STUDIA METODOLOGICZNE
NR 33 • 2014

MICHAŁ SOCHAŃSKI

Twierdzenie Gödla a spór o mechanicyzm

Wstęp

Pytania, czy człowiek jest maszyną oraz czy każde jego zachowanie można modelować przez pewną skończoną maszynę, nie przestają być aktualne i szeroko dyskutowane. Jeden z argumentów przeciwko mechanicyzmowi, który na powyższe pytania odpowiada twierdząco, bazuje na twierdzeniu Gödla o niezupełności arytmetyki, najpewniej najszerzej dyskutowanym logiczno-matematycznym twierdzeniu ostatniego stulecia. W najkrótszym sformułowaniu wniosek z tego rozumowania może brzmieć następująco: „umysł ludzki nie działa mechanicznie" lub „umysł nie może być równoważny maszynie cyfrowej". W poszczególnych wersjach argument z twierdzenia Gödla miał jednak także dowodzić platonizmu matematycznego, a nawet istnienia niematerialnej duszy. Stąd za jeden z celów tej pracy stawiam sobie uważną eksplikację założeń argumentu, uściśleń i konwencji terminologiczno-pojęciowych, jak również idealizacji, których przyjęcie jest warunkiem sformułowania samego argumentu. Przyjrzę się tu trzem głównym (a w każdym razie najbardziej znanym) metodom wykorzystania twierdzenia Gödla do argumentacji przeciw mechanicyzmowi – metodzie Johna Lucasa, Rogera Penrose'a i samego Kurta Gödla, spróbuję też naświetlić podobieństwa i różnice pomiędzy nimi. Argument Lucasa pojawił się w jego pracy *Minds, Machines and Gödel* w 1961 roku. Mimo iż przez pewien czas cieszył się on dobrą sławą, z czasem pojawiało się

coraz więcej głosów krytycznych. W latach 90. ubiegłego wieku dyskusję (znów przeważnie krytyczną) wzniecił na nowo Penrose w książkach *Nowy umysł cesarza*, a następnie *Cienie umysłu*. Gödel opisuje swoje rozumowanie w pracy *Some Basic Theorems on the Foundations of Mathematics*. Praca ta, mimo iż została napisana na początku lat pięćdziesiątych, ujrzała światło dzienne stosunkowo niedawno, stąd też można powiedzieć, iż dyskusja nad nią wciąż trwa. Różne aspekty „gödlowskiego" argumentu zostały poddane szerokiej i wnikliwej analizie w książce Stanisława Krajewskiego *Twierdzenia Gödla i jego interpretacje filozoficzne. Od mechanicyzmu do postmodernizmu*. Do niej też będę się w mojej pracy najczęściej odnosił, w szczególności przy omawianiu argumentu Lucasa.

Twierdzenie Gödla o niezupełności

Na początek należy przyjrzeć się bliżej samemu twierdzeniu Gödla. Jego treść jest przedstawiana w różnych popularnonaukowych opracowaniach w sposób mniej lub bardziej dokładny, przy czym im krótszy opis, tym większe ryzyko nieścisłości oraz nietrafionych interpretacji. Mimo to zaryzykuję przedstawienie jego matematycznej treści w ogólnych zarysach (nie dbając o pełny zestaw szczegółów formalnych), gdyż zrozumienie argumentu przeciw mechanicyzmowi będzie bez tego niemożliwe.

Twierdzenie Gödla mówi najogólniej, iż każdy sformalizowany system aksjomatyczny zawierający arytmetykę liczb naturalnych jest niezupełny. Oznacza to, iż dla każdego zbioru aksjomatów, o ile w aksjomatach tych ujęta jest arytmetyka liczb naturalnych, znajdziemy poprawnie sformułowane zdania, których na gruncie tych aksjomatów nie możemy ani udowodnić, ani odrzucić (to znaczy udowodnić ich negacji). Kluczowym elementem rozumowania Gödla jest konstrukcja zdania nierozstrzygalnego na gruncie arytmetyki liczb naturalnych, to znaczy takiego właśnie, które jest niezależne od aksjomatów (mówiąc o sformalizowanej arytmetyce, będę dalej nawiązywać do arytmetyki Peana – teorii sformalizowanej, która jest „standardowym" aksjomatycznym ujęciem arytmetyki liczb naturalnych; teorię tę będę oznaczał skrótowo przez *PA*).

Konstrukcja ta jest niezwykle pomysłowa: Gödel poprzez arytmetyzację metajęzyka[1] formułuje zdanie, które jest tak skonstruowane, iż można je rozumieć jako mówiące „ja nie mam dowodu", czyli „ja nie jestem twierdzeniem arytmetyki Peana"[2]. Następnie dowodzi, iż to właśnie zdanie jest nierozstrzygalne, to znaczy ani ono nie ma dowodu na gruncie arytmetyki Peana, ani jego negacja. Intrygujące i kluczowe dla argumentu w wersji Lucasa jest przy tym to, iż zdanie Gödla jest *prawdziwe*. Intuicyjnie można to wytłumaczyć, zauważając, iż jest tak, jak ono głosi, to znaczy że nie ma ono dowodu. Jest to ujęcie nieco metaforyczne, ma ono jednak również bardzo ścisły sens – zdanie Gödla jest mianowicie prawdziwe zgodnie z używaną w matematyce definicją prawdy Alfreda Tarskiego[3].

Dla dalszego wywodu istotne są jeszcze dwa fakty: po pierwsze, twierdzenie Gödla stosuje się do wszystkich systemów formalnych, które zawierają arytmetykę, czyli w praktyce prawie wszystkich systemów aksjomatycznych rozważanych w matematyce oraz logice. Oznacza to, iż dodanie nowych aksjomatów do systemów nic nie pomoże – mielibyśmy do czynienia z nowym systemem niezupełnym, posiadającym nowe zdania

---

[1] Arytmetyzacja metajęzyka jest jednym z głównych pomysłów Gödla, który umożliwił mu sformułowanie jego twierdzenia. Przypomnijmy najpierw, iż metajęzyk dla arytmetyki jest językiem, w którym mówi się *o arytmetyce*. Arytmetyzacja polega na kodowaniu jego symboli oraz wyrażeń za pomocą liczb naturalnych. Dzięki niej „zamiast o formułach można mówić o liczbach naturalnych i w konsekwencji metamatematyczne wypowiedzi na temat arytmetyki, jej formuł i twierdzeń dają się przetłumaczyć na wypowiedzi o liczbach naturalnych. […] Stąd w arytmetyce *PA* możemy mówić o niej samej!" [Murawski, 2000, s. 83––84].

[2] Łatwo zauważyć, iż zdanie Gödla jest w pewnym sensie zdaniem samoodnośnym, nie jest to jednak typ samoodnośności, który prowadzi w „bezpośredni" sposób do paradoksu takiego, jak na przykład paradoks kłamcy.

[3] Matematycy powiedzieliby dokładniej, iż jest ono prawdziwe w modelu standardowym. Przypomnijmy, iż zgodnie z definicją prawdy Tarskiego prawdziwość polega na spełnieniu danego zdania (a więc formuły zdaniowej, w której każda zmienna jest związana przez pewien kwantyfikator) przez każdy ciąg obiektów z interpretacji. Zdanie Gödla jest zdaniem poprzedzonym dużym kwantyfikatorem, wiążącym pewną zmienną. Mówiąc krótko, jest ono spełnione dla każdego podstawienia liczby naturalnej pod tą zmienną (stąd jego prawdziwość) – nie istnieje jednak dowód tego zdania oparty na aksjomatach i regułach dowodzenia dostępnych w systemie formalnym (dla szczegółów matematycznych zob. Murawski, 2000, s. 92–93).

nierozstrzygalne. Po drugie, dowód twierdzenia bazuje na założeniu, iż badany system jest niesprzeczny[4]. Jest to kluczowe założenie twierdzenia, które okaże się problematyczne dla antymechanicystów.

Mechanicyzm

W dalszej kolejności należy uściślić, co rozumiemy przez tezę mechanicyzmu, aby podjąć próbę ustalenia, w jakiej może ona pozostawać relacji do twierdzenia Gödla. Wcześni mechanicyści, jak Kartezjusz czy La Mettrie, wyrażając myśl, iż człowiek jest maszyną, odwoływali się do takich metafor, jak automatycznie poruszające się koła zębate w zegarze czy mechanizmy w młynie[5]. Jednak dopiero w ubiegłym stuleciu pojawił się aparat pojęciowy, w ramach którego możliwe stało się ścisłe sformułowanie tego, co rozumiemy pod takimi pojęciami jak algorytm lub automatyczna, mechaniczna procedura. Dokładniej, pojawiło się jednocześnie kilka takich aparatów pojęciowych, większość w okresie międzywojennym – były to między innymi rachunek λ Churcha, algorytm Markowa, funkcje rekurencyjne, a przede wszystkim maszyny Turinga. Co najważniejsze, wszystkie te ujęcia okazały się w ściśle formalnym sensie równoważne. Dostarczają one bardzo ogólnej charakteryzacji procedury algorytmicznej – na tyle ogólnej, iż każdy współczesny komputer cyfrowy jest równoważny pewnej maszynie Turinga.

Pozostaje pytanie, czy owe matematyczne ujęcia w pełni wyrażają intuicyjne pojęcie automatycznej procedury, czy algorytmu. Teza, iż każda procedura, którą nazwalibyśmy algorytmiczną, automatyczną, jest w istocie sprowadzalna do pewnej maszyny Turinga – nazywana jest tezą Churcha-Turinga lub tezą Churcha. Nie można jej, jak się wydaje, ściśle matematycznie udowodnić – jest ona natomiast szeroko potwierdzona[6]. Teza

[4] System jest niesprzeczny, gdy nie istnieje taka formuła, że na gruncie tego systemu można udowodnić zarówno ją, jak i jej negację.

