] Sami moglibyśmy ją zapoczątkować! Zachodzi tam darwinowska ewolucja. Możemy sobie wyobrazić, iż w którymś pokoleniu powstaje robot, którego możliwości matematyczne są dokładnie równoważne możliwościom Lucasa. Załóżmy, że miałby na imię Luke. Wtedy mielibyśmy przykład na najdziwniejsze ze wspominanych wcześniej sytuacji.

Po pierwsze znając maszynę, a nawet mogąc, być może, z nią konwersować, nie znalibyśmy jej kodu ani numeru w przeliczeniu maszyn Turinga. Nie budziłoby wątpliwości, że jest to maszyna Turinga, ale odkrycie jej kodu nie byłoby możliwe z powodu nadmiernej komplikacji i braku opisu jej powstania (nawet gdyby praprzodek rodu robotów był dokładnie opisany i miał ustalony numer w przeliczeniu maszyn). Po drugie, stwierdzenie równoważności naszego miłego robota z Lucasem nie byłoby możliwe, nawet gdyby było ściśle prawdziwe. Mógłby to uczynić jakiś hipotetyczny supermózg, zdolny zanalizować ludzkie potencje matematyczne, ale nie byłby w stanie tego okazać w sposób zrozumiały dla Lucasa lub dla naszego robota. Wreszcie po trzecie, nie można by było wykluczyć, że zarówno Lucas, jak i robot są sprzeczni, choć ze wszystkich sił starają się naprawić każdą napotkaną sprzeczność.

Naprawianie pojawiających się sprzeczności musiałoby – jeśli uznać poprawność twierdzenia o sprzeczności z II.K.2 i jego konsekwencji – skłonić Lucasa do porzucenia chęci dowodzenia swojej wyższości. Skłoniłoby nie tylko Lucasa, ale i robota Luke'a. Być może Lucas mimo wszystko chciałby twierdzić, że jeśli Luke jest niesprzeczny, to on wie, że zdanie Gödla dlań, które przecież istnieje, jest prawdziwe. Nie mógłby jednak stwierdzić niesprzeczności Luke'a. Co więcej, sam Luke mógłby też powiedzieć, że jeśli on, Luke, jest niesprzeczny, to jego zdanie Gödla jest prawdziwe. A wreszcie Luke mógłby to samo powiedzieć o Lucasie! Mógłby zapewne próbować „wygödlować" Lucasa. Nie wiem tylko, co by odpowiedział na temat sprzeczności robotów płci żeńskiej i księżycowych polityków…

L. Rozważania Penrose'a

[122] Von Neumann [1966], cz. II, pkt. 1.8 (w wydaniu ros. z 1971 s. 149). P. też Smart [1959], w Anderson [1964], 104.

[123] W tekście z 1953 roku, p. Anderson [1964], 38.

[124] Rucker [1982], 181.

Roger Penrose odnowił argument w stylu Lucasa.[125] Jego ujęcie jest nowym słowem o tyle, że autor – również, jak Lucas, z uniwersytetu w Oxfordzie – jest znanym matematykiem i fizykiem,[126] a ponieważ ma najwyraźniej łatwość pisania, przedstawił problem obszernie, atrakcyjnie, czasem w formie literackiej, wielokrotnie wracając do tematu z różnych stron. Jego autorytet naukowy i jego pióro spowodowały więc wrażenie, że z twierdzeń o niezupełności wyprowadzone zostały nowe wnioski.

Penrose zwraca się zarówno przeciw AI, jak i przeciw tezie, że nie da się pojąć umysłu w kategoriach naukowych, czyli fizycznych. Pisze: „procesy zachodzące w świadomości są nieco inne od tego, co dzieje się w komputerze. Nie wydaje mi się też, że świadomość wykracza poza zakres fizyki, choć sądzę, że nie mieści się w ramach znanych nam praw."[127] Jako fizyk Penrose dodał do argumentu Lucasa nowy ważny element, a mianowicie spekulacje na temat możliwych sposobów osiągnięcia mocy nieobliczalnych przez organizm ludzki. Nie chodzi tu o umysł jako przedmiot nieuchwytny czy duchowy, ale o pewne fizyczne hipotezy dotyczące jego podstaw fizjologicznych. Od tego zaczniemy, by potem wrócić do podstawy całego budowanego przezeń gmachu, jaką jest argument logiczny przeciw mechanicyzmowi. Wedle pracy Grush i Churchland, streszczenie koncepcji Penrose'a można przedstawić następująco.[128] Otóż dokonuje trzech kroków (część logiczna stanowi tylko pierwszy punkt kroku A, ale zajmuje połowę objętości w [1994]):

(A) Dzięki Gödlowi, wiemy, że ludzkie myślenie jest niealgorytmiczne; a zatem musimy być świadomi treści myślenia; nie powinniśmy tego wyjaśniać procesami zawierającymi elementy losowe, więc potrzebne jest nowe rozumienie strony fizycznej myślenia świadomego.

(B) Aby wyjaśnić redukcję funkcji kwantowej, użyteczna może być kwantowa teoria grawitacji; może to objąć procesy niealgorytmiczne (których przejawem wydają się quasi-kryształy).

(C) Mikrotubule, malutkie rurki o średnicy kilku nanometrów, obecne w komórkach, mogą ujawniać zjawiska kwantowe (pierwszy sugerował to Hameroff), co może być podstawą niealgorytmicznych procesów myślenia. Mikrotubule są ważne dla funkcjonowania neuronów, a te dla zjawisk świadomości i poznania, więc w tych zjawiskach pełnią ważną funkcję; w ten sposób, dzięki efektom kwantowym, wprowadzają element niealgorytmiczny do procesów myślowych i poznawczych.

1. Fizyka umysłu

Książki Penrose'a zawierają dobre wprowadzenie do fizyki współczesnej. Jednak jego oryginalne pomysły dotyczące „fizyki umysłu" (punkty (B) i (C) powyżej) są powszechnie krytykowane.

---

[125] We wspomnianych już książkach [1989] i [1994] oraz w kilku artykułach, rozdziałach książek i polemikach. Na szczególną uwagę zasługuje internetowy artykuł [1996] będący bardzo obszerną odpowiedzią na dogłębne krytyki m.in. Davida Chalmersa, Solomona Fefermana, Daryla McCullogha, Drew McDermotta w tym samym piśmie *PSYCHE*. Pierwsza książka ukazała się po polsku jako *Nowy umysł cesarza* (Penrose [1995]). Druga, *Cienie umysłu* (Penrose [2000]), jest dokładniejszym, polemicznym rozwinięciem tych samych wątków.

[126] W 1988 otrzymał, wraz z Hawkingiem, prestiżową Nagrodę Wolfa.

[127] W streszczeniu swoich dociekań w: Brockman [1996], 335.

[128] Grush i Churchland [1995], w: Churchland i Churchland [1998], 208-9. Książki Penrose'a są chwilami przegadane. Wedle Williama Robinsona (stwierdzenie ustne), uwodzi go jego własna retoryka.

Poznanie umysłu jest dla Penrose'a tożsame z poznaniem świadomości. Zakładając jako udowodnioną nieobliczalność, czyli niealgorytmiczność umysłu (innymi słowy: nierównoważność z maszyną Turinga), suponuje, że do wyjaśnienia jego tajemnic potrzeba nowej teorii fizycznej. Przyjmuje, że nieobliczalność w mechanice kwantowej (przy redukcji funkcji falowej) i teorii grawitacji kwantowej (której na razie nie ma!) dostarcza odpowiedniej podstawy dla teorii umysłu. Fizyczną podstawą nieobliczalności świadomości mają być owe mikrotubule, w których mają zachodzić efekty kwantowe, dające skutki na poziomie wyższym. Jest to spekulacja, którą niemiłosiernie skrytykowali biolodzy. Np. słynny noblista Francis Crick: „Penrose mówi o rzeczach, o których nie ma pojęcia."[129] Wedle Grush i P.S. Churchland, którzy rozważają zarówno filar logiczny, jak i fizyczny i neurobiologiczny koncepcji Penrose'a, choć nie da się dowieść fałszywości tych spekulacji, są one „całkowicie nieprzekonywujące i prawdopodobnie fałszywe."[130] Inny biolog, też laureat nagrody Nobla, Gerald Edelman nie jest łaskawszy. Według niego proponowanie użycia teorii grawitacji kwantowej jest wynikiem „karkołomnego przeskoku myślowego". Wyjaśnienie świadomości przez niepoznane elementy fizyczne to wprowadzanie następnego „ducha w maszynie – być może bardziej racjonalnego niż duchy religijne czy okultystyczne, w ostatecznym rozrachunku jednak nie bardziej użytecznego" (Edelman [1998], 296). Zarazem wyraża wdzięczność Penrose'owi za to, że zwrócił uwagę na częstą „pomyłkę kategorialną – czyli na porównywanie umysłu z komputerem" (*ibidem*, 297). Odnosi się to zapewne do dyskusji wokół mocnej AI (o której była wzmianka w II.B). Badacze i konstruktorzy w zakresie sztucznej inteligencji są, oczywiście, nie mniej niechętni. Np. Daniel Hillis uważa, że podejście Penrose'a „przypomina trochę argumentację stosowaną przez witalistów" (Brockman [1996], 346).

