https://doi.org/10.18778/1689-4286.08.06

HYBRIS nr 8 (2009)
ISSN: 1689-4286

John R. Lucas

Umysły, maszyny i Gödel [1]

Twierdzenie Gödla zdaje się, moim zdaniem, dowodzić, że mechanicyzm jest fałszywy, tj. że nie można pojmować umysłów jako maszyn. Podobne wrażenie odniosło wielu innych ludzi: prawie każdy, kto zajmuje się logiką matematyczną, gdy zapoznałem go z tą sprawą, przyznawał się do podobnych spostrzeżeń, choć odczuwał pewne opory wobec zajęcia ostatecznego stanowiska zanimby cały spór został mu przedstawiony, wszelkie zarzuty wyczerpująco wyłożone i przekonująco odparte[2]. To właśnie postaram się uczynić.

Twierdzenie Gödla oznajmia, iż dla każdego niesprzecznego systemu, wystarczająco mocnego, by zawierać prostą arytmetykę, istnieją formuły, które nie mogą być dowiedzione-w-systemie, lecz których prawdziwość my widzimy. Zasadniczo rozważamy formułę, która – w istocie – brzmi: "Ta formuła jest niedowodliwa-w-systemie". Jeśli formuła owa byłaby dowodliwa-w-systemie, otrzymywalibyśmy

[1] Tekst odczytany przez Johna Randolpha Lucasa na posiedzeniu Oksfordzkiego Towarzystwa Filozoficznego w 1959 roku; pierwotnie opublikowany w „Philosophy", vol. XXXVI, (1961), ss. 112-127; przedrukowany w: K. M. Sayre, F. J. Crosson (red.), *The Modelling of Mind*, Indianapolis 1963, ss. 255-271 oraz A. R. Anderson (red.), *Minds and Machines*, Engelwood Cliffs 1954, ss. 43-59.

[2] Zob. A. M. Turing, *Computing Machinery and Intelligence*, „Mind", 50 (1950), ss. 433-460, przedruk w: J. R. Newman (red.), *The World of Mathematics*, Nowy Jork 1956, ss. 2099-2123. Zob. również K. R. Popper, *Indeterminism in Quantum Physics and Classical Physics*, „British Journal for Philosophy of Science", vol. I, 2 (1950), ss. 117-133. Problem podjęty jest również przez Paula Rosenblooma w: P. C. Rosenbloom, *Elements of Mathematical Logic*, Nowy Jork 1950, ss. 207-208, Ernesta Nagela i Jamesa R. Newmana w: E. Nagel, J. R. Newman, *Gödel's Proof*, Nowy Jork 1958, ss. 100-102, oraz przez Hartleya Rogersa w: H. Rogers, *Theory of Recursive Functions and Effective Computability* (kserogram), vol. I, 1957, ss. 152 i kolejne.

sprzeczność: gdyby bowiem była dowodliwa-w-systemie, wówczas nie byłaby niedowodliwa-w-systemie, a zatem formuła “Ta formuła jest niedowodliwa-w-systemie” byłaby fałszywa; jednocześnie, gdyby rzeczona formuła była dowodliwa-w-systemie, nie byłaby fałszywa, lecz prawdziwa, jako że dla każdego niesprzecznego systemu nic fałszywego nie może być dowiedzione-w-systemie – jedynie prawdy. Zatem formuła “Ta formuła jest niedowodliwa-w-systemie” nie jest dowodliwa-w-systemie, lecz niedowodliwa-w-systemie. Co więcej, jeśli formuła “Ta formuła jest niedowodliwa-w-systemie” jest niedowodliwa-w-systemie, wówczas prawdą jest, że owa formuła jest niedowodliwa-w-systemie, tj. “Ta formuła jest niedowodliwa-w-systemie” jest prawdą.

Powyższy argument jest bardzo zawiły i trudny do zrozumienia, pomocnym jest więc wyłożyć go od drugiej strony – rozważyć ewentualność, w której “Ta formuła jest niedowodliwa-w-systemie” mogłoby być fałszem, pokazać, że jest to niemożliwe i – w ten sposób – że formuła ta jest prawdą, a co za tym idzie, że jest niedowodliwa. Mimo wszystko, argument wciąż pozostaje nieprzekonujący: czujemy, że gdzieś w nim musi tkwić jakiś haczyk. Tym, co z pomocą twierdzenia Gödla można uczynić, jest wykazanie, że nie ma żadnego haczyka i że taki wynik uzyskany być może na drodze najbardziej ścisłej dedukcji. Dotyczy ono [twierdzenie – MZ] wszystkich systemów formalnych, które są:

(i) niesprzeczne,
(ii) wystarczająco mocne, by zawierać prostą arytmetykę, tj. liczby naturalne i działania dodawania i mnożenia

i pokazuje, że są one niezupełne, tj. zawierają niedowodliwe, choć zupełnie sensowne formuły, spośród których pewne, jeśli stoimy poza systemem, widzimy jako prawdziwe.

Twierdzenie Gödla musi stosować się do maszyn cybernetycznych, ponieważ w istocie maszyny tkwi, że jest ona konkretną realizacją systemu formalnego. Z tego wynika, że jeśli daną mamy jakąkolwiek maszynę, która jest niesprzeczna i zdolna generować prostą arytmetykę, istnieje formuła, której [maszyna – MZ] nie jest w stanie przedłożyć

jako prawdziwej, tj. formuła, która jest niedowodliwa-w-systemie, lecz której prawdziwość my widzimy. Z tego zaś wynika, że żadna maszyna nie może być zupełnym lub adekwatnym modelem umysłu; że umysły są istotowo różne od maszyn.

Przez cybernetyczną maszynę rozumiemy urządzenie, które wykonuje pewien zestaw operacji zgodnie z określonym zestawem reguł. Zazwyczaj "programujemy" maszynę, tj. wydajemy jej zestaw instrukcji, mówiących co robić w poszczególnych sytuacjach i wprowadzamy początkowe "informacje", w oparciu o które maszyna ma przeprowadzać swoje obliczenia. Gdy rozważamy możliwość, że umysł jest cybernetycznym mechanizmem, przed oczami mamy właśnie taki model. Zakładamy, że mózg składa się ze skomplikowanych obwodów nerwowych i że informacje, których dostarczają zmysły, są "przetwarzane" i służą za podstawę do działania lub są przechowywane dla późniejszego użytku. Jeśli mózg faktycznie jest takim mechanizmem, to wówczas, przy danym sposobie jego zaprogramowania – sposobie, w jaki został "podłączony" – a także [przy danej – MZ] informacji, której mu dostarczono, jego reakcja – to, co uzyskamy "na wyjściu" – jest z góry ustalona i, przy odpowiedniej ilości czasu, możliwa do obliczenia. Nasze wyobrażenie maszyny zakłada po prostu, że jej zachowanie jest całkowicie zdeterminowane przez sposób, w jaki jest zbudowana, i przez napływające "bodźce" – nie ma możliwości, by zrobiła coś na własną rękę – przy danej konstrukcji i pewnym wkładzie informacji musi zadziałać w pewien określony sposób. My jednak nie powinniśmy rozważać tego, co maszyna musi zrobić, lecz to, co zrobić może. Tzn. zamiast rozważać cały zestaw reguł, które razem określą dokładnie, co maszyna zrobi w danych okolicznościach, winniśmy zająć się ogólnie tymi regułami, które ustalą możliwe reakcje maszyny, lecz nie całkowicie. Ścisłe reguły ściśle wyznaczają jej operacje na każdym etapie. Na każdym etapie dane będzie określone polecenie, np. "Jeśli liczba jest pierwsza i większa niż dwa, dodaj jeden i podziel przez dwa; jeśli nie jest pierwsza, podziel przez najmniejszy dzielnik"; my jednak rozważymy sytuację, w której dopuszczalne są instrukcje w postaci alternatywy, np. "W ułamku zwykłym możesz podzielić to, co pod i nad kreską, przez dowolną liczbę stanowiącą dzielnik zarówno licznika, jak i mianownika". W ten oto sposób, łagodząc