[5] Warto przy tym pamiętać, iż Kartezjusz – zgodnie ze swoim dualizmem – przyjmował mechanicyzm tylko w odniesieniu do świata materialnego.

[6] Dokładniej, teza Churcha stwierdza iż klasa funkcji obliczalnych jest dokładnie równa klasie funkcji rekurencyjnych. Pojęcie funkcji obliczalnej jest intuicyjne, niesformalizowane – funkcja jest obliczalna, gdy istnieje pewna mechaniczna, z góry określona, metoda obli-

Churcha jest dla naszych rozważań ważna między innymi dlatego, iż jeśli na nią przystaniemy, mówiąc o mechaniczności człowieka, maszynach, które symulują zachowania człowieka, czy wreszcie o jakichkolwiek algorytmach, wystarczy, jeśli będziemy mówić po prostu o maszynach Turinga. Tezie o mechaniczności człowieka (czy umysłu) można więc nadać konkretniejszą postać, a głównym kandydatem na model ludzkiego umysłu jest maszyna Turinga.

Można powiedzieć, iż współczesną wersją mechanicyzmu jest teza, iż możliwe jest stworzenie sztucznej inteligencji, to znaczy maszyny w jakiś sposób równoważnej człowiekowi. Program zwolenników tej tezy jest w ogólnych zarysach jasny, jednak dokładne sformułowanie tezy o możliwości stworzenia maszyny odpowiadającej człowiekowi napotyka na niemałe trudności. Teza taka powinna postulować jakąś relację pomiędzy maszynami i umysłem; pojęcie maszyny jest sformułowane jasno, problem tkwi w tym, co należy rozumieć pod pojęciem umysłu, i wreszcie – przez relację, jaka zachodzi między umysłem a maszyną. Mówimy przecież z jednej strony o jasno zdefiniowanym pojęciu logiczno-matematycznym, a z drugiej niejasno, mocno filozoficznie zabarwionym pojęciu umysłu. Stąd tezy zwolenników istnienia sztucznej inteligencji w swoim najogólniejszym sformułowaniu brzmią niezbyt jasno. Oto najważniejsze dwie z nich:

A) Myślenie to obliczanie.

B) Umysł jest maszyną (to znaczy funkcjonuje w ogólności jak maszyna).

Cóż miałoby jednak oznaczać, iż „myślenie to obliczanie"? Takie sformułowanie można traktować jako metaforę, luźną analogię. W reakcji między innymi na te właśnie trudności współczesna filozofia umysłu wypracowała dwie wersje tezy o sztucznej inteligencji. Silna (silna teza AI – od *Artificial Intelligence*) głosi, iż działanie umysłu można w zupełny sposób sprowadzić do działania maszyny. Słabsza teza AI głosi natomiast, iż maszy-

---

czania jej wartości dla każdego argumentu. Funkcje rekurencyjne są z kolei ściśle zdefiniowaną klasą funkcji. Potwierdzenie tezy Churcha polega między innymi tym, iż nie odnaleziono jeszcze funkcji obliczalnych, które nie były rekurencyjne, oraz na tym (o czym wspominam powyżej), iż wszystkie dotychczas sformułowane formalizacje pojęcia obliczalności okazały się równoważne [zob. Murawski 2000, s. 14, 63–65]).

na może *symulować* każde ludzkie zachowanie, z różnych względów jednak maszyna nigdy nie może być równoważna człowiekowi[7]. Mamy więc:

C) Każda działalność umysłu może być symulowana przez maszyny.

Sformułowanie to kryje jednak jeszcze jeden niuans, który będzie bardzo istotny dla dalszych rozważań. Możemy mianowicie odróżnić dwie tezy (sformułuję je, odnosząc się już do maszyn Turinga):

> **T1**. Istnieje *jeden* algorytm (czyli można skonstruować jedną maszynę), który jest w stanie modelować każdą (możliwą) działalność umysłu.
>
> **T2**. Możliwe jest modelowanie każdego zachowania umysłu poprzez *pewną* maszynę[8].

Teza **T1** jest o wiele mocniejsza niż **T2**. Ta druga jest szeroko potwierdzona i rzeczywiście wygląda na to, iż dużą część zachowań ludzi można obecnie symulować komputerowo. Należy jednak uważać – **T1** nie wynika z **T2**[9]! Jak dalej zobaczymy, zwolennicy argumentu z twierdzenia Gödla argumentują najczęściej przeciw tezie **T1**, czasem jednak również przeciw **T2**. Wydaje się również, iż argumenty gödlowskie – jeśli „działają" – obalają też słabą tezę AI, co pociąga oczywiście odrzucenie silnej tezy AI.

Jak do wyrażonych w powyższych tezach ujęć mechanicyzmu ma się twierdzenie Gödla? Jest ono przecież twierdzeniem czysto matematycznym, może więc powiedzieć coś bezpośrednio tylko o aktywności matematycznej. Aby bezpośrednio zastosować nasze twierdzenie, należy nieco przeformułować tezę **T1**:

---

[7] Silniejsza teza jest głoszona w ostatnich latach coraz rzadziej. W latach dziewięćdziesiątych została ona skrytykowana z różnych punktów widzenia oraz przez wielu filozofów, jak Searle czy Putnam. Nie jest moim celem ich analiza, a jedynie ocena, czy oraz na ile istotne jest w tej dyskusji twierdzenie Gödla.

[8] Powyżej podaję tylko niektóre – moim zdaniem ważniejsze – sformułowania tezy o mechaniczności umysłu. Można ich jeszcze wymienić co najmniej kilka.

[9] Logiczną strukturę pierwszej tezy można przybliżyć następująco: *dla każdej* czynności ludzkiej *istnieje* maszyna, które je symuluje; drugiej tezy natomiast: *istnieje* maszyna, która *dla każdej* czynności ludzkiej czynność tę symuluje. Tezy te różnią się więc kolejnością występowania w nich kwantyfikatorów dużego i małego. Zdanie „T2→T1" nie jest prawdziwe, gdyż nie jest tezą odpowiadająca mu formuła języka rachunku predykatów [zob. Krajewski, 2003, s. 133].

T1'. Istnieje maszyna Turinga równoważna umysłowi (czy symulująca umysł) pod tym względem, że dowodzi dokładnie tych samych twierdzeń matematycznych.

Relację umysłu do maszyn oceniać więc będziemy pod względem ich zdolności dowodzenia twierdzeń matematycznych. Można więc powiedzieć, iż teza **T1'** odpowiada tezie **T1,** o ile w miejsce terminu „umysł" podstawimy termin „aktywność matematyczna". Możemy jednak przyjąć, iż każdy wniosek, który otrzymamy co do ludzkich umiejętności w dziedzinie matematyki, można też (przynajmniej w pewnym sensie) przenieść na człowieka jako całość – mówimy wtedy: jeśli człowiek przewyższa maszynę (w szczególności) w jakimś fragmencie matematyki, to przewyższa ją w ogóle. Tak też rozumują zwolennicy argumentu z twierdzenia Gödla. Muszę tu zaznaczyć, iż mimo powyższych uściśleń będę dalej używał takich niejasnych sformułowań, jak „jesteśmy równoważni pewnej maszynie/algorytmowi", czy „pewna maszyna w pełni nas opisuje"; należy pamiętać, iż powinno się na ich miejsce zawsze podstawiać tezę **T1'**.

Warto wspomnieć, iż można rozpatrywać drugą grupę tez poprzez podstawienie w tych powyższych zamiast słowa „umysł", słowa „ciało". W takiej sytuacji pojawiłoby się dodatkowe pytanie: „czy umysł jest *niemechaniczny w odróżnieniu od ciała*?". Tak przecież uważał Kartezjusz. Jest to temat na osobną dyskusję; można w każdym razie założyć, iż jeśli przyjmiemy którąś z powyższych tez w odniesieniu do umysłu, będziemy musieli przyjąć ją również dla ciała (głębszym problemem filozoficznym jest pytanie, czy zachodzi implikacja odwrotna).

Założenia „gödlowskiego" argumentu

Możemy wreszcie połączyć wszystkie dotychczasowe rozważania i sformułować sam argument w „klasycznej" wersji Lucasa. Podsumujmy więc – zakładamy, iż:

Z1. Każdą algorytmiczną procedurę można sprowadzić do pewnej maszyny Turinga (teza Churcha).

Co więcej,

Z2. Każda maszyna (a więc również każdy komputer) jest równoważna pewnej maszynie Turinga[10].

Do tego dołączymy również bardzo ważny fakt natury czysto matematycznej. Otóż można przyjąć, iż istnieje bezpośrednia odpowiedniość pomiędzy maszynami Turinga a systemami formalnymi. Polega ona na tym, iż dla każdego systemu aksjomatycznego *S* możemy skonstruować odpowiadającą mu maszynę Turinga *T(S)* – a więc taką, która dowodzi dokładnie tych samych twierdzeń. Bardziej intuicyjnie, analogia ta polega na tym, iż „obliczenie, a ogólniej – sekwencja operacji dokonywanych przez maszynę M, odpowiada dowodowi formalnemu w systemie S" [Krajewski, 2003, s. 94][11]. Do naszych rozważań dołączamy więc kolejne ważna założenie:

Z3. Maszyny Turinga ≈ Systemy aksjomatyczne.