Wielki fizyk Murray Gell-Mann ma niezwykle zdecydowaną opinię: „Penrose napisał dwie głupie książki oparte na dawno skompromitowanym, błędnym poglądzie, że twierdzenie Gödla ma coś wspólnego ze świadomością" (Horgan [1999], 266). Niezależnie od naszego stosunku do kwestii świadomości, logiczny argument Penrose'a wymaga omówienia. Na nim spoczywa cała reszta, więc jeśli on jest nieudany, wszystko staje się bardzo wątpliwe, niezależnie od krytyk uderzających bezpośrednio w fizyczne czy biologiczne elementy tej koncepcji.

2. Rozumowanie Penrose'a

Rozumowanie przedstawione przez Penrose'a (tzn. część logiczna całej pracy, której się teraz przyjrzymy) jest w zasadzie wariacją na temat argumentu Lucasa. Trzeba od razu powiedzieć, że Penrose nie jest logikiem i co najmniej do wydania drugiej książki jego wiedza logiczna była niedostateczna i niedokładna. W zawartych tam rozważaniach zrobił błędy matematyczne, mówił bowiem o zdaniu Gödla tak, jakby to było zdanie wyrażające ω-niesprzeczność. Tymczasem jeśli schemat ω-niesprzeczności wyrazić jako jedno zdanie, to nie będzie ono Π_1, ale Π_3, a np. 1-niesprzeczność da się wyrazić jako zdanie Π_2. W odpowiedzi na krytykę Fefermana w [1995] Penrose nie tylko przyznaje się do tego, ale zgadza się, że wprowadzanie „Ω(F)" było całkowicie zbędne. „W gruncie rzeczy prezentacja w *Shadows* byłaby pożytecznie uproszczona, gdyby w ogóle nie wspominać o ω-niesprzeczności" (Penrose [1996], 2.2). Dodaje, że w części nakładu wymieniono „Ω(F)" na

[129] Wypowiedź dla Highfielda w 1994, p. Coveney, Highfield [1997], 399 i tamże przypis 122.

[130] Grush i Churchland [1995]; p. Churchland i Churchland [1998], 208.

zwykłe zdanie Gödla „G(F)".[131] Feferman wymienia więcej błędów z zakresu logiki matematycznej: mylenie pełnej adekwatności [soundness] teorii (w [1994], 90-92) z adekwatnością dla zdań Π_1 ([1994], 74-75), mylenie sytuacji, w których potrzeba założenia niesprzeczności, z tymi, w których potrzebna jest ω-niesprzeczność, nieprawdziwe twierdzenie, że dla dowolnego F niesprzeczność F pociąga niesprzeczność F+$Cons_F$ ([1994], 108) i inne niedokładności.[132] Są też błędy historyczne i w odniesieniach do literatury. Należy więc zapytać, czy okazany brak kompetencji dyskwalifikuje cały argument.

Otóż sądzę, że taki wniosek byłby przedwczesny. Wszystkie te przekłamania można uznać za pomyłki, które da się naprawić, a więc w sumie nieistotne dla zasadniczej linii rozumowania. Tak też broni się autor i uznaje, że nie ma powodu do zmiany stanowiska.

a) Pierwsza książka

W pierwszej książce ([1989], po polsku [1995]), która jest mniej „zawansowana", nie ma wspomnianych błędów logicznych. Główne punkty tej miłej w czytaniu opowieści to: popularyzacja matematyki (liczby zespolone, zbiór Mandelbrota, pokrycia płaszczyzny, rekurencyjna przeliczalność), obrona platonizmu matematycznego i potrzeby intuicyjnego wglądu, no a potem dużo o fizyce, kwantach, kosmologii i trochę o neurofizjologii. Wgląd potrzebny do zobaczenia prawdy uzasadniony jest po pierwsze doświadczeniem autora jako matematyka,[133] co jest niewątpliwe (i może to potwierdzić chyba każdy, kto zna matematykę z autopsji), ale nie dowodzi przecież nieistnienia algorytmu nam równoważnego. Po drugie – i to jest tu najistotniejsze – wgląd tłumaczy się przez odwołanie do twierdzenia Gödla (przedstawione w powiązaniu z twierdzeniem Turinga). Tak więc prawdziwość pewnych twierdzeń jest oparta na „świadomej kontemplacji"; co więcej, „w istocie, algorytmy, same w sobie, *nigdy* nie pozwalają wykryć prawdy" (Penrose [1995], 452, jego podkreślenie). Są to raczej naiwne stwierdzenia, spotykane często przy prezentacji problematyki gödlowskiej. Pierwsze jest obalone w II.G, drugie w II.J. Mówiąc najkrócej, cała procedura wygödlowywania właśnie *jest* algorytmiczna, ale zależy od niesprzeczności odpowiedniej teorii, a tej niesprzeczności możemy nie znać lub nie być pewni i żadna kontemplacja nie pomoże. Chyba, że *założymy* niesprzeczność, albo wgląd niealgorytmiczny, ale wtedy wpadamy w błędne koło w dowodzeniu.

Penrose rozważa hipotezę (która, jak wiemy, pochodzi od Gödla, ale autor nie jest jeszcze tego świadom), że nasze moce matematyczne są równoważne pewnemu algorytmowi, który jednak „jest tak skomplikowany i niejasny, że nigdy nie będziemy wiedzieć, czy jest poprawny." Jego odpowiedź jest rozbrajająca: jest to „sprzeczne z samą ideą matematyki!" Bo matematykę budujemy „z prostych i oczywistych elementów" (Penrose [1995], 458). Tak jakby to naprawdę absolutnie wykluczało istnienie jakiegoś ukrytego algorytmu. Nie chodzi przecież o algorytm, którego nauczamy na studiach matematycznych, ale np. o program hipotetycznego Luke'a z poprzedniego podrozdziału (II.K.3.b). Penrose rozważa zresztą ideę „doboru naturalnego algorytmów" (*ibidem*, 454-456), ale kwestionuje ją na podstawie praktycznego nieprawdopodobieństwa takiej ewolucji, np. tego, że „najmniejsza 'mutacja' algorytmu (…) na ogół powoduje, że staje się on całkowicie bezużyteczny" (*ibidem*, 455). A nam przecież chodzi o logiczną możliwość, a nie o praktyczne prawdopodobieństwo.

---

[131] Niestety, w polskim tłumaczeniu są wszędzie błędne użycia terminu „Ω(F)" i tezy go dotyczące (począwszy od s. 125).

[132] Feferman [1995], cz. 3. Tylko dla 1-niesprzecznych teorii F niesprzeczna musi być i F+$Cons_F$ (por. I.C.4).

[133] Np. Penrose [1995], 453.

b) Druga książka

W drugiej książce, [1994], Penrose podtrzymał w zasadzie wszystkie swoje opinie i odpowiedział obszernie na krytyki jego argumentacji, które pojawiły się po ukazaniu się pierwszej książki. „Uważam, że moje sformułowanie jest bardziej odporne na taką krytykę, z którą spotkał się dowód Lucasa, i pozwala ujawnić błędy krytyków." (Penrose [2000], 74). W artykule [1996] broni się przed następną falą negatywnych ocen, sformułowanych w związku z drugą książką. Jest ostrożniejszy w sformułowaniach. Np. celem książki ma być jasny argument, że „proces świadomego myślenia zawiera element nieobliczalny" (*ibidem*). Pozaświadome mechanizmy są więc pominięte, co faktycznie utrudnia krytykę. Jest tak jednak raczej w teorii, bo, jak zobaczymy, chce i je wykluczyć.[134] W swych rozważaniach Penrose odniósł się do zasadniczych elementów krytyki argumentu Lucasa oraz do tez proponowanych przez Gödla, a w szczególności Alternatywy Gödla (por. niżej II.M.2), wedle której nie da się wykluczyć, że jesteśmy maszyną, ale nie możemy tego stwierdzić, ani nawet ustalić jej niesprzeczności. Ujmując rzecz schematycznie, przy założeniu, że maszyna, lub teoria formalna T, jest nam równoważna, mamy trzy możliwości: I, II i III.[135] Mianowicie:

I: T jest nam znana i wiemy, że jest nam równoważna.

II: T jest nam znana, ale nie wiemy, że jest nam równoważna.

III: T nie jest nam znana.

