Obliczanie jest tylko interpretowalną manipulacją symbolami; poznanie nią nie jest*

Stevan Harnad

14 kwietnia 2013

University of Southampton

Streszczenie

Obliczanie jest interpretowalną manipulacją symbolami. Symbole są przedmiotami, którymi się manipuluje na podstawie ich kształtów, które to kształty są arbitralne w odniesieniu do tego, co może być zinterpretowane jako ich znaczenie. Nawet jeśli się akceptuje tezę Turinga/Churcha mówiącą, że obliczanie jest wyjątkowe, uniwersalne i bardzo bliskie wszechmocności, nie wszystko jest komputerem, ponieważ nie wszystkiemu można nadać systematyczną interpretację; a z pewnością nie wszystkiemu można nadać każdą systematyczną interpretację. Ale nawet po odróżnieniu komputerów i obliczania od innych rodzajów rzeczy, stany umysłowe nie będą jedynie realizacjami prawidłowych systemów symboli – z powodu problemu ugruntowania symboli: interpretacja systemu symboli nie jest wewnętrzną częścią systemu; jest ona nadana (narzucona z zewnątrz) systemowi przez interpretatora. Nie jest to prawdą, natomiast, o naszych myślach. Musimy, w związku z tym, być czymś więcej niż komputerami. Moim domysłem jest to, że znaczenia naszych symboli są ugruntowane w podłożu (*substrate*) naszych zdolności robotycznych do interakcji z realnym światem przedmiotów, zdarzeń i stanów rzeczy, a nasze symbole są interpretowalne jako ich dotyczące.

Słowa kluczowe: Przyczynowość, poznanie, obliczanie, świadomość, ciągłość, realizacja (implementation), robotyka, przetwarzanie sensomotoryczne, semantyka, systemy symboli, składnia, maszyna Turinga, test Turinga

Ojcowie nowoczesnej teorii obliczeniowej (Church, Turing, Gödel, Post, von Neumann) byli matematykami i logikami. Nie popełnili oni błędu i nie uznawali się za psychologów. Ich następcy potrzebowali paru więcej dziesięcioleci aby zacząć mylić obliczanie z poznaniem (Fodor 1975, Newell 1980, Pylyshyn 1984, Dietrich 1990). Zanim będziemy mogli uporządkować to pomieszanie, musimy najpierw dojść do porozumienia, czym jest obliczanie (dużo trudniej będzie się zgodzić na to, czym jest poznanie, ale będąc kartezjańskimi podmiotami poznawczymi, przyjmijmy ostensywną definicję poznania). Pozwólcie mi od razu oznajmić, że podpisuję się pod tym, co przyjęło się nazywać tezą Turinga - Churcha (Church/Turing Thesis, CTT) (Church 1956), która jest oparta na zbieżnym dowodzie, że wszystkie niezależne próby formalizacji tego, co matematycy rozumieją przez "obliczanie" lub "efektywną procedurę", nawet jeśli zewnętrznie wyglądały na różne, okazywały się równoważne (Galton 1990).

Systemy manipulacji symbolami i teza Turinga-Churcha Zgodnie ze wszystkimi notacyjnymi odmianami tego, co moglibyśmy słusznie nazwać Maszynami Turinga, obliczanie jest po prostu manipulacją egzemplarzami symboli [*symbol tokens*] opartą