wymogi dotyczące naszego modelu, tak że nie jest już całkowicie deterministyczny, lecz nadal w zupełności mechanicystyczny, powinniśmy być w stanie skierować naszą uwagę na atrybut postulowany często dla mechanicznych modeli umysłu; to mianowicie, że powinny one zawierać urządzenie losujące. Można by zbudować maszynę, dla której wybór pomiędzy pewną liczbą alternatywnych rozwiązań byłby ustalany na podstawie, powiedzmy, liczby atomów radu, które rozpadły się wewnątrz danego pojemnika w przeciągu ostatnich trzydziestu sekund. Wydaje się, *prima facie*, że nasze mózgi podlegałyby efektom losowym – promieniowanie kosmiczne równie dobrze mogłoby lub nie- wystarczyć do tego, by wywołać impuls nerwowy. Jest jednak oczywiste, że urządzenia losującego nie można by zainstalować w maszynie tak, by wybierało dowolną możliwość; dopuszczalny byłby jedynie wybór pomiędzy pewną liczbą dozwolonych możliwości. Dozwolonym jest dodać dowolną, losowo wybraną liczbę do obu stron równania, jednak nie dodać pewną liczbę do jednej strony równania i różną od niej do drugiej strony równania. Dozwolonym jest zdecydować się dowodzić raczej to twierdzenie Euklidesa niż inne lub użyć raczej tej metody niż innej, lecz nie "dowieść" czegoś, co nie jest prawdą lub użyć "metody dowodzenia", która nie jest właściwa. Wszelkie urządzenia losujące mogą umożliwiać wybór między tymi tylko operacjami, które nie będą prowadzić do sprzeczności, co zresztą jest wprost określone przez złagodzone wymogi naszego modelu. W istocie możemy tę sprawę przedstawić następująco: zamiast rozważać, w jaki sposób całkowicie zdeterminowana maszyna musi zadziałać, powinniśmy rozważyć, jak owa maszyna będzie mogła zadziałać, jeśli posiadać będzie urządzenie losujące, które uruchamia się, ilekroć możliwe będą co najmniej dwie operacje, spośród których żadna nie prowadzi do sprzeczności.

Gdyby taka maszyna została zbudowana, by generować twierdzenia z zakresu arytmetyki (pod wieloma względami najprostszego działu matematyki), składałaby się ona ze skończonej ilości części i z tego względu skończona byłaby liczba typów operacji dla niej możliwych i skończona liczba początkowych założeń, na których owe operacje mogłaby wykonywać. W istocie możemy iść dalej i stwierdzić, że liczba typów operacji i początkowych założeń, które można by do niej wprowadzić, byłaby ściśle określona.

Maszyny są ściśle określone – cokolwiek, co jest nieokreślone lub nieskończone, nie powinno być uznane za maszynę. Należy zwrócić uwagę, że mówimy o liczbie typów operacji, nie zaś liczbie operacji. Przy wystarczającej ilości czasu i założeniu, że maszyna nie zużyłaby się, mogłaby ona powtarzać operację w nieskończoność; idzie więc jedynie o to, że możliwa jest jedynie określona liczba różnych typów operacji, które może wykonywać.

Jeśli mamy ściśle określoną liczbę typów operacji i początkowych założeń, wprowadzonych do systemu, możemy wszystkie je przedstawić na kartce papieru za pomocą odpowiednich symboli. Operacje przyrównać możemy do reguł ("reguł inferencji" lub "reguł zastępowania"), które pozwolą nam przejść od jednej lub większej liczby formuł (lub nawet od żadnej formuły) do innej formuły, zaś początkowe założenia (o ile takie istnieją) przyrównać możemy do zbioru formuł początkowych ("sądów pierwotnych", "postulatów" lub "aksjomatów"). Gdy już wypiszemy je wszystkie, możemy odtworzyć każdą pojedynczą operację – wszystko, co musimy zrobić, to odnaleźć formuły reprezentujące stany [maszyny – MZ] przed i po oraz regułę, która została zastosowana. Tak oto możemy odtworzyć na kartce papieru każdy możliwy ciąg operacji, który maszyna może wykonać. Jakkolwiek długo maszyna by działała, my, mając wystarczającą ilość czasu, papieru i cierpliwości, bylibyśmy w stanie przedstawić na kartce odpowiednik jej operacji. Ów odpowiednik stanowiłby w istocie formalny dowód – każda operacja maszyny przedstawiona jest przez zastosowanie jednej z reguł, natomiast warunki, które określają, czy w danej sytuacji maszyna może wykonać pewne operacje stają się w naszym przedstawieniu warunkami, które określają, czy reguła może zostać zastosowana do danej formuły, tj. formalnymi warunkami stosowalności. Tak oto, interpretując nasze reguły jako reguły inferencji, powinniśmy otrzymać dowód – ciąg formuł, spośród których każda zapisana została jako rezultat zastosowania pewnej formalnej reguły do zdania lub zdań ją poprzedzających (za wyjątkiem, oczywiście, formuł początkowych, które dane są jako reprezentujące początkowe założenia wprowadzone do systemu). Wnioski, które maszyna może przedłożyć jako prawdziwe, stanowić więc będą odpowiednik

twierdzeń, które można udowodnić w odpowiadającym jej formalnym systemie. Teraz skonstruujemy formułę gödlowską dla tego systemu. Nie może ona zostać dowiedziona-w-systemie. Z tego względu maszyna nie jest w stanie przedłożyć odpowiadającej jej formuły jako prawdziwej. My jednak widzimy, że owa formuła gödlowska jest prawdziwa; każda rozumna istota może prześledzić argument Gödla i przekonać się, że ta konkretna formuła gödlowska, choć niedowodliwa-w-systemie, jest mimo to – a nawet z tego właśnie powodu – prawdziwa. Każdy mechaniczny model umysłu musi zawierać mechanizm, który może oznajmiać prawdy arytmetyki, bo to jest coś, co umysły umieją czynić; co więcej, nietrudno jest zbudować mechaniczny model, który pod wieloma względami oznajmia prawdy arytmetyki dużo lepiej niż istoty ludzkie. Lecz pod tym jednym względem nie będzie lepszy; dla każdej bowiem maszyny istnieje zdanie prawdziwe, którego ona nie może przedłożyć jako prawdziwego, zaś umysł może. To pokazuje, że maszyna nie może być zupełnym i adekwatnym modelem umysłu. Nie jest w stanie zrobić wszystkiego, co jest w stanie zrobić umysł, bo choć jest w stanie zrobić wiele, zawsze pozostanie coś, czego zrobić nie jest w stanie, zaś umysł jest. To nie znaczy, że nie możemy zbudować maszyny, która symulowałaby dowolny wskazany obszar działania umysłu; znaczy to jedynie tyle, że nie jesteśmy w stanie zbudować maszyny, która symulowałaby każdy obszar działania umysłu. Jesteśmy w stanie (lub będziemy pewnego dnia) zbudować maszyny zdolne symulować okruchy działania umysłu i – w istocie – przewyższać ludzki umysł wydajnością. Jednakże jakkolwiek dobra byłaby maszyna i jakkolwiek dalece przewyższałaby ludzki umysł we wszystkich niemal aspektach, zawsze mieć będzie tę jedną słabość, tę jedną rzecz, której nie będzie w stanie zrobić, a umysł ludzki będzie. Formuła gödlowska jest piętą achillesową cybernetycznej maszyny. Z tego też względu nie możemy mieć nadziei na to, że kiedykolwiek wyprodukujemy maszynę zdolną do wszystkiego, do czego zdolny jest umysł. Nigdy nie zbudujemy – jest to z zasady niemożliwe – mechanicznego modelu umysłu.