Penrose uwzględnił więc zarówno możliwość (II), że nie wiemy, iż poznawalna maszyna, lub teoria F, jest nam równoważna, jak i możliwość (III), że taka maszyna (czy teoria) jest, ale jest dla nas niepoznawalna. Można by rzec, że III – to Luke na księżycu, a II – to Luke rozłożony na czynniki pierwsze w ludzkim laboratorium. Obie te możliwości odrzuca (rozdz. 3 w [1994]), twierdząc, iż pozostaje tylko możliwość I, że system jest nam znany i wiemy, że jest nam równoważny. Ponadto rozważa kwestię błędów i ewentualnych sprzeczności i odrzuca możliwość, że taki system mógłby być nieadekwatny semantycznie [unsound], a tym bardziej sprzeczny, więc z twierdzenia Gödla wnioskuje, że nie istnieje system „adekwatny w sposób poznawalny" [knowably sound], a równoważny naszym mocom matematycznym (w zakresie Π_1-zdań). Ten wniosek wydaje się być w porządku. Faktycznie bowiem, znany nam, tzn. przejrzysty dla nas, niesprzeczny system nie może być nam równoważny.

Penrose uważa jednak, że załatwił możliwości I, II, III, czyli wszystkie możliwe sytuacje. Ponieważ, jak sądzi, wykluczył też możliwość sprzeczności (a nawet nieadekwatności) algorytmu T, więc jeśli T jest nam równoważny, to dzięki odrzuceniu III, T musi być poznawalny, a z powodu odrzucenia II musimy wiedzieć, że jest nam równoważny, a zatem, dzięki wygödlowywaniu, dochodzimy do sprzeczności, jak u Lucasa. To ma dowodzić, że nie ma F.

Otóż wydaje mi się jasne, że do tej sytuacji stosuje się zasadnicza krytyka, oparta na Twierdzeniu o sprzeczności z II.K.2. Wydaje się bowiem zupełnie niewątpliwe, że Penrose akceptuje warunki (W1) – (W4) (z II.K.1). Zatem Penrose byłby sprzeczny, a jego odrzucenie

---

[134] Rozróżnienie świadomego i nieświadomego użycia algorytmu, który by miał być nam równy, jest w recenzji W. Robinsona [1992]. Penrose w komentarzu umieszczonym bezpośrednio pod tą recenzją podejmuje to „bardzo pożyteczne" rozróżnienie, które jest ważne w następnej książce. W niej jednak recenzja Robinsona nie jest wspomniana.

[135] Penrose [2000], 171. Penrose wszędzie mówi o „świadomie poznawalnym" algorytmie, ale nie widzę powodu, by tutaj tego nie uprościć do: „poznawalnym". Zakładamy, że chodzi o poznawanie wyidealizowane, które jest świadome.

możliwości, że nam równoważny algorytm T jest sprzeczny (tym bardziej: nieadekwatny) byłoby błędne.

Występuje tu jednak pewna subtelność. Co z niesprzecznymi, ale nieadekwatnymi semantycznie (np. 1-sprzecznymi) teoriami (algorytmami) T? Penrose utrzymuje, że trzeba „wygödlowywać" tylko teorie adekwatne. Założenie (ii) Twierdzenia o sprzeczności jest więc za mocne. Da się jednak je zmodyfikować.

c) Twierdzenie o nieadekwatności

Załóżymy, że procedura w stylu Lucasa stosuje się tylko do teorii (maszyn) adekwatnych semantycznie. Znaczy to, że tylko dla n odpowiadających maszynom adekwatnym funkcja F musi być określona. Oznaczmy: **S** = numery maszyn (semantycznie) adekwatnych (sound). Oczywiście $\mathbf{S} \subseteq C$. Otóż spróbujmy to uwzględnić. Zakładamy:

(i') F jest częściowo rekurencyjna,

(ii') $\mathbf{S} \subseteq \mathrm{dom}(F)$,

(iii') dla każdego $n \in \mathbf{S}$: $S(M_n)$ non $\vdash F(n)$.

Te założenia wystarczają do dowodu modyfikacji Twierdzenia o niesprzeczności.

Twierdzenie (***o nieadekwatności***): Przy powyższych założeniach zbiór wartości funkcji F jest (semantycznie) nieadekwatny [unsound].

Dowód: Przypuśćmy, że zbiór

$$A = \{F(n): n \in \mathrm{dom}(F)\}$$

wartości funkcji F jest adekwatny. Ponieważ dzięki (i') jest on rekurencyjnie przeliczalny, to jest produkowany przez pewną maszynę Turinga. Możemy przyjąć, że dla pewnej liczby k: $A = S(M_k)$. Ponieważ A jest adekwatny, więc $k \in \mathbf{S}$, a zatem na mocy (ii') F(k) jest określona. Z (iii') wynika, że $S(M_k)$ non $\vdash F(k)$, a więc $F(k) \notin S(M_k)$, czyli $F(k) \notin A$, a to przeczy definicji A. Otrzymana sprzeczność dowodzi, że zbiór A musi być nieadekwatny.

A priori zbiór A może być niesprzeczny, choć jako semantycznie nieadekwatny zawiera fałszywe zdanie. To jest wystarczające dla naszych celów, bo pokazuje, że Penrose jest nieadekwatny, tzn. przyjmuje fałszywe twierdzenie arytmetyczne. Jego wiara w metodę pokazywania nieobliczalności umysłów, oparta na pewności co do zasadniczej poprawności metod używanych przez niego i innych matematyków, prowadzi go do sprzeczności z tym przekonaniem, bo założył, że nie zaakceptuje zdania fałszywego. Odpowiedź na pytanie tytułowe rozdz. 3.4 w [1994] „Czy matematycy bezwiednie korzystają z błędnego algorytmu?" brzmi „Czasem tak. Np. Penrose".

A zatem Penrose popada w fałsz i w sprzeczność z przekonaniem o adekwatności swoich metod dowodowych, jeśli tylko stosuje opartą na twierdzeniu Gödla metodę obalania mechanicyzmu. Na tym można by poprzestać, ale pożytecznie jest rozważyć pewne wątki bardziej szczegółowo. Zobaczmy więc, na czym polega odrzucenie możliwości II i III oraz dlaczego Putnam zarzucił Penrose'owi, że przeoczył możliwość IV, która sprawia zasadniczy kłopot. Następnie rozważymy tzw. nowy argument.

d) Przeoczona możliwość

Penrose dowodzi najpierw, że żaden znany nam algorytm, o którym wiemy, że jest adekwatny [sound], nie może symulować całej ludzkiej kompetencji matematycznej (Penrose [2000], 105). Otóż, jak stwierdziliśmy, owszem, tak się rzeczy mają, ale jest to do pogodzenia

z możliwością istnienia jakiegoś naśladującego nas programu, o którym wszakże nie wiedzielibyśmy, że jest nam równoważny (bo byłby zbyt nieprzejrzysty). Pomyślmy o Luke'u.

Następnie Penrose twierdzi, że gdybyśmy stosowali jakąkolwiek regułę niepoprawną, dającą fałszywe twierdzenie, to byłoby to „zasadniczo wątpliwe". (Jest to wspomniany wyżej „rozbrajający" argument, który stosował w pierwszej książce.) Dlatego ogranicza dyskusję do reguł adekwatnych. To eliminuje I. Na dodatek zakłada, że są one dostatecznie proste, byśmy mogli ich działanie „w pełni świadomie zrozumieć [appreciate]" (Penrose [2000], 173). Tu, wedle Putnama, popełnia ten sam błąd, co Lucas. Powód jest następujący. Penrose rozważa możliwość II, czyli dokładne zrozumienie algorytmu T, ale bez pewności, że jest nam równoważny. To ma być „bardzo mało prawdopodobne", bo i tak stwierdzilibyśmy, że T musi być adekwatny [sound].[136] Dodajmy, iż Penrose jest o tyle fair, iż stwierdza, że nie ma „jasnego sposobu wykluczenia możliwości II w ściśle logiczny sposób" (*ibidem*, 175). Penrose rozważa potem możliwość III, czyli program, którego specyfikacja leży poza naszym zasięgiem. Odrzuca ją, bo to by nic nie dało praktycznym przedsięwzięciom AI, a i tak redukuje się do I lub II (*ibidem*, 188.) Jest to bardzo wątpliwy krok, bo chodzi o teoretyczną możliwość istnienia odpowiedniej maszyny, a nie o praktyczne przedsięwzięcia. Co najważniejsze, zauważa Putnam, cały ten wywód pomija możliwość, że może być program, który da się napisać, ale nie da się go zanalizować, „w pełni świadomie zrozumieć". A zatem, można powiedzieć, że mamy możliwość IV. Jest to tak, jakby Luke został zanalizowany przez ludzkie laboratorium i choć cały jego program mielibyśmy wypisany, nie bylibyśmy w stanie stwierdzić, co on robi. Taka sytuacja wydaje się zresztą zupełnie normalna, bo, jak już wspominaliśmy, praktycznie używane wielkie programy zawierają różne błędy [bugs].