Argument Lucasa i jego krytyka

Możemy sformułować gödlowski argument w „klasycznej" wersji, którą jako pierwszy sformułował oksfordzki filozof John Lucas. Załóżmy, iż ktoś twierdzi, iż pewna konkretna maszyna jest równoważna człowiekowi (na mocy powyższych rozważań maszynie tej odpowiada pewien system

[10] Nie wszyscy zgadzają się, iż każda maszyna jest sprowadzalna do pewnej maszyny Turinga. Niektórzy autorzy analizują możliwość istnienia maszyn, które hiperobliczają (*hypercompute*), a więc takich, które potrafią w szczególności obliczać wartości funkcji, które nie są obliczalne według Turinga [Copeland, 2004, s. 251]. Copeland wymienia tu kilka takich możliwości, między innymi: „maszyny interakcyjne", to znaczy w jakiś sposób „podłączone do środowiska", maszyny „przyśpieszające", a więc zdolne wykonywać nieskończoną ilość kroków, czy wreszcie O-maszyny, a więc maszyny Turinga wzbogacone o dodatkowe moduły („wyrocznie" – *oracles*), zdolne do wykonywania pewnych zadań niedostępnych dla zwykłych maszyn Turinga. Nie mam zamiaru tu oceniać tych propozycji, które są dość abstrakcyjne i jeszcze niezastosowane w praktyce. Należy jednak pamiętać, iż do maszyn nie będących maszynami Turinga rozumowanie gödlowskie się nie stosuje.

[11] Równoważność maszyn Turinga i systemów formalnych idzie jeszcze dalej – twierdzenie Gödla, stosujące się do tych drugich, ma swój analogon w teorii maszyn Turinga. Pewna wersja tego twierdzenia będzie omówiona przy okazji analizy argumentu Penrose'a.

formalny). Otóż na mocy twierdzenia Gödla istnieją zdania, których maszyna ta nie może udowodnić. Co więcej, my wiemy („widzimy", jak pisze Lucas), że te twierdzenia są prawdziwe – wynika to bowiem z konstrukcji dowodu twierdzenia Gödla. Maszyna jednak nie jest w stanie tego dowieść. Wniosek – jesteśmy „lepsi" od dowolnej maszyny i w konsekwencji nie może istnieć maszyna równoważna umysłowi w sensie tezy **T1**, naszego poznania zdania nierozstrzygalnego nie można przy tym modelować przez maszynę, co obala również **T2**.

Takie rozumowanie przeprowadził Lucas w swojej głośnej pracy *Minds*, *Machines and Gödel* i trzeba przyznać, iż na pierwszy rzut oka może się ono wydawać całkiem przekonujące. Warto dodać, iż Lucas sugerował, że jego argumentacja uderza w tezę materializmu. Wydaje się, iż wnioskuje on tutaj następująco: wszystko, co materialne, zachowuje się w sposób mechaniczny (dodatkowe założenie), ale pokazaliśmy, iż umysł jest niemechaniczny, a zatem umysł jest niematerialny. Jest to chyba najmocniejszy wniosek, który próbowano wyciągnąć z twierdzenia Gödla.

Argument przeciw mechanicyzmowi można, jak widać, wyrazić w dość prosty i na pierwszy rzut oka przekonujący sposób. Przy głębszej analizie pojawia się tu jednak dużo luk i niedopowiedzeń. Zastanówmy się na początek, na czym polega wyjątkowa natura zdania nierozstrzygalnego i naszej wiedzy o nim. Lucas sugeruje, iż fakt, że my widzimy jego prawdziwość, a komputer nie, ustanawia naszą nad nim wyższość. Należy tu jednak być uważnym. Prawdziwość zdania nierozstrzygalnego jest przecież zwyczajną prawdziwością matematyczną, prawdziwością w modelu standardowym, o której była mowa powyżej. Co więcej, konstrukcja zdania nierozstrzygalnego jest procedurą czysto mechaniczną – a skoro tak, może ją przeprowadzić każdy komputer, jeśli wyposażyć go w odpowiednią procedurę[12]. Nasza wiedza o prawdziwości zdania Gödla nie wydaje się więc polegać na jakiejś szczególnej umiejętności „widzenia prawdy", szu-

[12] Dokładniej, można pokazać, iż istnieje funkcja rekurencyjna g, która dla każdej maszyny Turinga podaje numer Gödla zdania nierozstrzygalnego skonstruowanego w systemie sformalizowanym odpowiadającej owej maszynie Turinga [Murawski, 1999, s. 328]. Na mechaniczność procedury konstrukcji zdania nierozstrzygalnego zwrócił uwagę już Putnam w: Putnam, 1986.

kanie naszej wyższości nad maszynami w tym zakresie wydaje się więc nieuprawnione. Skoro argumentacja, która prowadzi do stwierdzenia prawdziwości zdań nierozstrzygalnych, nie jest ze swej natury, z zasady, czymś nieosiągalnym dla komputera, można stwierdzić, że twierdzenie Gödla na pewno nie obala tezy **T2**.

Dodajmy, iż pytanie, czy maszyna może w jakiś sposób ujmować semantykę, a więc operować pojęciami takimi jak prawdziwość i rozumieć znaczenie wyrazów, jest, jak się zdaje, szerszą kwestią, niezależną od rozważań wokół twierdzenia Gödla. W słowach Krajewskiego „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" [Krajewski, 2003, s. 104]. Inaczej mówiąc, jeśli założymy, że komputer nie może operować pojęciami semantycznymi, to z góry odrzucamy tezę sztucznej inteligencji i twierdzenie Gödla jest zbędne[13].

Rzekoma wyjątkowość natury i procesu poznania zdań nierozstrzygalnych prowadzi do jeszcze innego problemu. Dziwnym i nieintuicyjnym wydaje się mianowicie stwierdzenie, iż akurat „widzenie" prawdziwości zdania Gödla (a nie jakaś inna aktywność ludzi) jest tym, co odróżnia człowieka od maszyny. John Barrow pisze o tym następująco: „Czy jeżeli nie dostrzegam prawdziwości Gödlowskiej myśli, to znaczy, że nie jestem równie świadomy jak ktoś, kto ją dostrzega, albo że mój mózg mógłby być symulowany przez jakiś algorytm, jego zaś nie?" [Barrow, 1996, s. 407]. Inaczej mówiąc, czyżby niealgorytmicznie myśleli tylko logicy i matematycy rozumiejący twierdzenie Gödla[14]? Problem ten ukazuje bardzo niein-

[13] Z drugiej strony nie możemy wykluczyć, że bardziej zaawansowane maszyny będą w stanie operować pojęciami semantycznymi, choć wielu filozofów jest co do tej możliwości nastawiona sceptycznie.

[14] Pozostając na gruncie matematyki, można również zapytać, dlaczego akurat poznanie prawdziwości zdania nierozstrzygalnego ma charakter niemechaniczny, a nie jakiekolwiek inne twórcze i żmudne dociekania matematyków – choćby te mające na celu dowiedzenia słynnego twierdzenia Fermata czy hipotezy Goldbacha? Przecież właśnie w kontekście takich twierdzeń, i takich badań matematyków myślimy często o „twórczym", „kreatywnym" charakterze działalności matematyków.

tuicyjny aspekt rozumowania Lucasa. Nie przesądza on jednak o jego wadliwości; Krajewski w nawiązaniu do tej kwestii pisze więc, iż „dzieło obalania argumentu Lucasa należy kontynuować nawet wtedy, gdy się je stosuje tylko w odniesieniu do logików, czy wręcz tylko do Lucasa" [Krajewski, 2003, s. 127].

Argument przeciwko mechanicyzmowi jednak jeszcze nie upadł. Można próbować użyć twierdzenia Gödla do obalenia tezy **T1** bez odwoływania się do jakichś szczególnych mocy poznawczych człowieka związanych z rozpoznawaniem prawdziwości zdania nierozstrzygalnego.

Ustaliliśmy, iż procedura konstrukcji zdania nierozstrzygalnego dla danej teorii oraz odpowiadającej jej maszyny jest mechaniczna. Jednak, aby dana maszyna skonstruowała „swoje" zdanie nierozstrzygalne, trzeba dodać do niej nową procedurę, która nie była w niej wcześniej zawarta. Tworzymy wtedy nową maszynę i możemy również dla niej skonstruować kolejne, nowe zdania nierozstrzygalne. Można więc powiedzieć, iż zawsze możemy w ten sposób „wygrać" z maszyną. Taką procedurę konstruowania zdania nierozstrzygalnego dla danej maszyny nazywa się często „wygödlowywaniem" tej maszyny. Proces w którym konstruujemy takie zdania dla kolejnych maszyn, nazywa się natomiast „grą w wygödlowywanie" – grą, którą prowadzi antymechanicysta ze zwolennikiem sztucznej inteligencji, który twierdzi, iż możliwe jest skonstruowanie maszyny równoważnej człowiekowi. Krajewski pisze w tym kontekście, iż omawiany argument „to argument dialektyczny, czyli warunkowy: jeśli ktoś twierdzi, że jakaś maszyna jest równoważna umysłowi, to w odpowiedzi pokazuje się, że popada on w sprzeczność" [Krajewski, 2003, s. 129]. Można go uznać za argument typu *reductio ad absurdum*: zakładamy, iż pewna maszyna w pełni modeluje wszystkie ludzkie zachowania – nazwijmy taką maszynę-kandydatkę na doskonałą sztuczną inteligencję maszyną **F;** wtedy jednak konstruujemy dla takiej maszyny zdanie nierozstrzygalne, którego nie może ona dowieść, otrzymując sprzeczność z założeniem.

Zanim zaczniemy grać w grę w wygödlowywanie z antymechanicystą, musimy odpowiedzieć sobie jeszcze na jedno pytanie: czy nie jest tak, że my sami jesteśmy sprzeczni? A więc, czy do modelowania umysłu nie jest

przypadkiem potrzebna maszyna sprzeczna? (taką hipotezę postawił między innymi sam Turing). Oznaczałoby to jednak, iż taka maszyna **F** nie spełnia (wymienionych powyżej) założeń twierdzenia Gödla, a całe rozumowanie Lucasa nie mogłoby być przeprowadzone. Trudno powiedzieć, co dokładnie miałoby znaczyć stwierdzenie, iż jesteśmy sprzeczni. Z jednej strony wydaje się naturalnym, iż zmieniamy zdanie oraz że każdemu zdarza się w trakcie swego życia wypowiadać sprzeczne sądy. Z drugiej strony dla wielu filozofów – dla Lucasa, ale również dla Gödla – przyjęcie naszej sprzeczności oznaczałoby stwierdzenie, iż nie jesteśmy istotami racjonalnymi, na co żaden z nich nie chciał się zgodzić[15]. Rozważania, czy jesteśmy sprzeczni, wydają się pozostawać na poziomie spekulacji i być skazane na niekonkluzywność. Sformułujmy więc kolejne założenie naszego argumentu:

Z4. Jesteśmy niesprzeczni.