Powyższy wniosek wyda się niektórym ludziom wysoce podejrzany. W pierwszej kolejności sprzeciwią się oni temu, że maszyna może symulować dowolny wskazany

obszar działania umysłu, a zarazem nie może symulować każdego obszaru jego działania. Dla niektórych stanowi to sprzeczność. W odpowiedzi wystarczy wskazać, że nie zachodzi sprzeczność między tym, że dla każdej liczby naturalnej można podać liczbę większą i tym, że nie można podać liczby większej niż każda liczba. Możemy użyć tej samej analogii również przeciwko tym, którzy po znalezieniu formuły, nie mogącej być przedłożoną jako prawdziwa przez ich pierwszą maszynę, przyznają, że – istotnie – jest ona nieadekwatna [jako model umysłu – MZ], lecz zaraz potem usiłują zbudować drugą, bardziej adekwatną maszynę, zdolną już przedłożyć tę formułę jako prawdziwą. To – owszem – mogą uczynić, lecz wówczas dla tej drugiej maszyny istnieć będzie jej własna formuła gödlowska, skonstruowana przez zastosowanie procedury Gödla do formalnego systemu, reprezentującego jej (owej drugiej maszyny) własny rozszerzony schemat operacji. I tej formuły druga maszyna nie będzie w stanie przedłożyć jako prawdziwej, podczas gdy umysł dostrzeże jej prawdziwość. I nawet jeśli zbuduje się trzecią maszynę, zdolną do tego, do czego niezdolna była druga maszyna, nastąpi dokładnie to samo: istnieć będzie trzecia formuła, formuła gödlowska dla systemu formalnego, reprezentującego schemat operacji trzeciej maszyny, której nie będzie ona w stanie przedłożyć jako prawdziwej, podczas gdy umysł nadal widzieć będzie jego prawdziwość. I tak w nieskończoność. Jakkolwiek skomplikowaną maszynę zbudujemy, będzie jej – skoro jest maszyną – odpowiadał pewien formalny system, który poddać można procedurze Gödla w celu znalezienia formuły niedowodliwej-w-tym-systemie. Owej formuły maszyna nie będzie w stanie przedłożyć jako prawdziwej, choć umysł dostrzeże jej prawdziwość. I tak oto maszyna wciąż nie będzie adekwatnym modelem umysłu. Usiłujemy zbudować model umysłu, który byłby mechaniczny – czyli zasadniczo “martwy” – lecz umysł, będąc w istocie “żywy”, za każdym razem może iść o krok dalej niż każdy formalny, skostniały, martwy system. Dzięki twierdzeniu Gödla to do umysłu zawsze należy ostatnie słowo.

Przedstawmy teraz drugi zarzut. Procedura, za pomocą której konstruuje się formułę gödlowską, jest standardowa – tylko pod takim warunkiem możemy być pewni, że dla każdego formalnego systemu można skonstruować formułę gödlowską. Jeśli jednak

jest to standardowa procedura, powinno być możliwym takie zaprogramowanie maszyny, by i ją [procedurę – MZ] wypełniła. Moglibyśmy zatem zbudować maszynę ze standardowym zestawem operacji i – dodatkowo – z operacją polegającą na wypełnieniu procedury gödlowskiej i przedłożeniu wniosku owej procedury jako zdania prawdziwego, potem zaś ponownym wypełnianiu tej procedury tak długo, jak to konieczne. Taka maszyna odpowiadałaby systemowi z dodatkową regułą inferencji, umożliwiającą dodanie do systemu, jako twierdzenia, zdania gödlowskiego skontruowanego dla tego systemu, później formuły gödlowskiej skonstruowanej dla nowego, wzmocnionego [o poprzednią formułę gödlowską – MZ] formalnego systemu itd. Byłoby to równoznaczne z dodawaniem do pierwotnego formalnego systemu nieskończonego ciągu aksjomatów – każdej formuły gödlowskiej otrzymanej do danego momentu. Nawet w tym przypadku sprawa nie jest rozstrzygnięta. Maszyna z "operatorem gödlującym", jak moglibyśmy go nazwać, różni się od maszyn takiego operatora pozbawionych. I choć maszyna z operatorem byłaby w stanie wykonywać te operacje, dzięki którym umysł przewyższa maszyny bez operatora, to jednak teraz, jak można się spodziewać, umysł, skonfrontowany z maszyną posiadającą "operator gödlujący", weźmie ten fakt pod uwagę i wygödluje tę nową maszynę, sam "operator gödlujący" i wszystko. Tak się w istocie dzieje. Nawet jeśli dołączymy do systemu formalnego nieskończony zbiór aksjomatów, składający się z kolejnych formuł gödlowskich, powstały w ten sposób system pozostanie niezupełny i będzie istnieć dla niego formuła niedowodliwa-w-systemie, mimo że istota rozumna, stojąc poza systemem, widzi jej [formuły – MZ] prawdziwość[3]. Spodziewaliśmy się, że nawet jeśli dodany by został nieskończony zbiór aksjomatów, musiałby on zostać ustalony na mocy pewnej skończonej reguły lub warunku i ta właśnie reguła lub warunek mogłaby zostać wzięta pod uwagę przez umysł, który rozważałby powiększony w ten sposób system formalny. W pewnym sensie ponieważ to do umysłu zawsze należy ostatnie słowo, może on zawsze znaleźć dziurę w każdym systemie formalnym, który przedstawi się mu jako model jego własnych działań. Mechaniczny model musi być, w

[3] Zastosowano tu oryginalny dowód Gödla. K. Gödel, *Lectures at the Institute of Advanced Study*, Princeton 1934, § 1 oraz § 6.

jakimś sensie, skończony i ściśle określony i dlatego umysł zawsze może iść o krok dalej.