Warto wyraźnie napisać, jak pojawia się możliwość IV, bo Penrose chciał wyczerpać wszystkie możliwości i choć nie napisał tego jasno, niektórzy czytelnicy mogą się z tym godzić. Otóż upraszczając wysłowienie, mamy takie możliwości dotyczące algorytmu T: I: T znany i wiemy, że umysł ≡T, II: T znany i nie wiemy, że umysł ≡T, III: T nie jest znany. To faktycznie wyczerpuje wszystkie możliwości! Luka pojawia się, gdy zauważymy, że w II zakłada się, że skoro T jest znany, to jest w pełni zrozumiany. Tymczasem algorytm może być znany i nie być rozumiany! To by znaczyło, iż mamy dany program, o którym nie wiemy, czy jest nam równy, chociaż jest, i który jest tak nieprzejrzysty, że nie potrafimy nic pewnego powiedzieć np. o jego niesprzeczności. To jest właśnie możliwość IV.

Według Putnama,[137] Penrose, który pośrednio przyznaje się[138] do istnienia sytuacji IV, nie ma racji, uważając, że ona podpada pod III, bo w książce jest powiedziane, że w sytuacji III nie znamy programu. Zatem „odrzucenie możliwości, że taki system może symulować produkcję wyidealizowanego matematyka (bo jest w niej coś „co wydaje się cudem" lub jest „zasadniczo wątpliwe") nie jest w ogóle żadnym argumentem" (Putnam [1995], 372). Putnam wyraża opinię, że pomimo ciekawej dyskusji różnych kwestii, smutne jest, że ta książka w ogóle się ukazała.

---

[136] To małe prawdopodobieństwo [plausibility] wynikać ma z tego, że uznalibyśmy aksjomaty i reguły systemu T (który ma być nam równoważny) za niewątpliwie poprawne; stąd zaś wynikałoby, że i twierdzenia też. Robinson [1996] zauważa, że myli się tu dwa poziomy: twierdzeń matematycznych przez nas rozumianych i mechanizmu działania maszyny, która produkuje twierdzenia. Nie mam powodu uważać, że jest niesprzeczność na drugim poziomie, nawet jeśli wierzę w nieobalalną prawdziwość twierdzeń z pierwszego poziomu.

[137] Putnam [1995], 372.

[138] W liście do *New York Timesa* z 15.01.1995, który nawiązuje do recenzji Putnama z *New York Times Book Review* z 20.11.1994, będącej wcześniejszą wersją [1995].

e) „Nowy" argument

Chalmers w [1995] twierdzi, że głęboko zagrzebany w rozważaniach rozdziału 3 jest „nowy" argument i zwięźle go przedstawia. Penrose w [1996] ochoczo podejmuje ten wątek, najwyraźniej mile poruszony tym, że ktoś dostrzega coś, czego w gruncie rzeczy nie dostrzegał sam autor, i wyraża żal, że mało kto to w ogóle zauważył (a w szczególności, że nie uwzględnił tego Putnam). Rzecz w tym, iż nie zakładamy, że możemy poznać adekwatność algorytmu T. Zamiast tego – wnioskujemy to. Jeśli wiem, że umysł ≡T, i wiem, że umysł jest adekwatny (semantycznie, tzn. dowodzi tylko prawdziwych zdań), to stąd wnioskuję, że T jest adekwatny. Czyli według Chalmersa[139] rozumujemy następująco:

(1) wiemy, że umysł ≡T,

(2) wiemy, że umysł jest adekwatny,

(3) więc T jest adekwatny,

(4) więc T' = (T + „umysł ≡T") jest adekwatny,

(5) więc *Cons*(T') jest prawdziwe, ale T' tego nie dowodzi (z GII);

(6) my wiemy, że *Cons*(T') jest prawdziwe;

(7) sprzeczność, bo jeśli my wiemy, umysł ≡T, to T, a tym bardziej T', dowodzi *Cons*(F').

Jak stąd wywnioskować, że nie ma T równoważnej umysłowi? Jeśli wnioskować ¬(1), czyli że nie możemy wiedzieć, że jesteśmy równoważni systemowi T, to to nie wystarczy. Ale zawsze to już coś – pisze Chalmers. Trzeba jednak dodać, że zamiast ¬(1) możemy wnioskować ¬(6), czyli, że nie znamy niesprzeczności. Co więcej, możemy wnosić ¬(2). I tu pojawia się problem (dostrzeżony przez Chalmersa): już *samo* założenie (2), że wiemy (w sposób bezsporny), iż jesteśmy adekwatni, prowadzi do sprzeczności. Jest to bardzo podobne do wywodu z II.F.2.a o niedowodliwości naszej niesprzeczności. Rzecz polega na naśladowaniu wywodu GII z warunków Löba. Chalmers konkluduje, że „może jesteśmy adekwatni, ale nie możemy tego wiedzieć na pewno" (Chalmers [1995], 3.14).

Penrose odpowiada[140], że (1) wystarczy zastąpić założeniem słabszym: umysł ≡T. Penrose twierdzi też, że uniknie się takiej sprzeczności, o jakiej mówi Chalmers, jeśli ograniczyć się do rozpatrywania arytmetycznych Π_1-zdań. Jednak myli się.

Powyższy argument można mianowicie nieco uprościć. Obok poprzednich założeń uwzględniamy nowe:

(1') umysł ≡ T; (To jest owo słabsze założenie, proponowane przez Penrose'a);

Niech A $=_{df}$ r.e. zbiór Π_1-zdań dowodzonych przez T.

(1) wiemy, że umysł ≡T; (Jak poprzednio);

Gdy (1), to wiemy, że A jest zbiorem Π_1-zdań dostępnych umysłowi (tzn. dowodliwych)[141].

(2) wiemy, że umysł jest adekwatny (choćby w zakresie Π_1–zdań);

(2') wiemy, że T jest adekwatny – w tym sensie, że A jest adekwatny;

---

[139] Chalmers [1995], 3.2.

[140] Penrose [1996], 3.4.

[141] Gdy ¬(1) i (1'), to zachodzi równość (zbioru A i zbioru zdań dostępnych umysłowi), ale tego nie wiemy.

(G) wiedza o adekwatności, a więc i niesprzeczności, r.e. zbioru implikuje wiedzę o prawdziwości jego zdania Gödla i jego nieprzynależeniu do tego zbioru.

Zakładamy I: (1), (2) i (G) albo też II: (1'), (2') oraz (G). W obu przypadkach otrzymujemy sprzeczność.

Dowód. Z (1') A jest r.e. Ponieważ, dzięki (2'), A jest adekwatny, więc, na mocy (G), G_A jest określone i nie należy do A, ale widzimy, znając rozumowanie Gödla, że jest prawdziwym Π_1–zdaniem, czyli umysł tego dowodzi. Stąd: G_A należy do A. Jeżeli założymy (1) i (2), to tym bardziej mamy (1') oraz (3), czyli (2').

Można więc odrzucić (1') – zgodnie z pierwotnym zamierzeniem Penrose'a – lub też, wbrew niemu, wnioskować $\neg$(2'), czyli naszą niewiedzę na temat adekwatności T. Nie widać natomiast, jak wyprowadzić sprzeczność z (1') i (2). Cały opis jest zgodny z naszym rozumieniem krytyki przedstawionej przez Putnama. Założenie (1)$\wedge$(2) odpowiada możliwości I. Założenie (1')$\wedge$(2') odpowiada możliwości II. Natomiast (1')$\wedge$(2) odpowiada możliwości IV; nie wynika z tego (2'), które oznacza nasze rozumienie algorytmu T. Natomiast zgodnie z obserwacją Chalmersa wiemy, że samo (2) jest problematyczne, niezależnie od jakichkolwiek założeń na temat T, a nawet niezależnie od istnienia T.

Poglądy Penrose'a na matematykę, jego platonizm, podkreślanie roli intuicji, itp. – to poglądy dość wśród matematyków typowe. Co więcej, można powiedzieć, że Gödel ma wizję matematyki, która jest bardzo podobna do tych poglądów. Analizy Gödla były jednak bez porównania bardziej dociekliwe.

M. Co sądził Gödel?

W 1951 roku, w odczycie im. Gibbsa, Gödel przedstawił pogłębione podejście do konsekwencji filozoficznych swojego twierdzenia, w tym do kwestii mechanicyzmu. Uważał, że przez 20 lat nie nastąpiło dostatecznie poważne przyswojenie jego wyników na poziomie filozoficznym. Jak było już wspomniane (w II.A), poglądy, do których doszedł, bardzo powoli torowały sobie drogę do szerszej publiczności; stało się tak dzięki wysiłkom Wanga, Putnama, Benacerrafa, a ostatecznie – w wyniku publikacji odczytu [1951] w [CW3].