Aby wygödlować maszynę **F**, antymechanicysta musi pokonać kilka, jak się okaże, poważnych problemów natury praktycznej. Po pierwsze, aby móc skorzystać z twierdzenia Gödla i skonstruować zdanie nierozstrzygalne, musi wiedzieć, iż hipotetyczna maszyna **F** jest niesprzeczna (niesprzeczność jest jednym z założeń twierdzenia). Jak się okazuje, jest to w praktyce bardzo trudne do wykonania, matematycy mają bowiem trudności ze ścisłym wykazaniem niesprzeczności wielu teorii obecnie nam znanych, a teoria odpowiadająca maszynie **F** byłaby zapewne niewyobrażalnie skomplikowana. Dodatkowo (zwraca na to uwagę między innymi Chalmers w: Chalmers, 1995, pkt. 2.11–2.14) sprawdzanie niesprzeczności

[15] Dodam, iż istnieje tu jeszcze jedna możliwość – że jesteśmy *sprzecznymi maszynami*. Oznaczałoby to przyjęcie specyficznej postaci mechanicyzmu; jednocześnie jednak wyłączamy z dyskusji twierdzenie Gödla – które nie może takiej tezy ani udowodnić, ani odrzucić. Zauważmy, iż konstruuje się systemy aksjomatyczne tolerujące sprzeczności, nie jest też przesądzone, iż każda sprzeczna maszyna (cokolwiek miałoby to oznaczać) będzie musiała być wadliwa; programy komputerowe, w których są małe błędy, potrafią dobrze wypełniać swoje zadanie – podobnie jak nasze małe „sprzeczności" nie uniemożliwiają nam funkcjonowania. W kwestii problematyki sprzeczności ludzi i maszyn zob. Krajewski, 2003, s. 107–117.

maszyn jest w ogólności trudniejsze niż sprawdzanie niesprzeczności teorii formalnych. Stworzenie teorii formalnej odpowiadającej danej maszynie jest bowiem często wyjątkowo trudne; jest to oczywiście możliwe w teorii – mówią o tym odpowiednie twierdzenia – lista aksjomatów dla takiego systemu może się jednak okazać ogromna i bardzo skomplikowana, trudności mogą się również pojawić w przypadku reguł dowodzenia.

Kolejny problem natury praktycznej jest związany z podawaniem w trakcie gry w wygödlowywanie zdań Gödla dla coraz to bardziej skomplikowanych teorii. Douglas Hofstadter argumentuje tu, iż konstrukcja kolejnych zdań nierozstrzygalnych jest coraz trudniejsza wraz z komplikacją systemu, w szczególności, gdy jego wielkość osiąga kolejne nieskończone liczby porządkowe [Hofstadter, 1979, s. 476]. Wreszcie można podać w wątpliwość założenie, iż możliwa jest w ogóle sytuacja, w której *wiemy*, iż jakaś maszyna jest nam równoważna. To właśnie założenie o tym, że **F** ma tę własność, prowadzi poprzez *reductio ad absurdum* (na mocy twierdzenia Gödla) do wniosku, iż jednak nie może jej mieć.

Ta kwestia wydaje się być pomijana, zwraca na nią jednak uwagę między innymi David Chalmers w: Chalmers 1995. Oczywiście argument z twierdzenia Gödla jest warunkowy, twierdzi się tu, iż *jeśli* **F** jest równoważna człowiekowi, *to* zachodzi sprzeczność, jednak skoro omawiamy praktyczne problemy związane z wygödlowywaniem, należy również o takim problemie wspomnieć[16].

Antymechanicysta może stwierdzić, iż wyżej wymienione problemy są co najwyżej problemami natury praktycznej. Może on mianowicie się upierać, iż przynajmniej *w teorii*, czy *z zasady*, może zawsze wygödlować każdą maszynę. Idealizacja możliwości człowieka idzie tutaj już bardzo daleko, ale trzeba przyznać, iż argument nie został jeszcze z formalnego punktu widzenia obalony. Niestety, właśnie najsilniejsze – i, jak się wydaje, defi-

[16] Podałem tu tylko niektóre – jak się wydaje, najważniejsze – trudności związane z wygödlowywaniem. Można tu jeszcze wspomnieć o jednym problemie: nie ma powodu, aby sądzić, że mamy w stosunku do maszyn jakąś wyróżnioną pozycję w grze w wygödlowywanie. Skoro tworzenie zdania nierozstrzygalnego dla danej teorii jest czysto automatyczne, komputer może również grać w nią z nami! Co gorsza, skoro maszyny znacznie wyprzedzają nas w zdolnościach obliczeniowych, w pewnym momencie to one będą miały w tej grze przewagę [zob. Krajewski, 2003, s. 132–133].

nitywne – argumenty przeciwko antymechanicyście, który uważa, że może zawsze wygödlować maszynę, mają charakter czysto formalny. Otóż, po pierwsze, jest matematycznym faktem, iż zbiór maszyn niesprzecznych (to znaczy odpowiadających niesprzecznym systemom formalnym) nie jest rekurencyjny, to znaczy nie istnieje efektywna procedura odróżniająca maszyny niesprzeczne od sprzecznych. Wygödlowanie maszyny **F** wymagałoby od nas, abyśmy byli niemechaniczni, ale to właśnie chcieliśmy przez wygödlowanie wykazać [zob. Murawski, 1999, s. 329]. Ta matematyczna obserwacja prowadzi do jeszcze głębszego wniosku. Okazuje się mianowicie, iż antymechanicysta, na przykład sam Lucas, twierdząc, iż dla każdej podanej maszyny może (nawet w teorii) podać zdanie dla niej nierozstrzygalne, popada w sprzeczność[17]. Konkludując, akt wygödlowywania jest w praktyce prawie niemożliwy, a w ogólności nie jest możliwy do wykonania bez posiadania umiejętności, które same zakładają, iż nie jesteśmy maszyną. Tym samym antymechanicysta nie obala nawet najogólniejszej tezy **T1** – przynajmniej o ile dowód swój opiera na dialektycznej procedurze wygödlowywania.

Argument Penrose'a

Ponad dwadzieścia lat temu angielski fizyki i matematyk Roger Penrose sformułował „gödlowski" argument przeciwko mechanicyzmowi w nieco odświeżony sposób. Uczynił to najpierw w książce *Nowy umysł cesarza*, a następnie w o wiele bardziej systematyczny sposób w *Cieniach umysłu,* której to książce spróbuję się poniżej dokładniej przyjrzeć. Penrose wzbudził swoimi książkami sporą dyskusję – trzeba przyznać, iż głównie krytyczną – ale chyba jednak ciekawą i stymulującą. Wielu komentatorów, jak Krajewski, uważa, iż argument Penrose'a nic nie wnosi. Mimo to

[17] Dokładniej można dowieść, iż jeśli procedura wygödlowywania zostanie określona formalnie jako funkcja określona na zbiorze kolejnych maszyn Turinga (dokładniej na liczbach naturalnych odpowiadających tym maszynom w ciągu, co można uczynić efektywnie), a której wartościami są zdania nierozstrzygalne to wśród tych zdań nierozstrzygalnych pojawią się zdania wzajemnie sprzeczne [zob. Krajewski, 2003, s. 144–146].

wydaje się wprowadzać do dyskusji pewne nowe elementy, kładzie też nacisk na nieco inne kwestie niż Lucas. Formułuje również swoje wnioski nieco ostrożniej niż Lucas, warto więc się rozumowaniu Penrose'a przyjrzeć.

Rozważania Penrose'a skupiają się wokół pojęcia świadomości. Jego wersja argumentu „gödlowskiego" ma pokazać, iż twierdzenie Gödla pociąga fakt, że „korzystając ze świadomości, możemy podejmować działania, które przekraczają granice jakiejkolwiek obliczalnej aktywności" [Penrose, 2000, s. 9). Przykładem aktywności umysłowej, która wykracza poza wszelkie obliczenia, ma być przy tym wgląd matematyczny, traktowany jako szczególny typ rozumienia. Rozumienie matematyczne ma być przy tym nieredukowalne do obliczeń nie z powodu takich czy innych trudności praktycznych, ale *z zasady*, fakt ten ma być udowodniony natomiast za pomocą twierdzenia Gödla. Penrose podkreśla, iż jego rozumowanie obala nawet słabą tezę AI, to znaczy ma z niego wynikać, iż pewnych (świadomych) procesów zachodzących w mózgu nie można nawet *symulować* komputerowo [zob. Penrose, 2000, s. 31][18].

Penrose korzysta ze sformułowania twierdzenia Gödla wyrażonego w terminach maszyn Turinga. Proponuje rozważyć algorytm **A**, który formalizuje wszystkie aktywności matematyków, to znaczy ujmuje wszelkie procedury, z których korzystają, aby dowodzić twierdzeń matematycznych. Angielski matematyk ma tu na myśli algorytm, który „obejmuje wszystkie procedury, z których mogą skorzystać matematycy w celu dowiedzenia, że obliczenia się nie zakończą" [Penrose, 2000, s. 102]. Każdy problem matematyczny można bowiem – według Penrose'a – sformułować w postaci

[18] Znamienne jest to, iż Penrose jest przekonany, że świadomość można badać naukowo, nie wystarczają do tego jednak metody obliczalne (a więc te, które możemy wyrazić za pomocą maszyn Turinga). Określenie, czym miałby ogólnie być nieobliczalny (to znaczy taki, którego nie można symulować maszyna Turinga) proces na poziomie świata fizycznego, napotyka pewne trudności natury zarówno fizycznej, jak i matematycznej. Penrose przyjmuje ogólne założenie, iż dany proces lub daną procedurę uważamy za obliczalną, jeśli można ją symulować za pomocą komputera, czyli rodzaju maszyny Turinga. Z tego też powodu układy chaotyczne, oraz procedury podające przybliżenia traktowane są jako obliczalne; również sam fakt, iż używamy jako aparatury pojęciowej liczb rzeczywistych oraz różnych metod niedeterministycznych, nie świadczy o tym, iż opisywane tymi metodami procesy nie są obliczalne.

stwierdzenia: „Algorytm *X* zakończy (lub nie) pracę”[19]. O algorytmie **A** zakładamy dodatkowo, iż jest *poprawny* (co oddaje się też terminem – adekwatny), to znaczy jeśli stwierdza, iż dany algorytm *K* kończy pracę – to tak istotnie jest[20].