*Minds and Machines vol.4, nr 4, 1994

na kształtach tych symboli (Turing 1990). Najczęstszym obrazkiem jest jakaś maszyna z taśmą zapisaną symbolami; głowica czytająca może przesuwać taśmę do przodu lub do tyłu, może ją również zatrzymać; potrafi także czytać, wypisywać lub zastępować symbole. Maszyna ta jest zbudowana w taki sposób, że to co robi (czyta, zapisuje, porusza, zatrzymuje) jest wyznaczone przez stan w jakim się aktualnie znajduje, w jakie inne stany może przechodzić oraz jaki symbol jest właśnie czytany (na przykład, stan ten może być czymś, co zachowuje się zgodnie z regułą "czytaj, jeśli odczytywany jest symbol 1, przesuń taśmę w lewo i powróć do tego stanu; jeśli czytanym symbolem jest 0, zatrzymaj taśmę"). Wszystko, co matematycy, logicy i informatycy zdołali dotąd osiągnąć za pośrednictwem wynikania logicznego, obliczeń i dowodu może być wykonane przez tą maszynę z pomocą takich elementarnych operacji, jak wyżej opisane. Mówię "dotąd", ponieważ pozostaje wciąż otwarte pytanie, czy ludzie mogą "obliczać" rzeczy, które nie są obliczalne w sensie formalnym: jeśli by mogli, wówczas CTT byłaby fałszywa. Z tego powodu teza ta nie jest twierdzeniem, podlegającym dowodowi, ale indukcyjnym przypuszczeniem wspieranym przez oczywistość; oczywistość ta jednak dotyczy raczej własności formalnych (składniowych), niż fizycznych czy empirycznych. Istnieje naturalne uogólnienie CTT do systemów fizycznych (CTTP). Według CTTP, wszystko, co może zrobić dyskretny, fizyczny system (czy też wszystko, co ciągły fizyczny system może - w takim przybliżeniu, w jakim sobie życzymy - wykonać) można osiągnąć przez obliczanie. CTTP występuje w dwóch odmianach: jako słaba i mocna CTTP, w zależności od tego, czy teza ta mówi, że wszystkie fizyczne systemy są formalnie równoważne komputerom, czy też głosi, że są one po prostu komputerami. Nie jest to szczególnie istotne do czasu, kiedy przejdziemy poniżej do własności niezależności realizacji [*implementation-independence*], a wtedy okaże się, że rozróżnienie to podnosi rzeczywiste problemy komputacjonalizmu (tezy mówiącej, że poznanie jest tylko rodzajem obliczenia, C=C - Cognition=Computation), tezy, która także występuje w słabej i mocnej wersji. Na razie jednak wystarczy jeśli będziemy brali pod uwagę tylko jedną, nie zróżnicowaną CTTP. Są ludzie wątpiący w prawdziwość CTTP i są także tacy, którzy wątpią zarówno w prawdziwość CTT, jak i CTTP. Dla takich ludzi obliczanie jest albo czymś, czego jeszcze nie udało się sformalizować, albo też jest czymś nieformalizowalnym. Oczywiście to, co się potwierdza, lub czemu się zaprzecza w hipotezach wyższego poziomu łączących obliczanie z poznaniem będzie odmienne w pojęciu tych, którzy akceptują i tych, którzy nie akceptują CTT. Ci, którzy tezę tą odrzucają, są dużo bardziej skłonni, na przykład, do zaprzeczania komputacjonalizmowi. W związku z tym chcę zaznaczyć, że akceptuję tezę Turinga-Churcha, zarówno w jej formalnej, jak i fizycznej wersji (CTT i CTTP), jednakże również będę argumentował przeciwko C=C.

Systematyczna interpretowalność Wciąż są dwa istotne składniki definicji obliczania, które opuściłem. Pierwszy z nich jest kontrowersyjny i zazwyczaj nie traktuje się go jako część tej definicji: formalne obliczanie jest czystą manipulacją symbolami, z operacjami na symbolach (czytanie, wypisywanie, przesuwanie, zatrzymanie) opartymi, jak już stwierdziłem, na kształtach tych symboli. Takie operacje w oparciu o kształty są zazwyczaj nazywane "syntaktycznymi", w odróżnieniu od "semantycznych" operacji, które byłyby oparte raczej na znaczeniach symboli, niż na ich kształtach. Znaczenie nie wchodzi w skład definicji formalnego obliczania. Przypomnij sobie, że kiedy po raz pierwszy uczono cię arytmetyki (czy algebry lub teorii zbiorów) formalnie, symbole (w arytmetyce "0", "1", "+", "=", etc.) wprowadzano jako "pierwotne, nie definiowane terminy" - chociaż oczywiście już intuicyjnie i praktycznie wiedziałeś co one znaczą - a następnie poznałeś reguły łączenia tych terminów ze sobą w poprawne formuły i wyprowadzania z nich innych poprawnych formuł za pomocą zasad logiki i dowodu. Wszystkie te reguły