Gödlowi również zależało na argumentacji uprawdopodobniającej tezę, że umysł nie jest maszyną. Stwierdził jednak, że samo jego twierdzenie do tego nie wystarczy. Można uzasadnić tylko słabszą tezę, tzw. alternatywę Gödla (p. poniżej II.M.2). Próbując zrozumieć podejście Gödla, należy stale pamiętać, iż był on przekonany, że jesteśmy niesprzeczni. Co więcej, uznawał, że dowodzimy twierdzeń obiektywnie prawdziwych, przynajmniej czasami. Od tego zaczął swój odczyt.

1. Matematyka obiektywna i subiektywna

W absolutnym sensie są prawdziwe nie tylko stwierdzenia takie jak '2+2=4', ale również niektóre implikacje postaci „z takich a takich aksjomatów dowodliwe jest takie a takie twierdzenie".[142] W sumie te absolutne stwierdzenia tworzą „matematykę właściwą" ([CW3], 305). Nie ma zgody co do tego, co wchodzi w jej skład, ani jakie są jej aksjomaty,

[142] Warto pamiętać, że Russell w [1903] określił czystą matematykę jako ogół stwierdzeń [propositions] postaci „p implikuje q".

które muszą być oczywiste. Mimo to jest pewne, iż ta matematyka właściwa jest niezupełna, a raczej „nieuzupełnialna" [incompletable] czy niewyczerpalna [inexhaustible]. Dobrą ilustracją jest aksjomatyczna teoria mnogości, w której można wyróżniać coraz mocniejsze aksjomaty nieskończoności. Twierdzenie Gödla pokazuje, że ta niezupełność jest całkowicie ogólna i dotyczy każdego dobrze określonego systemu dowodzenia, czyli mówiąc językiem technicznym – systemu rekurencyjnie przeliczalnego. W tym miejscu należy przytoczyć bardzo ciekawą uwagę Boolosa. Czy niewyczerpalność lub nieuzupełnialność jest jednym ogólnym fenomenem, który się przejawia na wspomniane dwa sposoby? I czy w ogóle istnieje trzeci przykład niewyczerpalności, oprócz aksjomatów teorii mnogości i twierdzenia Gödla?[143]

GII wskazuje, że bez popadnięcia w sprzeczność nie możemy być pewni, iż dobrze określony system zawiera prawdziwe [correct] aksjomaty, i uważać, że z nich wynika cała matematyka właściwa. Przekonanie takie pociągałoby bowiem taką samą pewność co do zdania o jej niesprzeczności.

W tym momencie Gödel zadaje kluczowe pytanie: „Czy to oznacza, że żaden dobrze określony system prawdziwych aksjomatów nie może zawierać całej matematyki właściwej?" Odpowiedź pozytywna (że nie może) jest w moim przekonaniu źródłem argumentacji w stylu Lucasa i innych opinii o dowiedzionej przez Gödla otwartości matematyki, a szerzej każdego systemu myślowego, a nawet świata. Intuicja jest taka, że skoro z każdego zbudowanego przez nas czy przez naturę systemu opisu świata, czy ustanawiania reguł, zawsze, o ile jest niesprzeczny, wymyka się jakieś prawdziwe stwierdzenie (a mianowicie fakt owej niesprzeczności czy też jego wyrażenie), to żaden skończony system nie wystarczy; my to rozumiemy, więc możemy ciągle wzbogacać systemy opisu świata. Takie myślenie jest zilustrowane na wiele sposobów w rozdz. IV. Sama teza o niemożności skończonego opisu wydaje się skądinąd sensowna, ale powołanie się na Gödla ma ją uczynić bezsporną, bo dowiedzioną matematycznie.

Dla wielu osób stykających się pobieżnie z twierdzeniem Gödla jest oczywiste, że daje ono jednoznaczną odpowiedź na powyższe pytanie o to, czy żaden skończony system nie zawiera całej matematyki właściwej, absolutnie prawdziwej. Jest tak nawet dla niektórych autorów, którzy poświęcili tej problematyce dużo czasu i wysiłku (Lucas jest najlepszym przykładem). Tymczasem Gödel zauważa, że odpowiedź zależy od tego, co rozumiemy przez matematykę właściwą. Są dwie możliwości: ujęcie obiektywne i subiektywne.

Matematyka właściwa w sensie obiektywnym to ogół wszystkich (obiektywnie) prawdziwych twierdzeń, natomiast matematyka właściwa w sensie subiektywnym to ogół wszystkich twierdzeń dowodliwych, czyli możliwych do udowodnienia przez ludzi jakimikolwiek metodami. Jest to rozróżnienie na – by tak rzec – matematykę w sobie i matematykę dla nas. Może być tak, że dostępna jest nam – nie tylko w danym momencie, ale nawet potencjalnie – tylko część absolutnej matematyki. Oczywiście dla Gödla matematyka subiektywna jest zawarta w obiektywnej, czyli twierdzenia dowodliwe przez ludzi są prawdziwe. Jest to całkowicie tradycyjny pogląd na prawdziwość matematyczną. Ponieważ niektóre koncepcje filozoficzne negują istnienie prawd matematycznych, Gödel zaczął od wspomnianego wyżej przedstawienia prawd absolutnych „matematyki właściwej", nie przesądzając tego, jaki jest jej zakres.

2. Alternatywa Gödla

[143] We wstępie do Gödla [1951], [CW3], 295.

Dzięki twierdzeniu Gödla wiemy, że matematyka w sensie obiektywnym nie może być ujęta przez określony (rekurencyjny) system aksjomatów, czyli nie może być wyprodukowana przez żadną maszynę Turinga. Natomiast matematyka w sensie subiektywnym – może! Nie da się wykluczyć, że wszystko, co może być udowodnione przez człowieka, może być wyprodukowane przez „skończoną regułę", czyli przez maszynę Turinga. Wtedy jednak „my, z naszym ludzkim rozumieniem, nie moglibyśmy nigdy wiedzieć na pewno, że tak jest." Także „nie moglibyśmy nigdy wiedzieć z matematyczną pewnością, że wszystkie produkowane przez nią zdania są prawdziwe [correct]" ([CW3], 309). Innymi słowy, umysł ludzki, co najmniej w zakresie matematyki, byłby „równoważny skończonej maszynie, która nie mogłaby całkowicie zrozumieć swego własnego funkcjonowania" ([CW3], 310). Tutaj „rozumienie" oznacza w szczególności umiejętność widzenia czy stwierdzenia niesprzeczności. Nawet znając pełny opis maszyny, nie moglibyśmy stwierdzić, że ogół produkowanych przez nią zdań jest niesprzeczny. Opis tych hipotetycznych sytuacji, gdy umysł jest równoważny maszynie, i wskazanie, że nie wyklucza tego twierdzenie o niezupełności, został sformułowany z wielką mocą przez Gödla, choć on sam zupełnie w to nie wierzył. Nie miał prekursorów, bo nikt przedtem nie przeprowadził takiej refleksji. W tych kilku pośpiesznie wygłoszonych zdaniach widać źródło wszystkich późniejszych rozważań w tym duchu, prowadzonych przez Putnama, Wanga, Benacerrafa i innych. Później Gödel powiedział Wangowi (było to już cytowane na początku tego rozdziału), że nie da się wykluczyć tego, iż maszyna równoważna naszej intuicji może istnieć „i nawet może być odkryta empirycznie" (Wang [1996], 184). W tej opinii widać źródło wszystkich późniejszych spekulacji o robotach-matematykach.

W [1951] Gödel dodaje, że w opisanej sytuacji „niezdolność do zrozumienia siebie robiłaby mylne wrażenie nieograniczoności czy niewyczerpalności". A to jedynie obiektywna matematyka pozostałaby nieuzupełnialna. Jej niezupełność byłaby wtedy szczególnie mocna: formalne wyrażenie niesprzeczności matematyki subiektywnej, czyli pewne zagadnienie dotyczące istnienia rozwiązań równania diofantycznego, byłoby obiektywnie prawdziwe, ale nie do udowodnienia żadnymi dostępnymi dla człowieka metodami – i w *tym* znaczeniu byłoby *absolutnie* niedowodliwe. To wszystko dzieje się przy założeniu, że matematyka subiektywna jest mechanizowalna. Ale tak być nie musi. Albo więc nie jest, albo jest, a wtedy istnieje problem diofantyczny absolutnie (dla nas) nierozwiązalny. To jest słynna Alternatywa Gödla, „która wydaje się mieć wielkie filozoficzne znaczenie" ([CW3], 310):

> Albo matematyka jest nieuzupełnialna w tym sensie, że jej oczywiste aksjomaty nie mogą nigdy być zawarte w skończonej regule, czyli umysł ludzki (nawet w dziedzinie czystej matematyki) nieskończenie przewyższa moce dowolnej skończonej maszyny, albo istnieją absolutnie nierozwiązalne problemy diofantyczne.