Rozważmy dalej, powiada Penrose, zbiór wszystkich algorytmów, to znaczy zbiór wszystkich maszyn Turinga; zbiór taki można efektywnie ustawić w ciąg indeksowany liczbami naturalnymi. Algorytm **A** „działa” na tych wszystkich algorytmach w ten sposób, iż stwierdza, czy kończą pracę, czy nie (jako algorytm nam równoważny powinien dostarczać informacji o wszelkich algorytmach – procedurach związanych z matematycznym rozumowaniem). Kilka pomysłowych zabiegów pozwala wskazać na algorytm – nazwijmy go **C,** o którym z *konstrukcji rozumowania wiemy*, że nie zakończy pracy, nasz algorytm **A** nie jest jednak w stanie tego stwierdzić[21]. My wiemy więc coś, czego **A** nie jest w stanie wykazać. Stąd **A** nie może być formalizacją procesów myślowych zachodzących w umyśle matematyka. Finalna teza, czy wniosek z tych rozważań, jest nieco oszczędniejsza i precyzyjniejsza niż wersja wniosku proponowanego przez Lucasa; Penrose nazywa ją tezą **G** [Penrose, 2000, s. 105]:

> **G**: W celu wykazania prawdy matematycznej matematycy nie posługują się algorytmami, o których wiedzą, że są poprawne.

---

[19] Na przykład zamiast powiedzieć: „Każda liczba naturalna da się przedstawić jako suma kwadratów czterech liczb naturalnych” możemy powiedzieć: „Algorytm, który ma dane następujące zadanie: ‘Znajdź liczbę naturalną, która nie jest sumą czterech liczb kwadratowych’, nigdy nie skończy pracy”. Twierdzenie o takiej właśnie treści udowodnił w 1770 r. matematyk francuski Joseph Lagrange [zob. Penrose, 2000, s. 95].

[20] Formalniej, oznacza to, iż **A** jest adekwatny semantycznie (*sound*). System formalny jest adekwatny semantycznie, gdy dowodzi tylko prawdziwych twierdzeń, to znaczy jeśli istnieje jakiś dowód (syntaktyczny) dla danej formuły, to formuła ta jest prawdziwa (w rozumieniu semantycznym).

[21] Algorytmem tym jest – w pewnym uproszczeniu – sam algorytm **A** zastosowany do liczby odpowiadającej pozycji tego algorytmu w ciągu algorytmów, który utworzyliśmy. Widać tu pewne analogie z oryginalnym sformułowaniem twierdzenia Gödla. Wtedy widzieliśmy (z konstrukcji dowodu), iż dane twierdzenie jest prawdziwe, ale jednocześnie wykazaliśmy, iż nie posiada ono dowodu. Tu wiemy, iż algorytm nie zakończy pracy, tymczasem nie jest to algorytmicznie stwierdzalne.

Tezę tę angielski fizyk uważa za niepodważalny wniosek z twierdzenia Gödla w swoim sformułowaniu oraz stawia ją w centrum swoich dalszych rozważań. Warto dodać, iż wydaje się ona słuszna wielu filozofom, na przykład Chalmersowi, czy nawet ostremu krytykowi Penrose'a Krajewskiemu [Krajewski, 2003, s. 155].

To jednak jeszcze nie koniec argumentacji Penrose'a – teza o możliwości istnienia sztucznej inteligencji nie została jeszcze bowiem obalona. Angielski fizyk przedstawia dalej następujące rozumowanie typu *reductio ad absurdum*: załóżmy, powiada, że istnieje taka maszyna (odpowiadająca algorytmowi **A**), która jest równoważna naszym zdolnościom matematycznym – oznaczymy ją znów przez **F**. Istnieją wtedy trzy możliwości – **F** może być:

I. świadomie poznawalna, przy czym jej rola jako rzeczywistego algorytmu będącego podstawą matematycznego rozumienia jest również poznawalna.

II. świadomie poznawalna, ale jej rola jako rzeczywistego algorytmu będącego podstawą matematycznego rozumienia jest nieświadoma i niepoznawalna;

III. nieświadoma i niepoznawalna [Penrose, 2000, s. 171][22].

Penrose uważa, iż żadna z tych tez nie jest to utrzymania i w konsekwencji **F** nie może istnieć. Możliwość (I) jest, według Penrose'a, sfalsyfikowana przez tezę **G** (zgodnie z nią nie możemy twierdzić, iż jesteśmy równoważni pewnemu algorytmowi jednocześnie wiedząc, iż jest on poprawny). Penrose odrzuca również III. Największe problemy powstają przy odrzuceniu II; angielski filozof tutaj sporo pomniejszych argumentów, podam za Chalmersem skróconą wersję tego rozumowania. Według Chalmersa Penrose rozumuje w następujący sposób:

[22] W nieco bardziej zwięzłym ujęciu Krajewskiego tezy te brzmią następująco: „I. F jest nam znana i wiemy, że jest nam równoważna. II. F jest nam znana, ale nie wiemy, że jest nam równoważna. III. F nie jest nam znana" [Krajewski, 2003, s. 154].

1) wiemy, iż jesteśmy adekwatni (to znaczy sami nie dowodzimy fałszywych twierdzeń);

2) wiemy, iż F jest nam równoważny (to znaczy jest podstawą naszego rozumienia matematycznego);

3) a więc wiemy, iż F jest adekwatny [zob. Chalmers, 1995, s. 2].

Przyjęcie (2) pociąga odrzucenie II. Możliwość zajścia (2) jest chyba jednak najbardziej dyskusyjna. Penrose sam tu przyznaje, że teza II jest bardzo mało prawdopodobna, ale nie ma ściśle logicznego sposobu wykazania jej fałszywości[23]. Nie będę tu dalej analizował rozumowania Penrose'a. Wielu komentatorów – w tym Krajewski – uważa w każdym razie, iż stosują się do niego wszystkie krytyki argumentu Lucasa [zob. Krajewski, 2003, s. 155–156][24].

Jak rozważania Penrose'a mają się do argumentacji Lucasa? Używa on nieco innej terminologii – w samym sformułowaniu rozumowania za tezą **G** mówi o algorytmach, a nie o systemach formalnych. Ponadto zamiast pojęcia niesprzeczności używa pojęcia adekwatności semantycznej. Nie są one równoznaczne, ale można chyba powiedzieć, iż pełnią w obu argumentach tę samą funkcję (i rodzą te same problemy). Skąd możemy na przykład wiedzieć, iż algorytm **A** jest adekwatny? Co więcej, skąd pewność, iż *my* jesteśmy adekwatni (to znaczy dowodzimy tylko prawdziwych twierdzeń)? Oczywiście różnic jest więcej, ale trzeba przyznać, iż te modyfikacje pozwoliły Penrose'owi sformułować swój argument w subtelniejszy i uważniejszy sposób niż Lucas. Argument Penrose'a napotyka jednak na wiele problemów związanych również ze sformułowaniem Lucasa. Penrose z jednej strony nie szuka tej wyjątkowej umiejętności umysłu w poznaniu prawdziwości zdania nierozstrzygalnego. Jednak – podobnie jak Lucas – nie jest on w stanie stwierdzić, która dokładnie czynność naszego umysłu

[23] Hilary Putnam uważał, iż ten zestaw trzech możliwości podany przez Penrose'a nie jest wyczerpujący, mianowicie, iż może jeszcze zachodzić możliwość (IV). Może mianowicie istnieć program, który możemy zapisać, ale nie będziemy w stanie go w pełni świadomie zanalizować. Byłby on więc znany, ale nie rozumiany. Penrose odpowiedział na ten zarzut, iż ta możliwość zasadniczo podpada pod możliwość (III) [Krajewski, 2003, s. 157].

[24] Krajewski formułuje również specyficzną dla Penrose'owskiego argumentu wersję twierdzenia o sprzeczności Lucasa.

jest nieobliczalna. Zrozumienie faktu, iż algorytm **C** nie kończy pracy, można przecież traktować jako procedurę mechaniczną tak samo, jak poznanie prawdziwości zdania nierozstrzygalnego. Penrose nie obala więc tezy **T2**. Co więcej, choć unika on wielu problemów związanych z wygödlowywaniem, wydaje się nie podawać żadnych definitywnych argumentów przeciw tezie **T1**.

Warto również zwrócić uwagę na inne różnice między dwoma typami rozumowania.

Penrose nie zamierza na przykład argumentować za mentalizmem. Łączy za to silnie swoją wersję argumentu przeciw mechanicyzmowi z platonizmem matematycznym. Uważa, że twierdzenie Gödla wspiera tezę platonizmu oraz że rozumienie, czy (nieobliczalny) wgląd matematyczny, za pomocą którego dowiadujemy się o prawdziwości zdań nierozstrzygalnych, jest właśnie rodzajem wglądu w platoński świat idei. Penrose, niestety, nie próbuje precyzować stanowiska platonizmu, filozof matematyki Richard Tieszen pisze, iż w analizowanej tu książce „nie jest zbyt jasne, jak powinniśmy rozumieć platonizm Penrose'a i jego rolę w argumencie przeciwko mechanicyzmowi" [Tieszen, 2005, s. 219][25]. Dodajmy na koniec, iż włączenie w dyskusję pojęcia świadomości wydaje się być chybione. Stwierdzenia typu „jestem świadomy, że X" można w rozważaniach Penrose'a zastąpić stwierdzeniami „wiem, że X". Angielski matematyk być może wzbogaca dyskusję o pewien wymiar epistemologiczny, nie wydaje się jednak, aby cokolwiek interesującego powiedział o świadomości jako takiej[26].