były zasadniczo syntaktyczne. W żadnym miejscu nie użyto znaczenia symbolu do uzasadnienia tego, co wolno z tymi symbolami robić. Jednak, chociaż o tym nie wspominano, istota uczenia się formalnej matematyki (czy logiki lub programowania) polega na tym, że wszystkie te manipulacje symbolami są w jakiś sposób znaczące ("+" faktycznie zgadza się z tym co mamy na myśli dodając rzeczy do siebie, a "=" rzeczywiście odpowiada temu, co rozumiemy przez równość). To nie była jedynie pozbawiona znaczenia gra syntaktyczna. Zatem, chociaż zazwyczaj pozostaje to nie stwierdzone, jest wciąż kryterialną, jeśli nie definicyjną cechą obliczania to, że manipulacje symbolami muszą być semantycznie interpretowalne - i to nie tylko lokalnie, ale globalnie: wszystkie te interpretacje symboli i manipulacji muszą sobie wzajemnie odpowiadać, jak to ma miejsce w arytmetyce, na poziomie symboli indywidualnych, formuł i ciągów formuł. Wszystkie te interpretacje muszą mieć systematyczny sens, zarówno w całości, jak i w części (Fodor i Pylyshyn 1988). To kryterium semantycznej interpretowalności (przezwane "zniewoleniem deszyfrantów" ["*cryptographers constraint*"], ponieważ wymaga, aby system symboli był dekodowalny tak, aby miał systematyczny sens) nieczęsto się spotyka: łatwo jest zebrać garść dowolnych symboli i sformułować arbitralne, ale systematyczne reguły manipulacji tymi symbolami, nie zapewnia to jednak istnienia jakiegokolwiek sposobu sensownej interpretacji (Harnad 1994a).

Obliczanie: trywialne versus nietrywialne Ta sytuacja jest w jakiś sposób analogiczna do przysłowiowych małp przy maszynie do pisania: zazwyczaj przywołujemy ten obrazek aby rozważyć prawdopodobieństwo przypadkowego napisania przez te małpy ustępu z Szekspira. Ale równie dobrze możemy sobie wyobrazić te małpy systematycznie wypisujące niezliczone, następujące po sobie, tworzące regularności reguły: jakie jest prawdopodobieństwo, że któryś z takich syntaktycznych systemów będzie dekodowalny w sensowny sposób? (Zauważ, że nie pytam jak moglibyśmy je zdekodować, ale czy w ogóle jakaś mająca znaczenie interpretacja jest możliwa, niezależnie od tego, czy moglibyśmy taką interpretację znaleźć). Innymi słowy, zbiór systematycznie interpretowalnych formalnych systemów symboli jest z pewnością znacznie mniejszy aniżeli zbiór formalnych systemów symboli, i jeżeli wytwarzanie nieinterpretowalnych systemów symboli jest obliczaniem w ogóle, z pewnością lepiej opisuje je określenie trywialne obliczanie, podczas gdy ten rodzaj obliczania, którym jesteśmy zainteresowani (czy jesteśmy matematykami, czy też psychologami), jest obliczaniem nietrywialnym (nontrivial computation): mianowicie rodzaj, który ma systematyczny sens. System symboli jedynie z dwoma stanami: "0" i "1" interpretowany odpowiednio: "życie jest jak obważanek" i "życie nie jest jak obważanek", jest trywialnym systemem symboli. Arytmetyka i język angielski są systemami nietrywialnymi. Systemy trywialne mają niezliczone arbitralne "dwójki": można zamieniać ze sobą wzajemnie interpretacje ich symboli i wciąż myśleć o spójnej semantyce (np.: zamienić ze sobą wzajemnie obważanek i nie-obważanek w przykładzie powyżej). Nietrywialne systemy symboli zasadniczo nie mają niesprzecznie interpretowalnych "dwójek", a jeśli mają, są one kilkoma ściśle określonymi, dowodliwymi wyjątkami (jak zamienność koniunkcji\negacji i dysjunkcji\negacji w klasycznym rachunku zdań). Generalnie nie można jednak dowolnie zamieniać interpretacji w arytmetyce, języku angielskim, czy LISP-ie, i dalej oczekiwać, że system będzie w stanie dźwigać ciężar systematycznej interpretowalności (Harnad 1994a). Spróbujmy na przykład w języku angielskim zamienić interpretacje prawdy vs. fałszu czy nawet słów czerwony i zielony, nie wspominając już o takich funktorach, jak "jeśli" i "nie": zbiór wyrażeń angielskich nie będzie niesprzecznie interpretowalny po takiej arbitralnej, niestandardowej interpretacji; aby go takim uczynić trzeba by było zmienić interpretację każdego symbolu, dostosowując te zmiany systematycznie do pierwszej zamiany interpretacji. To właśnie