To jest odpowiedź na zarzut wysunięty przez Turinga[4]. Utrzymuje on, że wskazywanie na ograniczenia możliwości maszyn nie na wiele się zda. Mimo że każda konkretna maszyna jest niezdolna udzielić poprawnej odpowiedzi na pewne pytania, to w końcu każda konkretna istota ludzka również jest omylna; i w każdym przypadku "naszą wyższość odczuwać możemy w danej sytuacji wobec tej maszyny, nad którą odnieśliśmy nasze błahe zwycięstwo. Nie ma mowy o jednoczesnym zwycięstwie nad wszystkimi maszynami." Nie o to tu jednak chodzi. Nie dyskutujemy o tym, czy maszyny, czy umysły są lepsze, ale o tym, czy są takie same. Pod pewnymi względami maszyny są niezaprzeczalnie lepsze niż ludzkie umysły, zaś problem, z którym nie potrafią sobie poradzić jest – przyznajmy – raczej drobny, a nawet trywialny. Jest jednak wystarczający – wystarczający, by pokazać, że maszyna nie jest taka sama jak umysł. To prawda – maszyna potrafi wiele rzeczy, których ludzki umysł nie potrafi; jeśli jednak jest coś, czego maszyna z konieczności nie potrafi, zaś umysł ludzki potrafi, wówczas, jakkolwiek trywialna nie byłaby to sprawa, nie możemy postawić między nimi znaku równości i nie możemy mieć nadziei na zbudowanie mechanicznego modelu, który byłby adekwatną reprezentacją umysłu. Oznacza to zatem, że zwycięstwa naszego nie odnieśliśmy nad konkretną tylko maszyną, lecz nad każdą, którą ktokolwiek zechce wskazać – po łacinie *quivis* lub *quilibet*, a nie *quidam* – a jakikolwiek mechaniczny model umysłu musi być konkretną maszyną. Mimo że prawdą jest, iż każde poszczególne "zwycięstwo" umysłu nad maszyną może zostać "przebite" przez inną maszynę, zdolną udzielić odpowiedzi, do której udzielenia pierwsza maszyna była niezdolna, a zatem "nie ma mowy o jednoczesnym zwycięstwie nad wszystkimi maszynami", to w tym momencie staje się to nieistotne. Nie chodzi o nierówny pojedynek pomiędzy umysłem a wszystkimi maszynami, lecz o to, czy możliwa jest jakakolwiek konkretna maszyna, która potrafi wszystko to, co umysł. By teza mechanicyzmu miała sens, musi być z zasady możliwe zbudowanie modelu,

[4] A. Turing, dz. cyt., ss. 444-445, J. R. Newman, dz. cyt., s. 2110.

konkretnego modelu, który potrafiłaby wszystko to, co umysł. Przypomina to grę[5]. Mechanicysta ma pierwszy ruch. Buduje on – jakikolwiek, ale ściśle określony – mechaniczny model umysłu. Ja wskazuję coś, czego ów model nie potrafi, ale co potrafi umysł. Mechanicysta może wówczas dowolnie ulepszyć swój wynalazek, lecz za każdym razem, gdy to uczyni, ja jestem uprawniony do tego, by szukać usterek w tak poprawionym modelu. Jeśli mechanicysta jest w stanie zaprojektować model, w którym nie znajdę usterki, wówczas jego teza jest dowiedziona, jeśli nie jest w stanie – wówczas teza pozostaje bez dowodu. A ponieważ, jak się okazuje, z konieczności nie jest w stanie tego zrobić – jego teza zostaje obalona. By zwyciężyć, musi być w stanie zbudować pewien ściśle przez niego określony model umysłu – całkowicie dowolnego, lecz takiego, który będzie umiał wskazać i którego będzie się trzymał. Ponieważ jednak nie może, z zasady nie może, zbudować żadnego mechanicznego modelu, który byłby adekwatny, nawet jeśli stopień niepowodzenia jest niewielki, mechanicysta musi ponieść porażkę, a mechanicyzm – okazać się fałszem.

Wciąż jednak wysunąć można poważniejsze zarzuty. Twierdzenie Gödla stosuje się do systemów dedukcyjnych, a istoty ludzkie nie ograniczają się do przeprowadzania jedynie dedukcyjnych inferencji. Twierdzenie Gödla stosuje się jedynie do systemów niesprzecznych, a przecież można mieć wątpliwości co do tego, na ile zasadne jest założenie, że istoty ludzkie są niesprzeczne. Twierdzenie Gödla stosuje się jedynie do systemów formalnych, a przecież na ludzką pomysłowość nie jest a priori nałożone jakiekolwiek ograniczenie, które wyklucza możliwość, że uda się nam skonstruować jakąś replikę człowieczeństwa, niereprezentowalną przez żaden system formalny.

Istoty ludzkie nie ograniczają się do przeprowadzania dedukcyjnych inferencji i dlatego C. G. Hempel[6] oraz Hartley Rogers[7] podkreślali, że porządny model umysłu musiałby

[5] Dla podobnego typu argumentacji zob. J. R. Lucas, *The Lesbian Rule*, „Philosophy", vol. XXX, 114 (1955), ss. 202-206 oraz tenże, *On Not Worshiping Facts*, „The Philosophical Quarterly", vol. VIII, 31 (1958), s. 144.

[6] W prywatnej rozmowie.

[7] Rogers, dz. cyt., ss. 152 i kolejne.

dopuszczać możliwość przeprowadzania nie-dedukcyjnych inferencji, a to mogłoby wytyczyć drogę ucieczki przed rezultatem Gödla. Hartley Rogers wysuwa konkretną propozycję, w myśl której maszyna miałaby być zaprogramowana tak, by rozważać różne formuły, które nie zostały dowiedzione ani obalone, i od czasu do czasu dopisywać je do listy aksjomatów. Ostatnie twierdzenie Fermata* lub hipoteza Goldbacha mogłyby zostać dopisane w taki właśnie sposób. Jeśli później okazałoby się, że włączenie ich w poczet aksjomatów prowadzi do sprzeczności, wówczas zostałyby odrzucone i w tych okolicznościach oczywiście ich negacje zostałyby dopisane do listy twierdzeń. Tak oto można by zbudować maszynę, która byłaby zdolna do przedkładania jako prawdziwych pewnych formuł, niemożliwych do wywiedzenia z jej aksjomatów w oparciu o jej reguły inferencji. Z tego też względu ta [tj. odwołująca się do twierdzenia Gödla – MZ] metoda dowodzenia wyższości umysłu nad maszyną mogłaby stracić rację bytu.

Konstrukcja takiej maszyny nastręcza jednak trudności. Nie mogłaby bowiem przyjmować wszystkich niedowodliwych formuł i dopisywać ich do listy aksjomatów, gdyż w efekcie przyjmowałaby zarówno formuły gödlowskie, jak i ich negacje i w ten sposób stawała się sprzeczna. Podobnie, gdyby przyjmowała jedynie pierwsze z pary zdań nierozstrzygalnych i, włączając je w poczet swoich aksjomatów, nie uważałaby już jego negacji za nierozstrzygalną, a co za tym idzie – odrzuciłaby ją. Mogłaby przecież postąpić tak z niewłaściwym elementem pary – mogłaby przyjąć negację formuły gödlowskiej, a nie samą tę formułę. System ustanowiony przez standardowy zbiór aksjomatów wraz z dołączoną negacją formuły gödlowskiej, mimo że niesprzeczny, jest systemem niesolidnym, uniemożliwiającym naturalną interpretację. To coś jakby geometrie nie-rzutowe* w dwóch wymiarach – niby niesprzeczne, ale raczej niewłaściwe, wystarczająco niewłaściwe, nie zajmować się nimi na poważnie. Maszyna, która podatna byłaby na tego rodzaju dolegliwości, nie stanowiłaby modelu ludzkiego

* Tzw. Wielkie Twierdzenie Fermata zostało ostatecznie dowiedzione w 1994 roku przez amerykańskiego matematyka Andrew Wilesa.