Chalmers uważa, iż w trzeciej części książki Penrose'a ukryty jest „nowy", niezauważony przez wielu (włączając w to chyba samego Penrose'a), argument przeciwko mechanicyzmowi. Podkreśla on najpierw, iż autor *Cieni umysłów* przyjmuje w „głównym argumencie" dwa założenia – po pierwsze to, że wiemy (umiemy pokazać), iż **F** jest adekwatna; po dru-

[25] Zwróćmy uwagę, iż w argumencie Lucasa platonizm nie grał właściwie żadnej roli.

[26] Co ciekawe, Penrose podkreśla, iż nie argumentuje przeciwko determinizmowi; według angielskiego fizyka fakt, iż pewne procesy są nieobliczalne, nie obala tej tezy. Relacja determinizmu i mechanicyzmu jest materiałem na odrębną i jak się wydaje, ciekawą, dyskusję.

gie to, że wiemy, iż jesteśmy równoważni **F**. „Nowy" argument – według Chalmersa – nie wymaga przyjęcia tych założeń. Argument ten był krytykowany przez Krajewskiego, niezwykle wnikliwie (oraz w sposób nieco bardziej przychylny) analizuje go również Shapiro w: Shapiro 2003. Analiza tego rozumowania jest jednak materiałem na odrębną dyskusję.

Mechanicyzm a poglądy Kurta Gödla

Bardzo interesujące filozoficzne wnioski z swoich twierdzeń wyciągał sam Gödel. W pracy *Some Basic Theorems on the Foundations of Mathematics and Their Implications* Gödel w przenikliwy sposób połączył swoje twierdzenia z problematyką mechanicyzmu oraz zagadnieniami związanymi z filozofią matematyki[27]. Z twierdzeń tych skorzystał przy tym w zupełnie inny sposób niż Penrose i Lucas. Posługując się nim, sformułował alternatywę (zwaną Alternatywą Gödla), którą uważał za ściśle matematyczny wniosek z tego twierdzenia, a która miała służyć do obalenia tezy mechanicyzmu. Gödel nie sądził przy tym, iż samo jego twierdzenie wystarcza do obalenia mechanicyzmu – są do tego potrzebne dodatkowe założenia.

Gödel wprowadza do dyskusji pytania o samą matematykę; w szczególności pyta, które twierdzenia matematyczne są w ogóle poznawalne (zauważmy, że teza **T1'**, stwierdzająca coś o relacji zbiorów twierdzeń, które może udowodnić człowiek i komputer, nie bierze pod uwagę takiego pytania). Wprowadza w tym duchu rozróżnienie na matematykę obiektywną i subiektywną. Matematyka subiektywna składa się z tych twierdzeń, które będziemy mogli udowodnić kiedykolwiek i jakimikolwiek metodami, matematyka obiektywna natomiast z tych, które są obiektywnie prawdzi-

[27] Praca *Some Basic Theorems…* powstała na podstawie wykładu wygłoszonego przez Gödla w grudniu 1951 roku. Był on 25. z serii prestiżowych wykładów ku czci amerykańskiego matematyka Josiaha Willarda Gibbsa. Gödel z początku miał zamiar opublikować ten tekst, czego jednak nigdy nie zrobił. Podobny los spotkał zresztą szereg tekstów austriackiego logika, wiele z nich zostało opublikowanych i udostępnionych szerszej publiczności dopiero w 1995 roku w 5-tomowym zbiorze *Kurt Gödel. Collected Works* [zob. Feferman, 2006, s. 134–135]).

we, niezależnie od naszych zdolności poznawczych – zawiera ona więc, można powiedzieć, matematykę „samą w sobie" (założenie istnienia takich twierdzeń jest oczywiście wyrazem realizmu matematycznego).

Podczas gdy większość autorów wyciąga z pierwszego twierdzenia Gödla wniosek, iż całej matematyki (nie odnosząc się do tego podziału) nie da się zawrzeć w jednym systemie aksjomatycznym, Gödel odnosił ten wniosek tylko do matematyki obiektywnej. Austriacki logik nie wykluczał przy tym, iż całą matematykę subiektywną można zawrzeć w jednym takim systemie, czy – jak to w tym kontekście ujmuje – skończonej regule [zob. Krajewski 2003, s. 162; Gödel, 1951, s. 309]. Jakie miałoby to jednak konsekwencje? Tu Gödel wykorzystuje drugie ze swoich słynnych twierdzeń. Jest ono wnioskiem z pierwszego twierdzenia, ma jednak nieco inny wydźwięk filozoficzny. Drugie twierdzenie Gödla głosi, iż nie możemy udowodnić niesprzeczności żadnego systemu aksjomatycznego (zawierającego arytmetykę liczb naturalnych) „w ramach samego tego systemu", to znaczy jedynie za pomocą środków dostępnych w tym systemie (to znaczy w szczególności aksjomatów i reguł dowodzenia)[28]. Zgodnie z tym twierdzeniem, jeśli wszystkie nasze umiejętności matematyczne można ująć w postaci jednego systemu aksjomatycznego, to nie będziemy mogli formalnie wykazać, że ten system jest niesprzeczny. Tu dochodzimy do słynnej alternatywy Gödla:

> 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 [Gödel, 1951, s. 310, tłumaczenie za: Krajewski, 2003, s. 163][29].

---

[28] Zdanie „wyrażające" niesprzeczność systemu w nim samym okazuje się być jednym ze zdań nierozstrzygalnych, których istnienie wykazuje pierwsze twierdzenie Gödla.

[29] Wang podaje również prostsze sformułowanie tej alternatywy, bezpośrednio odnoszące się do pojęć matematyki subiektywnej i obiektywnej: „Albo matematyka subiektywna przewyższa możliwości wszystkich [każdego] komputerów, albo też matematyka obiektywna przekracza matematykę subiektywną" [Wang, 1997, s. 186, tłumaczenie za: Krajewski, 2003, s. 163]. Równoważność tych dwóch sformułowań okaże się jaśniejsza w dalszej części tekstu.

Przyjrzyjmy się bliżej jej treści. Oznaczmy pierwszy jej składnik przez A, drugi natomiast przez B. Najpierw należy bliżej objaśnić, co oznacza zdanie B. Otóż można wykazać, iż zdanie stwierdzające niesprzeczność danej teorii (w tym przypadku teorii nam równoważnej), można przedstawić w postaci pewnego równania diofantycznego [zob. Gödel, 1951, s. 308][30]. Tak więc założenie, iż jesteśmy niesprzeczną maszyną, prowadzi do wniosku, że istnieją absolutnie nierozwiązalne problemy matematyczne [zob. Feferman, 2006, s. 140].

Zauważmy dalej, iż A jest pewnym sformułowaniem tezy o niemechaniczności umysłu.

Głosi, iż kompetencje matematyczne człowieka zawsze przerastają te komputerów. Wynika stąd, iż ¬A pociąga B, a zgodnie z tym twierdzenie Gödla pociąga fałszywość następującej koniunkcji [zob. Tieszen, 2006, s. 230]:

> (**TG**) wszystkie równania diofantyczne są rozwiązalne (¬B) oraz jesteśmy równoważni (przynajmniej w zakresie matematyki subiektywnej) pewnej maszynie Turinga (¬A).

Dlatego też zachodzi Alternatywa Gödla – A i B nie mogą być jednocześnie odrzucone. Gödel nie wyklucza, iż oba człony alternatywy są prawdziwe, formułuje jednak swoją alternatywę jako dysjunkcję. Aby definitywnie odrzucić mechanicyzm, argumentuje on przeciwko tezie B[31].

Drugi składnik alternatywy, a więc stwierdzenie, iż istnieją absolutnie nierozwiązalne równania diofantyczne, Gödel odrzuca, powołując się na swój optymizm poznawczy. Austriacki logik pyta tu mianowicie, dlaczego miałyby istnieć problemy matematyczne, które możemy sformułować, a których nie możemy *z zasady,* a więc w żaden możliwy sposób rozwią-

---

[30] Równania diofantyczne dotyczą wyłącznie liczb całkowitych – a więc obiektów matematycznych które są najbliższe naszej intuicji (w przeciwieństwie do na przykład przestrzeni 100-wymiarowych). Prostym przykładem takiego równania jest: $x^3 + y^2 = 5$.

[31] Alternatywa Gödla pozostawia możliwość jednoczesnego zachodzenia ¬A oraz B. Wykazanie ¬B pociągać więc będzie A, a więc tezę antymechanicyzmu [zob. Krajewski, 2003, s. 163–164].

zać [zob. Murawski, 1999, s. 333–334][32]. W szczególności jeśli są to problemy dotyczące pojęciowo łatwo nam dostępnych liczb całkowitych (o których mówią równania diofantyczne), a nie, na przykład, wyższych typów nieskończoności czy choćby przestrzeni wektorowych? Gödel odrzuca więc tezę B, a „teza antymechanicystyczna, którą tak bardzo chcą mieć Lucas i Penrose, wynika dla Gödla z optymizmu matematycznego" [Krajewski, 2003, s. 164][33].