Należy jeszcze raz podkreślić, że „absolutnie" oznacza „przez dowolne metody dowodu dostępne ludzkiemu umysłowi". Prostsze sformułowanie swojej alternatywy Gödel przekazał Wangowi:[144]

> Albo matematyka subiektywna przewyższa możliwości wszystkich [każdego] komputerów, albo też matematyka obiektywna przekracza matematykę subiektywną.

Chociaż powyższa teza jest sformułowana jako alternatywa wykluczająca (zapewne dlatego, że jak wskazałem przed chwilą wywodzi się z takiej alternatywy: albo matematyka subiektywna nie jest mechanizowalna, albo jest, a wtedy…), to jednak Gödel od razu dodaje, że nie jest wykluczone, iż prawdziwe są oba człony tej alternatywy. Oznaczając pierwszy

[144] Wang [1996], 186, jako cytat 6.1.4.

człon alternatywy Gödla (niemechaniczność umysłu) przez ‘(a)’, a drugi (istnienie absolutnie nierozwiązalnego problemu) przez ‘(b)’, mamy więc *a priori* trzy możliwości: $\neg(a)\wedge(b)$, $(a)\wedge(b)$, $(a)\wedge\neg(b)$.

Możliwość $\neg(a)\wedge(b)$, czyli mechaniczność umysłu, a zarazem istnienie arytmetycznych absolutnie nierozwiązalnych problemów – to jest zapewne stanowisko wielu zwolenników sztucznej inteligencji. Prawdziwość obu, $(a)\wedge(b)$, czyli przyjęcie, że umysł jest niemechaniczny, ale są problemy absolutnie nierozwiązalne, to postawa zdrowego rozsądku, a także zapewne wielu filozofów. Pozostawiając na chwilę na boku problem mechanicyzmu, rozważmy wspólną dla obu możliwości tezę (b). Czy w ogóle można powątpiewać w to, że są prawdziwe stwierdzenia arytmetyczne, również dotyczące równań diofantycznych, których prawdziwości w żaden sposób nie da się dowieść? Boolos zauważa, że istnienie zdań, których nie możemy rozstrzygnąć, wydaje się obecnie naturalne również logikom – „właśnie dlatego, że osiągnęliśmy tak dobre zrozumienie twierdzeń o niezupełności”[145], a więc dzięki Gödlowi. Istnienie nierozwiązalnych dla nas problemów jest naturalne z punktu widzenia zwyczajnych doświadczeń matematycznych. Np. Feferman i Solovay uważają, że są proste kwestie matematyczne, które „prawdopodobnie nigdy nie zostaną rozstrzygnięte przez umysł ludzki, bo wykraczają poza choćby w wyobrażeniu osiągalną moc obliczeniową i nie dają żadnego zaczepienia pojęciowego” ([CW2], 292). Jako przykład podają (konkretny) bardzo odległy wyraz rozwinięcia liczby π. Czy w ogóle jakiś matematyk uważa, że wszystkie problemy da się (w zasadzie) rozwiązać? Otóż tak. Na przykład Gödel.

Gödel nie był pierwszy, bo to Hilbert w Królewcu, w 1930 roku,[146] wygłosił słynne *dictum* „Wir müssen wissen, wir werden wissen”, które znalazło się na jego nagrobku. Chodziło mu dokładnie o to, że nie ma problemów nierozwiązalnych [ein unlösbares Problem]. Gödel uważał tak samo. To właśnie jest dlań podstawą do przyjęcia $(a)\wedge\neg(b)$, bo przecież jego alternatywa stwierdza, że $\neg(b)\rightarrow(a)$. Teza antymechanicystyczna, którą tak bardzo chcą mieć Lucas i Penrose, wynika dla Gödla z optymizmu matematycznego – z przekonania, że, używając innego sformułowania Hilberta, w matematyce nie ma miejsca na *ignorabimus*. Tak więc twierdzenie Gödla, według jego autora, ma konsekwencje dla tezy o mechanicznej naturze umysłu. Obala ją, o ile przyjąć dodatkowe założenie: rozwiązalność problemu (nie)istnienia rozwiązań dla równań diofantycznych.

Według Gödla, opisana alternatywa jest ustanowiona ściśle, natomiast raczej bez pretensji do podobnej ścisłości[147] można z obu jej członów wyciągnąć wnioski „zdecydowanie przeciwstawne filozofii materialistycznej”: jeśli zachodzi (a), to nie da się utożsamić umysłu z mózgiem, który wydaje się „skończoną maszyną ze skończoną liczbą neuronów i powiązań między nimi”. Jeśli zaś (b), to obalony jest pogląd, że „matematyka jest naszym własnym tworem” ([CW3], 311). Na uwagę zasługuje sposób, w jaki Gödel uzasadnia to, że z (b) wynika, iż matematyka nie jest naszym tworem, a ma w sobie element obiektywny, od nas niezależny. Gdyby tak nie było, to my, jako twórcy, powinniśmy znać wszystkie własności naszych tworów. Dlaczego? Otóż Gödel wydaje się mieć na uwadze dwa modele twórczości. Po pierwsze – twory skombinowane ze skończonej liczby elementów składowych, takie jak zegary stworzone przez zegarmistrza. To dlatego odrzuca zarzut, że konstruktor nie zna przecież *każdej* własności swojego dzieła. Mianowicie materiały, z

[145] Boolos we wstępie do Gödla [1951], [CW3], 294.

[146] Dnia 5 września, kiedy akurat Gödel na konferencji w tymże mieście wspomniał po raz pierwszy publicznie o swoim odkryciu niezupełności.

[147] Por. Wang [1996], 186.

których konstruuje, nie są jego tworem, więc oczywiście nie zna pewnych ich właściwości, a zatem również pewnych cech całej konstrukcji ([CW3], 312). (Np. zegarmistrz może nie wiedzieć, w jakiej temperaturze kółka zębate się stopią.) Drugi model twórczości to snucie fabuły. Tu twórca ma swobodę – w granicach logiki. Powinien znać własności tego, co stworzył. Tak nie jest, gdy przyjmiemy (b) – stąd konkluzja, że matematyka nie jest „swobodnym wytworem".

Przeciwko Gödlowi można wytoczyć argument na innej płaszczyźnie niż okoliczności rozpatrywane dotychczas. Korzysta on z teorii chaosu, która pokazuje przykłady, jak w zupełnie określonych deterministycznych strukturach bardzo (dowolnie) małe zmiany na wejściu powodują kolosalne różnice na wyjściu. Choć teoretycznie opis jest zupełny, więc odpowiednie równania można potraktować jako nasz twór, nie ma możliwości ustalenia konsekwencji zmian, które są poza zasięgiem możliwości pomiarowych. Czy gdyby minimalna zmiana grubości jednego zęba koła zębatego powodowała wybuch zegarka, zegarmistrz miałby jeszcze szansę na pełną znajomość swego dzieła? Rozwinięta teoria chaosu jest dość nowa i być może Gödel nie zdążył jej poznać, ale jej podwaliny położył Poincaré. O tym Gödel na pewno wiedział. Poza tym podobna nieprzewidywalność pojawia się, gdy rozpatrzyć uniwersalną maszynę Turinga. Nie da się efektywnie opisać jej zachowania w zależności od danych wejściowych, choć samą maszynę, tzn. jej program, możemy opisać całkowicie, i to nawet praktycznie, a nie tylko w zasadzie. Wydaje mi się, że to podważa tezę, że twórca koniecznie musi znać własności swoich tworów.

Powyższy argument nie uderzy jednak we własne przekonania Gödla. Sam Gödel, choć był przeciwko (b), również uważał, że matematyka nie jest naszym własnym tworem.[148] Argumentom na rzecz tej tezy i innym rozważaniom wspierającym matematyczny realizm poświęcona jest pozostała część odczytu [1951] (por. III.C.4.e).

W gruncie rzeczy, niezależnie od swojego twierdzenia i powyższej interpretacji, która używa implikacji $\neg(b)\rightarrow(a)$, Gödel był skądinąd przekonany, iż zachodzi teza (a), czyli że umysł wykracza poza maszyny, a nawet poza materię. Bardzo mu zależało na uzasadnieniu tego poglądu, na pewno nie mniej niż autorom w rodzaju Lucasa czy Penrose'a. Nie chciał tylko wyciągać pochopnych wniosków.