* Tj. odrzucające tzw. twierdzenie Desarguesa, które jest niezależne od aksjomatów geometrii Euklidesa.

umysłu.

Staje się więc jasne, że niezbędne byłyby bardziej wnikliwe kryteria selekcji formuł niedowodliwych. Hartley Rogers podsuwa pewne możliwości. Lecz nawet jeśli mamy już reguły dołączania nowych aksjomatów, nawet jeśli dołączone aksjomaty przyjmowane są tylko tymczasowo i mogą zostać odrzucone, gdyby okazało się, że prowadzą do sprzeczności, możemy zabrać się do gödlowania tego systemu, jak każdego innego. Jesteśmy w tym samym położeniu, co wówczas, gdy dysponowaliśmy regułą włączającą nieskończony zbiór formuł gödlowskich w poczet aksjomatów. Krótko mówiąc, jakkolwiek zaprojektowana jest maszyna, musi działać losowo lub w oparciu o ściśle określone reguły. Tak długo, jak jej działania są losowe, nie możemy jej przechytrzyć. Jej postępowanie nie będzie jednak przekonującym odzwierciedleniem zachowania inteligentnej istoty. Tak długo natomiast, jak działa ona zgodnie ze ściśle określonymi regułami, możemy użyć metody Gödla do wygenerowania zdania, którego maszyna, w zgodzie z tymi regułami, nie może przedłożyć jako prawdziwego; my natomiast, stojąc poza systemem, widzimy jego prawdziwość[8].

Twierdzenie Gödla stosuje się jedynie do niesprzecznych systemów. Wszystko zaś, czego możemy formalnie dowieść, to że jeśli system jest niesprzeczny, ta formuła gödlowska jest nidedowodliwa-w-systemie. By móc kategorycznie stwierdzić, że formuła gödlowska jest niedowodliwa w systemie, i dlatego prawdziwa, musimy nie tylko mieć do czynienia z niesprzecznym systemem, lecz także być w stanie orzec, że jest niesprzeczny. A – jak pokazał Gödel w swym drugim twierdzeniu, będącym następstwem pierwszego – niemożliwym jest dowiedzenie w niesprzecznym systemie tego, że jest on niesprzeczny. Tak oto, by wskazać usterkę maszyny poprzez skonstruowanie formuły, co do której wiemy zarazem, że jest prawdziwa i że maszyna

[8] Oryginalny dowód Gödla można zastosować, gdy dana reguła jest taka, że generuje pierwotnie rekurencyjny zbiór dodatkowych formuł, K. Gödel, dz. cyt., § 1 oraz § 6. Właściwie wystarczy, by zbiór był rekurencyjnie przeliczalny. Zob. B. Rosser, *Extensions of some theorems of Gödel and Church*, „Journal of Symbolic Logic", vol. I, 3 (1936), ss. 87-91.

nie może przedłożyć jej jako takiej, wpierw musimy być w stanie stwierdzić, że maszyna (lub raczej, odpowiadający jej system formalny) jest niesprzeczna, a na to nie możemy przeprowadzić niepodważalnego dowodu. Wszystko, co jesteśmy w stanie zrobić, to sprawdzić maszynę dokładnie i zobaczyć, czy wygląda na niesprzeczną. Zawsze istnieje jednak ryzyko, że jakaś sprzeczność pozostała niewykryta. W najlepszym razie moglibyśmy powiedzieć, że maszyna jest niesprzeczna, pod warunkiem, że sami jesteśmy. Ale jakim prawem mielibyśmy to zrobić? Drugie twierdzenie Gödla zdaje się pokazywać, że człowiek nie może orzec o sobie samym, że jest niesprzeczny. Dlatego Hartley Rogers[9] argumentuje, że w istocie nie możemy użyć pierwszego twierdzenia Gödla, by odrzucić tezę mechanicyzmu, dopóki nie będziemy w stanie powiedzieć, że "są pewne dystynktywne cechy, które umożliwiają istocie ludzkiej wykroczenie poza owo ostatnie ograniczenie i stwierdzenie swej własnej niesprzeczności, a przy tym pozostanie niesprzecznym."

Odruchową reakcją ludzi, gdy kwestionowana jest ich niesprzeczność, jest zajadła jej obrona; lecz to, z uwagi na drugie twierdzenie Gödla, brane jest przez niektórych filozofów za dowód ich rzeczywistej sprzeczności. Profesor Putnam[10] zasugerował, że istoty ludzkie są maszynami, lecz maszynami sprzecznymi. Gdyby maszyna była tak skonstruowana, by reprezentować sprzeczny system, wówczas nie istniałaby poprawnie zbudowana formuła, której nie mogłaby
przedłożyć jako prawdziwej. I tak oto nie można by wykazać, że jest gorsza niż istota ludzka. Nie moglibyśmy czynić jej zarzutu z jej sprzeczności – czyż ludzie nie są tak samo sprzeczni? Z pewnością kobiety tak. I politycy. Nawet mężczyźni nie będący politykami czasem sobie zaprzeczają, a przecież pojedyncza sprzeczność wystarcza do tego, by usprzecznić system.

To, że od czasu do czasu wszyscy bywamy sprzeczni, nie ulega wątpliwości, lecz z tego nie wynika jeszcze, że jesteśmy równoważni sprzecznym systemom. Nasze sprzeczności

[9] H. Rogers, dz. cyt., s. 154.

[10] W prywatnych rozmowach na Uniwersytecie w Princeton w stanie New Jersey, USA.

są raczej pomyłkami niż ustalonymi zasadami. Odpowiadają przypadkowym awariom maszyny, nie zaś normalnemu schematowi działań. Świadczyć o tym może fakt, że wystrzegamy się sprzeczności, gdy tylko rozpoznamy je jako takie. Gdybyśmy rzeczywiście byli maszynami sprzecznymi, powinniśmy pozostać zadowoleni z naszych sprzeczności i ochoczo uznawać oba ich człony. Co więcej, bylibyśmy gotowi stwierdzić absolutnie wszystko, a to przecież nie ma miejsca. Łatwo można pokazać[11], że w sprzecznym systemie formalnym wszystko jest dowodliwe, zatem warunkiem niesprzeczności systemu staje się fakt, że nie wszystko można w nim udowodnić – że nie jest przypadkiem, w którym "wszystko wolno". To z pewnością charakteryzuje działania umysłów istot ludzkich: są one selektywne – odróżniają stwierdzenia pożądane-prawdziwe od niepożądanych-fałszywych; gdy dana osoba gotowa jest stwierdzić absolutnie wszystko i zaprzeczać sobie bez żadnych skrupułów i odrazy, wówczas uznaje się, że "straciła rozum". Istoty ludzkie, mimo że niezupełnie niesprzeczne, są nie tyle sprzeczne, ile omylne.