Gödel wyciąga ze swojej alternatywy jeszcze inne, daleko idące wnioski. Uważa on mianowicie, iż przyjęcie każdego z dwóch jej członów zmusza nas do przyjęcia wniosków, które są „w zdecydowany sposób przeciwstawne wobec filozofii materialistycznej" [Gödel, 1951, s. 311]. Przyjęcie drugiego składnika alternatywy wspiera, według Gödla, obiektywność matematyki: skoro możemy sformułować problemy, których nie możemy (z zasady) rozwiązać, to musi być tak, iż nie *tworzymy*, nie *konstruujemy* matematyki. Skoro byśmy ją bowiem tworzyli, to jako twórcy musielibyśmy być w stanie (przynajmniej teoretycznie) dla każdego z jej zdań móc stwierdzić, czy jest prawdziwe. Tymczasem B pociąga istnienie zdań, które są prawdziwe, przy czym ich prawdziwość leży poza zasięgiem naszych umiejętności poznawczych. Z drugiego składnika alternatywy wyciąga Gödel jeszcze dalej idące wnioski, to znaczy odrzucenie materializmu. Czyni to jednak przez przyjęcie dodatkowego założenia: austriacki logik odróżnia mózg od umysłu, zakładając jednocześnie, iż mózg jako skończony układ fizyczny jest najprawdopodobniej maszyną. Ale zgodnie z pierwszym członem alternatywy mózg nie pozna wszystkich prawd ma-

[32] Zwróćmy uwagę, że byłaby to silniejsza nierozwiązywalność niż na przykład ta związana z „gödlowskimi" zdaniami nierozstrzygalnymi – ta druga bowiem jest relatywna do systemu oraz nie wyklucza zastosowań środków metamatematycznych, ta pierwsza z kolei wyklucza wszelkie możliwe metody. Feferman podaje pewne powody, dla których należy odrzucić istnienie problemów z zasady nierozwiązalnych; istnieje jednak według niego wiele zagadnień matematycznych, które są *w praktyce* (najprawdopodobniej) nie do rozwiązania [Feferman, 2003, s. 147–149].

[33] Zwróćmy uwagę, iż niezależnie od tego, czy przyjmiemy optymizm poznawczy Gödla, jego alternatywa ma pewne znaczenie dla filozofii matematyki. Jak pisze Tieszen: jeśli zaakceptować tezę **TG**, „nie możemy być mechanicystami w odniesieniu do umysłu, […] i jednocześnie nieograniczonymi optymistami w kwestii rozwiązywalności matematycznych problemów" [Tieszen, 2006, s. 233].

tematycznych, a więc musi je według Gödla poznawać umysł, który miałby wtedy jakąś umiejętność wychodzenia poza skończone umiejętności tego pierwszego [Gödel, 1951, s. 311][34]. Ten wniosek jest chyba najbardziej dyskusyjny (warto też zauważyć, iż Gödel argumentuje tu bardzo podobnie do Lucasa).

Jeśli chodzi o naturę poznania matematycznego w ogóle, Gödel nie sądził, aby jego twierdzenie przesądzało o jego niemechanicznej naturze (jak sugerowali Lucas i Penrose). Potwierdzało ono jedynie jego pogląd na matematykę, który żywił niezależnie od tegoż twierdzenia[35]. Gödel nie doszukiwał się więc niemechaniczności akurat w naturze poznania zdania nierozstrzygalnego, czy postrzeganiu, iż pewien algorytm się nie zatrzymuje – te akty nie mają wyróżnionego statusu. Mimo to był przekonany, iż poznanie matematyczne w ogóle dokonuje się (w pewnej przynajmniej części) za pomocą niemechanicznie funkcjonującego umysłu. Zdania matematyczne są prawdziwe niezależnie od nas, matematyka ma obiektywnie istniejący przedmiot, a kontakt z nim umożliwia nam pewna szczególna kompetencja poznawcza, którą nazywał Gödel intuicją[36]. Przy tym wszystkim Gödel kładł nacisk na niewyczerpywalność matematyki, rolę intuicji w ciągłej analizie pojęć matematycznych; sądził on, iż „nasza intuicja pod-

[34] Gödel niejednokrotnie wyrażał pogląd o istnieniu świata niematerialnego i o istnieniu umysłu oderwanego od ciała oraz od mózgu, żywił go również niezależnie od swoich twierdzeń. Uważał, iż przekonanie, że nie istnieje umysł oderwany od ciała, jest przesądem naszych czasów, który zostanie kiedyś naukowo obalony [zob. Murawski, 2000, s. 172]. W rozmowie z Wangiem wyraził nawet pogląd, iż „mózg to komputer podłączony do umysłu" [Wang, 1997, s. 196]. Zauważmy, iż Gödel jest tutaj w swoich poglądach o wiele odważniejszy niż na przykład Penrose

[35] Gödel wielokrotnie wyrażał swój sprzeciw wobec mechanicyzmu. Na przykład w jednej z notatek pochodzących z lat trzydziestych i najprawdopodobniej przygotowanych do wykładu, stwierdza on: „mechanizacja matematycznego rozumowania nie jest możliwa, to znaczy nigdy nie będzie możliwym zastąpienie matematyka przez maszynę, nawet jeśli ograniczymy się do problemów teorioliczbowych" [za: Murawski, 2002, s. 110]. Austriacki logik szukał różnych sposobów na obalenie mechanicyzmu. Tieszen, powołują się na rozmowę z Wangiem, wspomina, iż do obalenia poglądu, iż umysł ludzki jest maszyną, Gödel chciał użyć na przykład idei zawartych w fenomenologii Husserla [Tieszen, 2006, s. 231].

[36] Poglądy Gödla na matematykę opisane są obszernie między innymi w: Wójtowicz, 2002.

lega rozwojowi, a rozwój ten dokonuje się dzięki analizie pojęć i uprawianiu matematyki" [Wójtowicz, 2002, s. 63].

Twierdzenie Gödla ma według jego autora potwierdzać, iż w rozwijającym się poznaniu matematycznym musimy odwoływać się właśnie do intuicji i znaczeń pojęć matematycznych. Skoro żaden system aksjomatyczny nie ujmie w pełni pojęcia liczby, do jego (między innymi) poznania potrzebne są wglądy bazujące na czymś więcej niż na skończonych, kombinatorycznych własnościach symboli, mianowicie na refleksji nad *znaczeniem*. Konieczność istnienia takiego rodzaju intuicji wywodzi z twierdzenia Gödla również Tieszen. Uważa on, iż dla dowodu na przykład niesprzeczności *PA* będziemy musieli użyć pojęć z „wyższego" poziomu (*higher level concepts*), to znaczy pojęć, które nie są finitystyczne, które nie odnoszą się do skończonych układów jakichś obiektów umiejscowionych w czasoprzestrzeni. Te pojęcia będą musiały dotyczyć znaczeń, odwoływać się do intuicji matematycznej. Nawiązując dalej do Gödla, pisze, iż poznanie pojęcia liczby pozwala nam na szukanie wciąż nowych aksjomatów dla *PA*. Jeśli dalej wyobrazimy sobie – pisze Tieszen – proces dodawania do *PA* nowych aksjomatów, które są kolejnymi zdaniami nierozstrzygalnymi, i tworzenia w ten sposób kolejnych rozszerzeń *PA*, to przez taki ciąg nowych teorii „przeprowadza" nas właśnie intuicyjne poznanie pojęcia liczby; to z niego – według Tieszena – korzystamy, tworząc kolejne teorie należące do takiego ciągu [zob. Tieszen, 2005, s. 220–222][37].

Jak porównać rozumowania Gödla do wcześniejszych rozumowań Lucasa i Penrose'a?

Wymieńmy tylko kilka różnic. Po pierwsze, Gödel analizuje głównie konsekwencje drugiego ze swoich twierdzeń. Dodatkowo, formułując swoją alternatywę, łączy rozważania nad mechanicyzmem z kwestią obiektywności matematyki w sposób na pewno subtelniejszy i głębszy niż Penrose. Jeśli chodzi o kwestie epistemologiczne, Gödel nie szuka niemechaniczności w procesie poznania prawdziwości zdania nierozstrzygalnego. Jednak,

[37] Przypomnijmy, iż podobną konsekwencję twierdzenia Gödla dla poznania matematycznego podkreślał Penrsoe. Pisał na przykład, iż „reguły mogą czasami być substytutem rozumienia, ale nigdy nie mogą go całkowicie zastąpić" [Penrose, 2000, s. 101].

podobnie jak Lucas i Penrose, przekonany jest o pewnym niemechanicznym pierwiastku w naszym poznaniu matematycznym, próbuje go również jakoś powiązać ze swoim twierdzeniem. Za Lucasem – a w przeciwieństwie do Penrose'a – Gödel argumentuje wreszcie za mentalizmem.

Podsumowanie: co bezpośrednio wynika z twierdzenia Gödla?

Wydaje się, iż twierdzenie Gödla nie obala w sposób definitywny ani tezy **T1**, ani tezy **T2** – przynajmniej, jeśli zastosujemy metody opisane w powyższym tekście. Jakie wnioski można z niego wywieść w sposób bezpośredni? Drugie twierdzenie Gödla pokazuje, iż jeśli jesteśmy maszyną (niesprzeczną), to nie możemy wykazać swojej niesprzeczności. Można również pokazać, iż jeśli jesteśmy równoważni pewnej konkretnej maszynie, to nie możemy odnaleźć kodu tej maszyny. W słowach Krajewskiego więc „twierdzenie Gödla nie wyklucza, że człowiek może być maszyną, ale wtedy nie możemy widzieć jaką; innymi słowy, nie jest wtedy ona dla niego dostatecznie zrozumiała, przejrzysta" [Krajewski, 2003, s. 274]. Wygląda również na to, iż „nie ma algorytmu nam równoważnego, świadomie poznawalnego (co oznacza, że działanie algorytmu da się zrozumieć) i niesprzecznego" [Krajewski, 2003, s. 274–275]. Wydaje się, iż jest to wniosek zgodny z tezą **G** Penrose'a[38]. Krajewski zaznacza jednak, iż to „nie przesądza nic na temat naszej mechaniczności czy niemechaniczności" [Krajewski, 2003, s. 265][39].