3. Natura umysłu

Umysł jest ujmowany przez Gödla, jak wszystkich w tej dyskusji, w sposób wyidealizowany jako „indywidualny umysł o nieograniczonym długości życia." Jeśli nawet jest maszyną w jakimś bardzo ogólnym sensie, to jest to maszyna, która „rozpoznaje siebie jako mającą słuszność"; poza tym maszyna składa się z części, a świadomość „ma charakter jedności [is connected with one unity]" (Wang [1996], 189). Tezy te są godne uwagi, choć raczej zdroworozsądkowe. Nie różnią się od przekonań wyrażanych przez Lucasa i podobnych mu autorów, jedyną różnicą jest ostrożność w formułowaniu rzekomych dowodów takich tez w oparciu o twierdzenie Gödla. Wyraża się też tu przekonanie o obiektywnej prawdziwości matematyki, a więc tym bardziej jej niesprzeczności. Dodatkowym argumentem jest to, że „jest bezpośrednio oczywiste, iż jestem niesprzeczny, o ile przyjmuje

[148] Gdyby uważał, jak intuicjoniści, że matematyka jest naszym tworem, to mógłby z *tego* powodu odrzucać tezę (b). Uważam, że przytoczony przez Boolosa (w [CW3], 294) cytat z Kanta pasuje mniej do Gödla, a bardziej właśnie do intuicjonistów i konstruktywistów. Kant w *Krytyce czystego rozumu* napisał, że w pewnych dziedzinach (jak właśnie w matematyce) z uwagi na ich naturę powinna być możliwa odpowiedź na każde pytanie, bo „odpowiedź musi wypływać z tych samych źródeł, z których rodzi się pytanie" (A 476/B 504; Kant [1957], t.II, 218).

się ‘dowód absolutny’ jako pojęcie” (Wang [1996], 188). To pojęcie nie jest jak na razie wyklarowane, ale Gödel miał nadzieję, że jest to możliwe.

Wang zauważa, że jeśli użyć terminu „paralelizm psychofizyczny” na tezę, że każde zjawisko mentalne ma swój własny odpowiednik fizyczny, w szczególności w specyficznym stanie mózgu, to zarówno Wittgenstein jak i Gödel mówili o tym paralelizmie jako o „przesądzie”.[149] Według Gödla, kwestia istnienia umysłu niezależnego od materii może być nawet rozstrzygalna empirycznie. Mianowicie może się okazać, że „nie ma dość komórek nerwowych, by wykonać obserwowalne działania umysłu” (Wang [1996], 190).

Choć Gödel uznawał, jak Turing, że mózg działa jak komputer cyfrowy, nie uważał, by niemożliwy był umysł poza materią. Turing zakładał, że zmiana stanu musi mieć charakter fizyczny. Gödel wydaje się tego nie akceptować, bo zarzucił Turingowi, że ten pochopnie założył, iż umysł (w ciele) człowieka może przyjąć tylko skończenie wiele stanów. Oczywiście w żadnej chwili nie może być nieskończenie wielu stanów, o ile każde dwa stany się różnią fizycznie, a możliwości fizycznie różnych stanów mózgu jest skończenie wiele. Na to Gödel się godzi, bo mechanika kwantowa przyjmuje tylko skończoną liczbę stanów i w ogóle fizyka dopuszcza tylko ograniczoną dokładność, a to się nie zmieni (Wang [1996], 196). Jednak Gödel uznał, że Turing popełnił „błąd”, bo „umysł, w swym działaniu, nie jest statyczny, ale stale się rozwija.”[150] Dlatego liczba stanów umysłu może z upływem czasu dążyć do nieskończoności. Najwyraźniej Gödel używa innego pojęcia stanu niż Turing. Najlepiej wyjaśnia to pewna jego wypowiedź zapisana przez Wanga: „Nawet jeżeli skończony mózg nie jest w stanie zawierać nieskończonej ilości informacji, duch może być w stanie. Mózg jest maszyną liczącą połączoną z duchem” (Wang [1996], 193). To ostatnie stwierdzenie brzmi wyjątkowo mało współcześnie. Wydaje się powtórzeniem kartezjańskiej wizji „ducha w maszynie”. W ten sposób, śledząc rozwój mechanicyzmu od Kartezjusza do Gödla, zataczamy krąg.

Według Webba, Gödel przyjmuje, iż coraz lepsze i precyzyjniejsze rozumienie pojęć abstrakcyjnych spowoduje nieograniczony wzrost liczby stanów, choć niekoniecznie stanów rozróżnialnych fizycznie.[151] Jest to nieco podobne do naszej możliwości „myślenia” o nieskończenie wielu liczbach naturalnych, pomimo skończoności mózgu i skończonej liczby stanów, jakie się da w nim wyróżnić. Można też w związku z tym przypomnieć, że uniwersalna maszyna Turinga jest przykładem tego, jak ustalona skończona liczba stanów może wystarczyć do naśladowania dowolnej maszyny Turinga, z dowolną (skończoną) liczbą stanów; cały wzrost komplikacji odbywa się na zewnętrznej taśmie. A człowiek *jest* (gdy dokonamy stosownej idealizacji co do możliwego czasu działania i dostępnej taśmy) uniwersalną maszyną Turinga – twierdzi Turing.[152]

---

[149] Wang [1996], 190. jeśli chodzi o wypowiedzi Gödla, to p. [1951] ([CW3], 309 i 311), Wang [1974], 326; por. też III.C; co do Wittgensteina p. [1999], 139.

[150] Tę obserwację Gödla pierwszy zacytował Wang [1974], 325; p. Gödel [1972] w [CW2], 306; por. Wang [1996], rozdz. 6.3. Pierwszą osobą, która usłyszała o tym od Gödla, był Morgenstern. Zanotował to w swoim dzienniku pod datą 9.12.1969 (p. Dawson [1997], 232).

[151] [CW2], 299.

[152] Por. Webb w [CW2], 301.

Gödel uważał więc, że uzasadniona teza Churcha-Turinga stanowi, iż każda *mechaniczna* procedura efektywna może być symulowana przez maszynę Turinga, a niekoniecznie każda procedura, która *dla nas* jest efektywna.[153]

Gödel wierzył, że możliwe jest rozwinięcie systematycznych metod pogłębiania naszego rozumienia nieskończoności i dowodzenia tak, byśmy byli w stanie zajmować się nimi „w sposób konstruktywny, ale nie mechaniczny" (Webb [1990], 303-4). Innymi słowy, że możliwe jest rozszerzenie pojęcia procedury mechanicznej do „procedury systematycznej", „na tyle ścisłej i zdefiniowanej na tyle precyzyjnie, że moglibyśmy dowieść, iż może dokonać więcej niż jakakolwiek procedura mechaniczna" (Wang [1996], 202). W szczególności dotyczy to hipotetycznej możliwości odkrywania kolejnych aksjomatów nieskończoności w teorii mnogości. By było to naprawdę nierekurencyjne, musielibyśmy mieć taki dostęp do zbiorów, o jakim może marzyć tylko tak konsekwentny platonik jak właśnie Gödel.

4. Podsumowanie

Pozostają trzy możliwości, jeśli chodzi o ograniczenia wprowadzane przez twierdzenie Gödla na maszyny i ludzi.

Po pierwsze, niektórzy uważają, że maszyna podlega ograniczeniom, a umysł nie; należy do nich sam Gödel, no i oczywiście Lucas i Penrose, którzy myślą, wbrew Gödlowi, że da się tego dowieść. Jak pokazaliśmy, popadają wtedy w sprzeczność.

Po drugie, wielu autorów stwierdza, że i maszyny, i umysł podlegają takim samym lub podobnym ograniczeniom. Na przykład Michael Apter: „Z pewnością jest prawdą że zarówno ludzie, jak i maszyny są przedmiotem twierdzenia Gödela w tym zakresie, w jakim funkcjonują jako układy formalne" (Apter [1973], 115). Michael Arbib cytuje *in extenso* odpowiedni fragment tekstu Putnama z [1960] o ewentualnej niemożności dostrzeżenia niesprzeczności, ale wyżej stawia kontrargument Scrivena, wedle którego „twierdzenie Gödla wskazuje na trudność, która nie jest większa w przypadku maszyny niż w przypadku człowieka".[154] Post w [1944] stwierdza, że dostępne człowiekowi metody są rekurencyjne, więc dowód Gödla pokazuje, że stosują się do nas te same ograniczenia co do maszyn. Podobnie Myhill: system nerwowy, a w domyśle umysł, „jest poddany wszystkim ograniczeniom [odnoszącym się do] maszyny Turinga" (Myhill [1950], 195). Dennett rozważa możliwość istnienia najbardziej „obszernej" maszyny Turinga, „zawierającej wszelkie możliwe zmiany programu", powstającej w wyniku hipotetycznego uwzględnienia najdrobniejszych składników fizycznego kształtu urządzenia, które ją wciela i warunków zewnętrznych, w których działa; przy założeniu pełnego determinizmu[155] możliwe jest wygödlowanie, ale nic nie wskazuje, że „ludzie byliby wyłączeni z takich konsekwencji" (Dennett [1972], 529-530). Podobnie piszą nie tylko zwolennicy AI, ale też niektórzy inni autorzy. Np. Wang uznawał coś zbliżonego: „Powiedzenie, że można mieć taki umysł

---

[153] A przynajmniej tak interpretuje to Webb (Webb [1980], 223). Sieg krytykuje Webba za to, że przyjął za Gödlem, iż Turing „popełnia błąd", mówiąc o skończonej liczbie rozróżnialnych stanów umysłu, podczas gdy Turing w [1937] twierdził tylko, że wtedy, gdy – „co jest krystalicznie jasne" (Sieg [1995], 98) – analizuje się procedury *mechaniczne*, trzeba wziąć pod uwagę skończoną liczbę stanów umysłu (p. Davis [1965], 136). Wydaje się, że Wang najpierw był zdania podobnego do opinii Siega, ale potem podzielił zdanie Gödla – p. Wang [1996], 197.