ta sztywność i wyjątkowość systemu w odniesieniu do standardowej , "zamierzonej" interpretacji, decydowała będzie, jak sądzę, o rozróżnieniu pomiędzy nietrywialnymi a trywialnymi systemami symboli. Podejrzewam również, że różnica pomiędzy nimi będzie raczej typu "wszystko albo nic", aniżeli kwestią stopnia. Komputer, w takim razie, będzie fizyczną realizacją (implementation) jakiegoś systemu symboli - systemu dynamicznego, którego stany i sekwencje stanów są interpretowalnymi obiektami (podczas, gdy w statycznym formalnym systemie symboli te obiekty są po prostu znakami (rysunkami) na papierze). Uniwersalna Maszyna Turinga jest abstrakcyjną idealizacją klasy realizacji systemów symboli; komputer cyfrowy jest konkretnym fizycznym urzeczywistnieniem tego systemu symboli. Myślę, że ściana, na przykład, jest jedynie realizacją trywialnego obliczania, i skutkiem tego - jeśli omawiane rozróżnienie: trywialny/nietrywialny można formalnie wyliczyć (opracować) - ścianę można wyłączyć z klasy komputerów (lub włączyć jedynie jako trywialny komputer).

Niezależność [od] realizacji [*Implementation Independence*] Jesteśmy zatem zainteresowani jedynie nietrywialnym obliczaniem. Oznacza to symbole, na których operacje przeprowadza się jedynie na podstawie ich kształtu, ale niemniej poddające się systematycznej interpretacji. Innymi słowy chodzi o znaczące systemy symboli. Tego kryterium sensowności (posiadania znaczenia, *meaningfulness*) potrzebujemy nie tylko po to, aby skupić uwagę na nietrywialnym obliczaniu, ale także, aby wskazać na drugą znaczącą własność obliczania, o której już wspominałem: Kształty egzemplarzy symboli muszą być arbitralne. Arbitralne w odniesieniu do czego? W odniesieniu do tego, co można zinterpretować jako ich znaczenie. W formalnej arytmetyce Peano, na przykład, symbolem równości "=" operuje się jedynie na podstawie jego kształtu, a jego kształt nie wiąże się z żadnym fizycznym odniesieniem do własności "równości" - ani nie przypomina jej z wyglądu, ani też nie jest z tą własnością przyczynowo powiązana. To samo jest prawdziwe o "3", która ani nie przypomina "troistości" (threeness) ani nie jest przyczynowo z nią powiązana w świecie. W tym miejscu należy zwrócić uwagę na fakt, że podobnie symbole naturalnego języka mają tą własność arbitralności w odniesieniu do tego, co znaczą (co Saussure nazwał "*l'arbitraire du signe*"). Czymkolwiek jest naturalny język, jest z pewnością także formalnym systemem symboli, ponieważ ma on w rzeczywistości składnię, która generuje wszystkie i tylko jego poprawne wyrażenia, i te wyrażenia są faktycznie systematycznie interpretowalne, jako znaczące to, co znaczą. Inne pytanie, jakie możemy zadać w odniesieniu do "małp przy maszynie do pisania" dotyczy prawdopodobieństwa tego, że mogłyby one przez przypadek odkryć syntaktyczne reguły, które generowałyby wszystkie i tylko takie łańcuchy symboli, które byłyby systematycznie interpretowalne jako wyrażenia w jakimś naturalnym języku. Ponieważ wszystkie formalne języki (jak arytmetyka, rachunek zdań czy LISP) są podzbiorami jakiegoś języka naturalnego, to czy prawdopodobieństwo tego, że wygenerowany system symboli jest systematycznie interpretowalny jak i język naturalny (lub język myśli, Fodor 1975) także kryje się w definicji formalnego obliczania? Stoimy tutaj przed niebezpieczeństwem pomieszania obliczania z poznaniem, któremu chcieliśmy zaradzić, zatem zaznaczmy, że kryterium interpretowalności-przez-poznającego nie może być bardziej niepożądane w teorii obliczeniowej, aniżeli jest kryterium obserwowalności-przez-poznającego w teorii kwantowej (i oczywiście bez jakiegokolwiek dziedziczenia paradoksów tej ostatniej): potrzebujemy jedynie odnotować, że pośród systemów symboli rzadkimi są te, które wykonują nietrywialne obliczenia. Możemy potrzebować pomyślnej ludzkiej interpretacji, aby dowieść, że dany system rzeczywiście wykonuje nietrywialne obliczenia, jest to jednak jedynie kwestia epistemiczna. Jeśli, w oczach Boga, istnieje potencjalna systematyczna interpretacja, wówczas ten system dokonuje obliczeń nietrywialnych, niezależnie