Omylna, lecz samokorygująca maszyna wciąż podlegałaby wynikom Gödla. Jedynie taka, która byłaby całkowicie sprzeczna, wymknęłaby się im. Czy moglibyśmy zatem otrzymać całkowicie sprzeczną, a zarazem samokorygującą maszynę, której nie dotyczyłyby wyniki Gödla i która jednocześnie nie byłaby trywialna i całkowicie niepodobna do istoty ludzkiej? Maszynę ze swego rodzaju ukrytą sprzecznością wkomponowaną w nią w taki sposób, że dla wszelkich normalnych celów byłaby niesprzeczna, lecz gdyby podsunąć jej formułę gödlowską, byłaby w stanie ją udowodnić?

Istnieje cała gama sposobów na wyeliminowanie niepożądanych dowodów. Moglibyśmy wprowadzić regułę, że ilekroć dowiedliśmy p i nie-p, sprawdzamy ich dowody i odrzucamy dłuższy z nich. Moglibyśmy też ułożyć aksjomaty i reguły inferencji w określonym porządku i gdy dany byłby dowód prowadzący do sprzeczności, sprawdzilibyśmy które aksjomaty i reguły są niezbędne do jego

[11] Zob. np. A. Church, *Introduction to Mathematical Logic*, vol. I, Princeton 1956, § 17, s. 108.

konstrukcji i odrzucali aksjomat w owym porządku ostatni. W taki lub podobny sposób otrzymalibyśmy system sprzeczny z tzw. “stop-regułą”, tak że sprzeczność nigdy nie wyszłaby na światło dzienne w postaci sprzecznej formuły.

Ta propozycja na pierwszy rzut oka prezentuje się kusząco – jest w niej jednak coś zasadniczo błędnego. Nawet jeśli zachować pozory możemy niesprzeczności dzięki regule mówiącej, że ilekroć pojawią się dwa sprzeczne zdania, winniśmy odrzucić to, którego dowód jest dłuższy, to taka reguła będzie odpychająca, jeśli odwołamy się do naszego zmysłu logicznego. Nawet mniej arbitralne propozycje są wciąż zbyt arbitralne. System nie wykonuje już operacji na pewnych ściśle określonych formułach zgodnie z pewnymi ściśle określonymi regułami inferencji. Zamiast tego reguły będą stosowalne, a aksjomaty pozostaną prawdziwe pod warunkiem, że... nie uznamy tego za niedogodne. Tak oto tracimy grunt pod nogami. Jednym razem zastosowanie reguły modus ponens może być dopuszczone, innym zaś – odrzucone. W jednym wypadku aksjomat może być prawdziwy, w innym – jak widać – fałyszywy. Taki system przestałby być formalnym systemem logicznym, a odpowiadająca mu maszyna nieszczególnie zasługiwałaby na miano modelu umysłu. Byłaby bowiem daleka od podobieństwa do umysłu w jego operacjach – umysł faktycznie sprawdza aksjomaty i reguły inferencji wzbudzające naszą wątpliwość, gdy jednak uzna, że prowadzą do sprzeczności, odrzuca je w całości. To prawda, że wstępnie wypróbowujemy aksjomaty i reguły inferencji, nie zachowujemy ich jednak, gdy już odkryjemy, że prowadzą do sprzeczności. Możemy próbować zastąpić je innymi, możemy dojść do wniosku, że nasza formalizacja jest błędna i dlatego – mimo że pewien aksjomat lub reguła inferencji danego rodzaju jest niezbędna – nie byliśmy w stanie poprawnie ich sformułować. Nie zachowamy jednak błędnych sformułowań bez zmiany, jedynie z zastrzeżeniem, że jeśli dowód prawadzi do sprzeczności, odmówimy jego przeprowadzenia. Takie zachowanie byłoby kompletnie nieracjonalne. Znaleźlibyśmy się w sytuacji, w której – w pewnych przypadkach, jeśli satysfakcjonowałyby nas przesłanki reguły modus ponens – stosowalibyśmy ją i dopuszczali jej wniosek, zaś w innych – odmawialibyśmy zastosowania tej reguły i nie dopuszczali wniosku. Osoba lub

maszyna, która uczyniłaby to bez podania dobrego powodu, zostałaby uznana za arbitralną i nieracjonalną. W pojęciach "argumentu" lub "powodu" zawiera się to, że są w pewnym sensie ogólne i powszechnie obowiązujące – tj. jeśli modus ponens jest właściwą metodą argumentowania, gdy ja wyprowadzam pożądany wniosek, pozostaje on właściwą metodą również wtedy, gdy ty – mój przeciwnik – wyprowadzasz wniosek, którego ja nie chcę przyjąć. Nie możemy dowolnie wybierać sytuacji, w których dana forma argumentowania jest właściwa; nie, jeśli mamy być rozsądni. Jest oczywiście prawdą, że w naszych prywatnych, nie w pełni sformalizowanych, dyskusjach odróżniamy od siebie pewne argumenty, które na pierwszy rzut oka wyglądają podobnie, podając powody, dlaczego w rzeczywistości nie są podobne. Można utrzymywać, że podobnie maszyna mogłaby być uprawniona do odróżniania od siebie pewnych argumentów, które na pierwszy rzut oka wyglądają podobnie, jeśli miałaby dla tego dobry powód. I można dalej utrzymywać, że maszyna ma dobry powód dla odrzucenia tych typów argumentów, które odrzuciła, a właściwie – najlepszy z powodów – tzn. uniknięcie sprzeczności. Ale jest to – jeśli w ogóle można to nazwać powodem – kiepski powód. Nie zyskuje uznania ten, kto unika sprzeczności jedynie poprzez odrzucanie tych argumentów, które by go do niej prowadziły i motywuje to jedynie tym, że w przypadku ich przyjęcia zostałby do sprzeczności doprowadzony. Takie rozumowanie nazwać można stosowaniem podwójnych standardów, a nie porządnym argumentem. Nie zyska uznania nikt, kto – wystarczająco bystry, by tok dyskusji przewidzieć kilka kroków naprzód – sytuacji, w której zmuszony by został do uznania własnej sprzeczności unika poprzez przerwanie owej dyskusji, gdy tylko zorientuje się, dokąd ta prowadzi. Bylibyśmy raczej skłonni określić go jako sprzecznego – w tym przypadku nie dlatego, że przyjął, a potem odrzucił to samo zdanie, lecz dlatego, że użył, a potem odmówił użycia tej samej reguły inferencji. "Stop-reguła", zapobiegająca wyprowadzeniu sprzeczności, to za mało, by uchronić sprzeczną maszynę od nazwania ją sprzeczną.