[154] Scriven [1960], p. Arbib [1968], 186.

[155] Tak daleko posunięte założenia wydają się całkiem nieprawdopodobne, i są u Dennetta tylko hipotetyczne, ale spekulacja dokładnie w tym stylu pojawiła się w sierpniu 2002, gdy kończyłem niniejszą książkę, w postaci publikacji Stephena Wolframa o tym, że świat *jest* automatem komórkowym.

[rozwiązujący nierekurencyjny zbiór problemów], a jest logicznie niemożliwe, by mieć taką maszynę, jest nieco niedookreślone." (Wang [1964], 107). Jednak przypomnijmy, że to u Gödla jest po raz pierwszy rozważona możliwość, iż „matematyka subiektywna", czyli pozostająca w zasięgu umysłu, może być równoważna jakiejś maszynie.

Wreszcie, po trzecie, niektórzy suponują, że może jesteśmy sprzecznymi maszynami, a przynajmniej, że jest to logicznie możliwe, więc z tego powodu nie podlegamy ograniczeniom gödlowskim. Wspomina o tym Putnam, a Arbib uznając twierdzenie Gödla za fascynujący, ale pozbawiony znaczenia fakt matematyczny, stwierdza: „Rozumujemy posługując się analogiami. Wciąż uczymy się nowych rzeczy. Popełniamy błędy. Nie jesteśmy spójni, w przeciwieństwie do aksjomatów."[156] Apter odróżnia poziom podstawowy (algorytmiczny) od wyższego, gdzie może być heurystyka itp. Dodaje, że dlatego ludzie i maszyny „mogą w pewnych warunkach przezwyciężyć ograniczenia" wynikające z twierdzenia o niezupełności, „tolerując zdarzające się niekonsekwencje i błędy, które są prawie nieuniknione przy zastosowaniu metod heurystycznych" (Apter [1973], 115). Podobnie piszą Grush i Churchland: „są powody, by wątpić, że ludzkie poznanie czy świadomość muszą wykorzystywać procesy niealgorytmiczne, bo nie w pełni adekwatne choć godne zaufania [unsound, albeit reliable] procesy algorytmiczne unikają gödlowskiej sieci." (Churchland i Churchland [1998], 227). Pierwszy mówił o tym Turing: „Możemy spowodować, by [maszyna] zamiast nieudzielania odpowiedzi, dawała niekiedy odpowiedź błędną. (…) nie mówią one [twierdzenia limitacyjne] jednak nic o tym, ile inteligencji można oczekiwać od maszyny, jeśli nie pretendujemy do jej bezbłędnego działania."[157] Cała ta linia obrony brzmi nieco tajemniczo – jakby pokładało się nadzieję w sprzeczności lub błędach. Gdyby jednak powiedzieć, że sprzeczność jest niemożliwa do wykluczenia, to ta uwaga wpisuje się w drugą metodę kontrataku wspomnianą przez Burgessa (por. II.A.2). Warto jednak podkreślić, że i ta możliwość, owa nieco smętna wizja, że jesteśmy sprzecznymi maszynami, jest jako pierwsza wzmiankowana przez samego Gödla.

Gödel stoi więc u podstaw każdego z tych stanowisk nie tylko z powodu samego twierdzenia, ale i dlatego, że rozważył wszystkie hipotetycznie możliwe warianty sytuacji. Choć mało kto wydaje się tego świadomy, wszystkie dyskusje o sprawach poruszanych w tym rozdziale są w dość ścisłym sensie przypisami do prac Gödla.

Na stosunek do argumentów Lucasa, Penrose'a i im podobnych decydujący wpływ ma ogólna wizja umysłu i maszyn. Zmienia się ona zresztą wraz z rozwojem cywilizacji. Jeśli sądzić po studentach, z którymi się stykam, nasz skomputeryzowany świat jest zapewne źródłem znacznie częstszego wśród młodych przekonania o mechanizowalności wszelkich czynności, nawet umysłu. Skoro podstawowe założenia są ważniejsze niż dowody, co w filozofii jest przecież normalne, to powinienem wnosić, że argument anty-lucasowski, w rodzaju prezentowanego przeze mnie, też nikogo nie nawróci. Szczególnie, że wskazując na sprzeczność lub błędne koło antymechanicystycznego argumentu Lucasa, nie twierdzimy, że da się udowodnić pozytywną tezę o mechaniczności umysłu (lub jej zaprzeczenie).

Zgadzam się w zasadzie z opiniami Penrose'a na temat niezbędności intuicji i wglądu w matematyce i w ogóle w myśleniu, na temat swoistości umysłu, ale Gödel pomaga w tym w ograniczony tylko sposób. Eliminuje naiwną – jak wiemy teraz – wiarę we wszechogarniający przez nas skonstruowany i zrozumiały system czy algorytm, obejmujący całą matematykę. Argument w stylu Lucasa czy Penrose'a nikogo nie nawraca. Ci, którzy i tak wierzą w

---

[156] Słowo 'spójni' jest użyte – jak w wielu polskich tłumaczeniach – zamiast 'niesprzeczni' [consistent] (w wywiadzie z sierpnia 1994, p. Coveney i Highfield [1997], 400, i przypis 125, 502).

[157] Odczyt 20.02.1947 na forum Londyńskiego Towarzystwa Matematyycznego. Wg Hodgesa [2002], 301.

niemechaniczność umysłu, chętnie widzą taki matematyczny dowód swojej słuszności, ale ci którzy i tak wierzą, że maszyna może być równoważna naszemu umysłowi, nie przejmą się. Jeśli docisnąć Lucasa, to pozostaje (chyba) następujące rozumowanie: Gdybym był maszyną, to wiem, że zdanie *Cons* dla mnie byłoby prawdziwe. Skąd? Bo wiem, że jestem niesprzeczny. Skąd? Bo tak czuję. Ale jak to dowieść? Czuję – przecież nie jestem maszyną! Błędne koło jest nie do uniknięcia. Z kolei jeśli ktoś wierzy, że w gruncie rzeczy jesteśmy skomplikowanymi maszynami, to – nawet zakładając niesprzeczność – fakt, że czegoś się nie da dowieść, czyli stwierdzić niesprzeczności, nie jest niepokojący. Przecież nie jesteśmy maszynami wszystkowiedzącymi! Nieznany nam, subtelniejszy algorytm, w rodzaju Luke'a, nie jest logicznie wykluczony. Podobne stanowisko wyraża Feferman, pod którego słowami mógłbym się podpisać: „Choć jestem przekonany o zupełnej niewiarygodności obliczeniowego modelu umysłu, dla mnie osobiście gödlowski argument Penrose'a w żaden sposób tego przekonania nie wspomaga, i podejrzewam, iż tak samo jest z innymi czytelnikami o podobnych przekonaniach. Z drugiej strony jestem pewien, że ci, którzy sympatyzują z obliczeniowym modelem umysłu, znajdą powody, by szybko zignorować argument gödlowski. (…) Jeśli mam rację, to ten [Penrose'a] wysiłek – nie odmawiając mu sumienności – jest w dużej mierze zmarnowany." (Feferman [1995], 1.2).

Mimo ścisłości przysługującej twierdzeniom matematycznym (a może właśnie z powodu niej?) ich zastosowania filozoficzne nie dają niewątpliwych, jednoznacznych wniosków. Rozdział IV zawiera więcej przykładów i rozważań o prawomocności pozamatematycznych wniosków. Rozdział III pokazuje, jaki był światopogląd Gödla i wpływ jego filozofii na dokonane przezeń odkrycia i ich interpretacje.