---

[38] Tezę **G** można przeformułować w następujący sposób: „jeśli istniałby algorytm formalizujący nasze umiejętności matematyczne, to nie moglibyśmy wiedzieć, iż jest poprawy". Jest to sformułowanie bardzo bliskie jednej z uwag Gödla. Pisze on, iż nie jest logicznie możliwe, by ktoś mógł dla danego systemu aksjomatycznego twierdzić, iż postrzega (z matematyczną pewnością) jego reguły i aksjomaty jako prawdziwe i jednocześnie twierdzić, iż system ten zawiera w sobie całą matematykę [Gödel, 1951, s. 309].

[39] Paragraf ten można skrócić następująco w słowach samego Gödla: „twierdzenia o niezupełności nie wykluczają możliwości, iż istnieje komputer dowodzący twierdzenia, który jest w istocie równoważny intuicji matematycznej. Implikują one jednak, że w takim – z innych przyczyn wysoce nieprawdopodobnym – przypadku, albo nie znamy dokładnej

Wyciągnięcie jakichkolwiek definitywnych wniosków co do mechanicznej natury umysłu jest z pewnością utrudnione przez dwie kwestie: po pierwsze, niejasność samej tezy mechanicyzmu; po drugie, kontrowersje co do sposobu zastosowania twierdzenia Gödla, a ogólniej – trudności w zestawianiu pojęć matematycznych i filozoficznych. Shapiro twierdzi, iż „nie dysponujemy żadną przekonującą (*plausible*) tezą mechanicyzmu, która byłaby wystarczająco ścisła, aby mogła być zakwestionowana przez twierdzenia o niezupełności" (za: Feferman, 2006, s. 147). Jeszcze bardziej pesymistyczny jest tu Lindström, który twierdzi, iż „może się okazać, że nasze pytanie [to znaczy o mechanicyzm] nie jest dobrze zdefiniowane, i w związku z tym nie ma dobrze określonej (*well-defined*) „odpowiedzi" [Lindström, 2001, s. 249].

Mimo to uważam, iż badania nad „gödlowskim" argumentem przyczyniły się do pogłębienia rozumienia problemu mechanicyzmu. Wydaje się, iż spór ten był dla wielu filozofów inspirujący – autor niezwykle inspirującej, ponad 700-stronicowej książki o twierdzeniu Gödla, komputerach i logice, Douglas Hofstadter (i jednocześnie jeden z pierwszych krytyków argumentu Lucasa) pisał o omawianym tu argumencie co następuje: „był [on] jednym z głównych czynników (*forces*), który zmusił mnie do przemyśleń nad zagadnieniami omawianymi w tej książce [*Gödel, Escher, Bach*]" [Hofstadter, 1979, s. 472]. Można więc chyba za Krajewskim przyjąć, iż „rozstrzygnięcie tego pytania [to znaczy 'czy umysł jest maszyną?'] przez użycie twierdzenia Gödla nie jest możliwe, ale rozważanie go w świetle tego twierdzenia – owszem" [Krajewski, 2003, s. 285–286][40].

---

specyfikacji takiego komputera, albo nie wiemy, że działa on poprawnie" [Wang, 1997, s. 186].

[40] Warto wspomnieć, iż twierdzenie Gödla próbowano zastosować do zagadnień mechanicyzmu również na inne sposoby. Bringsjord oraz Arkoudas sformułowali parę lat temu nowy, *modalny* argument „gödlowski" przeciwko tezie, iż umysł jest maszyną Turinga – a za tezą, że jest hipermaszyną! (o hipermaszynach, czyli maszynach niebędących maszynami Turinga, wspominałem w przypisie nr 10). Wykorzystując powyższe twierdzenie, konkludują oni, iż: *jeśli* możliwe jest, że umysłu są hiperkomputerami, to faktycznie nimi są (a nie – zwykłymi komputerami!); zob. Bringsjord, Arkoudas 2004, s. 169. Wydaje się, iż świadczy to po raz kolejny o uniwersalności i „elastyczności" rozumowania bazującego na twierdzeniu Gödla.

Inną kwestią jest to, iż nawet jeśli twierdzenie Gödla ma jakieś znaczenie dla zagadnienia mechaniczności umysłu, nie ma ono żadnego praktycznego wymiaru. Eugeniusz Szumakowicz i Ewa Bryła twierdzą, iż w przypadku omawianego argumentu „mamy generalnie do czynienia z przeidealizowanym porównywaniem przeidealizowanego umysłu z przeidealizowanym komputerem (maszyną Turinga) w uprawianiu przeidealizowanej (sformalizowanej) matematyki" [Szumakowicz, Bryła, 2004, s. 33–34]. Nawet jeśli jest to zbyt ostra ocena omawianego tu argumentu (jest on w końcu argumentem filozoficznym!), to najprawdopodobniej analizy jemu poświęcone nie będą miały faktycznie większego wpływu na badania nad sztuczną inteligencją.

Bibliografia

Barrow J., (1996), *II razy drzwi*, Warszawa, Prószyński i S-ka.

Bringsjord S., Arkoudas K., (2004), „The modal argument for hypercomputing minds", *Theoretical Computer Science*, 317, s. 167–190.

Chalmers D.J., (1995), „Minds, machines, and mathematics. A review of shadows of the mind by Roger Penrose", *Psyche*, 2(9); http://psyche.cs.monash.edu.au.

Copeland B.J., (1998), „Turing's *O*-machines, Searle, Penrose and the brain", *Analysis*, 58.2, s. 128–138.

Copeland B.J., (2004) „Hypercomputation: philosophical issues", *Theoretical Computer Science*, 317, s. 251–267.

Feferman S., (1995), „Penrose's Gödelian argument. A review of shadows of the mind by Roger Penrose", *Psyche*, 2(7); http://psyche.cs.monash.edu.au.

Feferman S., (2006), „Are there absolutely unsolvable problems? Gödels's dichotomy", *Philosophia Mathematica* (3), Vol. 14, s. 134–152.

Gödel K., (1951), „Some basic theorems on the foundations of mathematics", [w:] *Kurt Gödel. Collected Works.* Vol. III, [ed.] Feferman S., New York, Oxford, Oxford University Press, 1995, s. 304–323.

Hofstadter D., (1979), *Gödel, Escher, Bach: an eternal golden braid*, New York, Basic Books, Inc. Publishers.

Krajewski S., (2003), *Twierdzenie Gödla i jego interpretacje filozoficzne. Od mechanicyzmu do postmodernizmu*, Warszawa, Wydawnictwo Instytutu Filozofii i Socjologii PAN.

Lucas J.R., (1961), „Minds, machines and Gödel", *Philosophy* 36, s. 120–124.

Lindström P., (2001), „Penrose's new argument", *Journal of Philosophical Logic*, Vol. 30, s. 241–250.
McCullough D., (1995), „Can humans escape Gödel? A review of shadows of the mind by Roger Penrose"*, Psyche*, 2(4); http://psyche.cs.monash.edu.au.
Murawski R., (1997), „Gödel's incompleteness theorems and computer science", *Foundations of Science* 2 (1997), s. 123–135.
Murawski R., (1999), *Recursive functions and metamathematics: problems of completeness and decidability, Gödel's Theorems*, Dodrecht–Boston–London, Kluwer Academic Publishers.
Murawski R., (2000), *Funkcje rekurencyjne i elementy metamatematyki*, wyd. 3, Poznań, Wydawnictwo Naukowe UAM.
Murawski R., (2002), „Truth vs. provability – philosophical and historical remarks", *Logic and Logical Philosophy*, Vol. 10, s. 93–117.
Penrose R., (2000), *Cienie umysłu*, Poznań, Zysk i S-ka.
Putnam H., 1986, „Minds and machines", [w:] *Philosophical Papers,* Vol. 2. *Mind, Language and Reality*, Cambridge, Cambridge University Press, s. 362–385.
Searle J.R., (1999), *Umysł na nowo odkryty*, Warszawa, Państwowy Instytut Wydawniczy.
Shapiro S., (2003), „Mechanism, truth, and Penrose's new argument", *Journal of Philosophical Logic* 32, s. 19–42.
Szumakowicz E., Bryła E., (2004), „Niealgorytmiczność myślenia", *Zagadnienia Filozoficzne w Nauce*, XXXV, s. 25–44.
Tieszen R., (2005), „Penrose on minds and machines", [w:] *Phenomenology, Logic and the Philosophy of Mathematics,* Cambridge, Cambridge University Press, s. 315––324.
Tieszen R., (2006), „After Gödel: mechanism, reason, and realism in the philosophy of mathematics", *Philosophia Mathematica* (3), Vol. 14, s. 229–254.
Wang H., (1997), *A Logical Journey. From Gödel to Philosophy*, Cambridge,The MIT Press.
Wójtowicz K., (2002), *Platonizm matematyczny*, Warszawa, Wydawnictwo Diecezji Tarnowskiej Biblos.

Gödel's theorem and the debate on mechanism

ABSTRACT. This paper discusses and summarizes the main aspects of the debate on the implications of Gödel's incompleteness theorems for mechanism. First, the general line of argument against mechanism is presented based on the incompleteness theorems. Secondly, three specific strategies of argumentation are outlined, i.e. those proposed by John R. Lucas, Roger Penrose, and Kurt Gödel himself. Each of these strands of

thought is analyzed by taking special care to bring out the underlying philosophical, logical and mathematical assumptions as well as to clarify the understanding of basic concepts such as those of ‘mind’, ‘machine’ or ‘mechanism’, and finally, to precisely formulate the conclusions that can be drawn from different arguments. The paper concludes with a brief assessment of the validity and philosophical significance of Gödel’s theorem-based arguments against mechanism and suggestions on what can in fact be concluded from this theorem as to the relation between the human mind and machines.

KEY WORDS: philosophy of mathematics, philosophy of mind

Michał Sochański, Wyższa Szkoła Uni-Terra, ul. Prądzyńskiego 53, 61-527 Poznań