Pozostaje więc możliwość, że to my jesteśmy sprzeczni i nie ma żadnej "stop-reguły", ale sprzeczność jest tak ukryta, że nigdy nie wychodzi na światło dzienne. Ostatecznie,

naiwna teoria mnogości, która była silnie osadzona w zdroworozsądkowym sposobie myślenia, okazała się sprzeczna. Czy możemy być pewni, że podobny los nie spotka również prostej arytmetyki? Właściwie nie – choć żywimy głębokie poczucie pewności, że nasz system liczb całkowitych, które można dodawać i mnożyć, nigdy nie okaże się sprzeczny. Jedyne, czego nie można wykluczyć, to że odkryjemy, iż nieprawidłowo go sformalizowaliśmy. Gdyby tak się stało, powinniśmy spróbować na nowo sformułować intuicyjne pojęcie liczby, tak jak to uczyniliśmy ze zbiorami. Gdybyśmy to zrobili, musielibyśmy oczywiście zmienić nasz system – nasze obecne aksjomaty i reguły inferencji zostałyby całkowicie odrzucone – nie byłoby wówczas kwestii korzystania i niekorzystania z nich w "sprzeczny" sposób. Po zmianie systemu bylibyśmy w takiej samej sytuacji, w jakiej znajdujemy się w tej chwili, posiadając system, wierząc w jego niesprzeczność, lecz nie mogąc jej dowieść. Lecz czy nowy system nie mógłby również być obciążony jakąś inną sprzecznością? Właściwie istnieje taka możliwość. Lecz znów, żadna sprzeczność, gdy już zostanie wykryta, nie będzie tolerowana. Jesteśmy zdeterminowani, by nie być sprzecznymi i zdecydowani rugować sprzeczność, gdy tylko się pojawi. Tak oto, choć nie możemy być zupełnie spokojni lub wolni od ryzyka konieczności zrewidowania naszej matematyki, ostateczne rozwiązanie musi być jednym z dwojga: albo posiadamy system z prostą arytmetyką, który, wedle naszej najlepszej wiedzy i przekonania, jest niesprzeczny, albo też żaden taki system nie jest możliwy. W pierwszym przypadku znajdujemy się w takiej sytuacji, jak obecnie; w drugim zaś, jeśli odkryjemy, że żaden system zawierający prostą arytmetykę, nie może być wolny od sprzeczności, będziemy musieli zaniechać paranią się nie tylko matematyką i naukami matematycznymi, ale wszelką myślą w ogóle.

Nadal można utrzymywać, że mimo iż człowiek w pewnym sensie musi zakładać swą niesprzeczność, to jednak nie jest w stanie zasadnie jej stwierdzić bez zaprzeczania własnym słowom. Możemy być niesprzeczni. Właściwie mamy wszelkie przesłanki, by wierzyć, że jesteśmy; jednakże wrodzona skromność nie pozwala nam mówić o tym otwarcie. Nie to jednak głosi drugie twierdzenie Gödla. Gödel pokazał, że dla niesprzecznego systemu formuła stwierdzająca niesprzeczność tego systemu, nie może

być w nim dowiedziona, a co za tym idzie, że maszyna, o ile jest niesprzeczna, nie jest w stanie przedłożyć jako prawdziwego twierdzenia o swej własnej niesprzeczności. Stąd też wynika, że również umysł, gdyby rzeczywiście był maszyną, nie mógłby wyprowadzić wniosku, że jest niesprzeczny. Dla umysłu, który nie jest maszyną, powyższy wniosek nie wynika. Wszystko, czego Gödel dowiódł, to że umysł nie może przeprowadzić formalnego dowodu niesprzeczności systemu formalnego wewnątrz tego systemu. Nie ma jednak przeszkód, by wyjść poza system i podać nieformalne argumenty na rzecz niesprzeczności systemu formalnego lub czegoś mniej formalnego i usystematyzowanego. Takie nieformalne argumenty nie będą mogły być całkowicie sformalizowane; lecz w ten oto sposób cały wydźwięk rezultatów Gödla jest taki, że nie powinniśmy szukać całkowitej formalizacji i nie możemy jej osiągnąć. I choć byłoby miło, gdyby była ona możliwa, jako że w pełni sformalizowane argumenty mają większą moc niż te nieformalne, to jednak nie jesteśmy w stanie sprowadzić wszystkich argumentów do sformalizowanej postaci. Nie wolno więc nieformalnym argumentom czynić zarzutów z tego, że są nieformalne lub uważać ich wszystkich za zupełnie pozbawione wartości. Dlatego też wydaje mi się zarówno właściwym, jak i rozsądnym, by umysł stwierdzał swą własną niesprzeczność. Właściwym, bo choć maszyny, jak mogliśmy się spodziewać, są niezdolne do pełnej refleksji nad swym własnym postępowaniem i możliwościami, to jednak tego rodzaju samoświadomości oczekujemy od umysłów. Rozsądnym zaś ze wspomnianych powodów. Nie tylko możemy uczciwie stwierdzić, że wiemy, iż jesteśmy niesprzeczni, pomimo błędów, które popełniamy, ale wręcz musimy w każdym wypadku zakładać, że jesteśmy, jeśli jakakolwiek myśl w ogóle ma być możliwa. Co więcej, cechujemy się selektywnością, nie będziemy mówić wszystkiego jak leci, tak jak mogłaby to zrobić maszyna. I wreszcie, możemy, w pewnym sensie, zdecydować się na bycie niesprzecznymi; w tym mianowicie, że możemy postanowić nie tolerować sprzeczności w naszym myśleniu i mówieniu oraz eliminować je, gdy tylko się pojawią, poprzez wyparcie się i odwołanie jednego z członów sprzeczności.

Teraz widzimy, dlaczego możemy oczekiwać, że twierdzenie Gödla odróżni istoty

samoświadome od przedmiotów martwych. Istotą formuły gödlowskiej jest jej samozwrotny charakter. Mówi ona, że “Ta formuła jest niedowodliwa-w-systemie”. Jeśli teraz odniesiemy to do maszyny, formuła owa wyrażona będzie w języku zależnym od tej konkretnej maszyny. Maszynie zadaje się pytanie o jej własne procesy. Żądamy od niej, by była samoświadoma i powiedziała nam, które rzeczy może robić, a których nie może. Takie pytania nieustannie prowadzą do paradoksów. Przy pierwszych i najprostszych próbach filozofowania, każdy uwikłany zostaje w pytania o to, czy kiedy ktoś wie coś, to czy wie o tym, że to wie i co właściwie zostaje pomyślane, gdy ktoś myśli o sobie, i co sprawia, że myślimy. Po tym, jak problem ów będzie go intrygował i męczył przez długi czas, nauczy się, by nie stawiać tych pytań otwarcie – jego pojęcie istoty świadomej niepostrzeżenie stanie się różne od jego pojęcia przedmiotu nieświadomego. Gdy mówimy, że istota świadoma coś wie, nie mamy na myśli jedynie tego, że wie, lecz także to, iż wie, że wie i wie, że wie, że wie – i tak dalej, tak długo, jak zechcemy stawiać pytanie. A zatem pojawia się tu nieskończoność; nie jest to jednak regres w nieskończoność w niepożądanym sensie, ponieważ to raczej pytania w końcu wyczerpują się, jako prowadzące donikąd, nie zaś odpowiedzi. Pytania te zdają się prowadzić donikąd, gdyż w samym pojęciu [istoty świadomej – MZ] zawiera się już zdolność do udzielania na nie odpowiedzi w nieskończoność. Mimo że istoty świadome mogą tak w nieskończoność, nie chcielibyśmy przedstawiać tego jako ciągu zadań, które są w stanie realizować. Nie pojmujemy też umysłu jako ciągu jaźni, nad-jaźni i nad-nad-jaźni. Sądzimy raczej, że istota świadoma jest jednością i mimo że mówimy o częściach umysłu, zwrotu tego używamy jako metafory i w żadnym wypadku nie powinien on być brany dosłownie.

Prowadzi to do paradoksów świadomości, jako że istota świadoma może być zarówno świadoma siebie samej, jak i innych rzeczy, a zarazem nie może być pojmowana jako rozkładalna na części. To oznacza, że istota świadoma potrafi poradzić sobie z pytaniami gödlowskimi wsposób dla maszyny nieosiągalny, gdyż istota świadoma może zarówno rozważać siebie samą, jak i swoje działanie i nie być różną od tej, która to działanie wykonuje. Można, w pewnym sensie, zbudować maszynę, która mogłaby

"rozważać" swoje działanie, ale nie mogłaby "wziąć tego pod uwagę", nie stając się tym samym inną maszyną, a dokładniej – starą maszyną z dołączoną "nową częścią". Natomiast nieodłącznie związane z naszym wyobrażeniem istoty świadomej jest to, że może ona dokywać autorefleksji i poddawać ocenie swoje działania, nie wymagając przy tym żadnej dodatkowej części – jest zupełna i nie ma żadnej pięty achillesowej.

Nasza teza staje się w ten sposób sprawą raczej analizy pojęciowej, aniżeli matematycznego odkrycia. Potwierdzi się to, gdy rozważymy kolejny argument, wysunięty przez Turinga[12]. Do tej pory konstruowaliśmy jedynie dość proste i przewidywalne artefakty. W miarę jednak, jak zwiększać się będzie złożoność naszych maszyn, mogą pojawić się niespodzianki. Turing przyrównuje tę sytuację do stosu atomowego. Poniżej pewnej masy "krytycznej" niewiele się z nim dzieje, lecz gdy masa krytyczna zostanie przekroczona, zaczyna iskrzyć. Być może podobnie jest z mózgami i maszynami. Większość mózgów i wszystkie maszyny są w tej chwili w stanie "podkrytycznym" – reagują na napływające bodźce w sposób nudny i mało interesujący, nie mają wyobrażenia siebie samych, ich reakcje są szablonowe; lecz niektóre mózgi w tej chwili, i być może pewne maszyny w przyszłości, są nadkrytyczne i grają na własną rękę. Turing sugeruje, że to tylko kwestia złożoności i że po osiągnięciu pewnego jej stopnia pojawi się różnica jakościowa, zatem "nadkrytyczne" maszyny będą zupełnie niepodobne do tych, jakie wyobrażaliśmy sobie do tej pory.

Możliwe, że tak jest w istocie. Złożoność często pociąga za sobą zmiany jakościowe. Choć brzmi to niewiarygodnie, to jednak może się okazać, że powyżej pewnego poziomu złożoności maszyna przestanie być przewidywalna, nawet z zasady, i zacznie działać na własną rękę lub, co zabrzmi bardzo wymownie, zacznie posiadać swój własny umysł. Stałoby się to z chwilą, w której nie byłaby już całkowicie przewidywalna i uległa, za to zdolna do wykonywania działań, które my uznajemy za inteligentne, nie zaś za pomyłki lub przypadkowe strzały, lecz do których zarazem nie została zaprogramowana. Lecz w takim wypadku przestałaby ona być maszyną w klasycznym

[12] A. Turing, dz. cyt., s 454.

rozumieniu tego słowa. W mechanicystycznej debacie nie idzie o to, jak umysły są, lub mogłyby być, powoływane do życia, ale jak działają. Podstawą tezy mechanicyzmu jest to, że mechaniczny model umysłu powinien działać w oparciu o "mechaniczne zasady", tj. że możemy zrozumieć działanie całej maszyny, odwołując się do działania jej części i że działanie każdej z części albo jest zdeterminowane przez jej stan początkowy i budowę maszyny, albo też jest wynikiem losowego wyboru spośród ściśle określonej liczby ściśle określonych operacji. Jeśli mechanicysta zbuduje maszynę, która będzie tak skomplikowana, że nie zachowa tego, co właściwe maszynie, wówczas dla celów naszej dyskusji nie będzie to już maszyna, bez względu na to, jak została zbudowana. Powinniśmy wtedy rzec raczej, że stworzył umysł w takim samym sensie, jak my płodzimy ludzi. Byłyby zatem dwa sposoby powoływania na świat nowych umysłów – sposób tradycyjny – poprzez płodzenie dzieci, rodzonych później przez kobiety i sposób nowy – poprzez budowanie szalenie skomplikowanych systemów, powiedzmy, zaworów i przekaźników. Gdy mówimy o tym drugim sposobie, musimy wyraźnie zaakcentować, że mimo iż to, co zostało zbudowane, wygląda jak maszyna, w istocie nią nie jest, gdyż nie jest prostą sumą swoich części. Nie można by przewidzieć, co [maszyna – MZ] zrobi, znając jedynie sposób, w jaki została zbudowana i początkowy stan jej części. Nie można by nawet wskazać granic tego, co może zrobić, jako że nawet jeśli zadalibyśmy jej pytanie w stylu Gödla, podałaby ona właściwą odpowiedź. Powinniśmy właściwie powiedzieć krótko, że każdy system, którego nie udało się przygwoździć pytaniem Gödla, nie jest *eo ipso* maszyną Turinga, tj. maszyną w klasycznym rozumieniu tego słowa.

Jeśli ten dowód fałszywości mechanicyzmu jest słuszny, to ma to olbrzymie konsekwencje dla całej filozofii. Od czasów Newtona widmo mechanicystycznego determinizmu prześladuje filozofów. Gdybyśmy mieli zachować standardy naukowości, wygląda na to, że musielibyśmy spojrzeć na istoty ludzkie jak na zdeterminowane automaty, nie zaś jak na niezależne podmioty moralne. Gdybyśmy zaś mieli być moralni, wygląda na to, że musielibyśmy wyprzeć się nauki, arbitralnie ustanowić kres jej postępu w zrozumieniu ludzkiej neurofizjologii i salwować się ucieczką w

obskurancki mistycyzm. Nawet Kant nie znalazłby rozwiązania sporu pomiędzy tymi dwoma stanowiskami. Teraz jednak, choć wciąż pozostaje wiele argumentów przeciw istnieniu wolności człowieka, argument mechanicystyczny, być może najpoważniejszy ze wszystkich, stracił swą moc. Na filozofach przyrody przestanie ciążyć obowiązek negowania wolności przy pomocy takich zarzutów [mechanicystycznych – MZ] w imię nauki. Moraliści przestaną odczuwać potrzebę zniesienia nauki w imię wiary. Zaczynamy widzieć, jak tworzy się przestrzeń dla moralności, niepociągającej za sobą konieczności zniesienia lub nawet ograniczania którejkolwiek z dziedzin nauki. Nasz argument nie ustanowił granic dla naukowych badań – nadal możliwe jest badanie pracy mózgu. Nadal możliwe jest budowanie mechanicznych modeli umysłu. Po prostu teraz widzimy, że żaden mechaniczny model, ani żadne wyjaśnienie sformułowane w czysto mechanicystycznej terminologii, nie jest zupełnie adekwatny. Możemy budować modele i wyjaśnienia, które z pewnością będą pouczające; jednakże, jakkolwiek zaawansowane będą, zawsze pozostanie coś więcej do powiedzenia. Nie ma arbitralnie ustalonych granic dla naukowych badań, ale żadne naukowe badania nigdy nie mogą wyczerpać nieskończonej różnorodności ludzkiego umysłu.

Przełożył Michał Zawidzki