od tego, czy człowiek kiedykolwiek znajdzie taką interpretację, czy nie. Wracając jednak do kwestii arbitralności kształtów - potrzebujemy semantyki, aby ją uszczegółowić: trywialne byłoby stwierdzenie, że każdy przedmiot, zdarzenie i stan rzeczy jest obliczeniowy ponieważ może być systematycznie interpretowany jako będący swoim własnym symbolicznym opisem: kot na macie może być interpretowany jako oznaczający kota na macie, z kotem będącym symbolem kota, matą będącą symbolem maty i ich przestrzennym zestawieniem jako symbolem bycia na. Dlaczego nie jest to obliczanie? Ponieważ kształty symboli nie są arbitralne w stosunku do tego, co według interpretacji jest ich znaczeniem, faktycznie są one dokładnie tym, co według tej interpretacji znaczą. Wydawać się może ponadto, że jest to trywialny przypadek, ale różnicuje się on na więcej problematycznych sytuacji: tych, w których te "symbole" fizycznie przypominają to, co znaczą, czy też są z tym przyczynowo powiązane. Proste operacje wykonywane przez Maszynę Turinga nie są oparte na jakimkolwiek bezpośrednim związku pomiędzy symbolami i ich znaczeniami; gdyby były oparte, wówczas pociągałoby to za sobą niezliczone wyjątki i formalne niezmienniki [*invariants*], tak, że zatracilibyśmy własności tego interesującego i potężnego fenomenu obliczania. Innym sposobem charakteryzowania tej arbitralności kształtów symboli w formalnym systemie symboli jest "niezależność realizacji": używane kształty symboli można zastąpić całkowicie odmiennymi, a jeśli system dotychczas przeprowadzał obliczanie, to dalej by przeprowadzał to samo obliczanie, jeśli manipulacje tymi nowymi kształtami przeprowadzano by w oparciu o te same syntaktyczne reguły. Ta niezależność realizacji obliczania wykluczałaby szczególne przypadki w których specyficzne kształty symboli czy ich fizyczne związki z tym, co znaczą, grałyby przyczynową rolę w takim "obliczaniu". Wyobraźmy sobie maszynę która mogłaby obliczać 2+2 tylko dopóki byłaby z powodzeniem łączona z dwoma parami rzeczy; lub maszynę która mogłaby generować "0" , gdy nie ma danych wejściowych lub "1" gdy ma jedną daną wejściową i tak dalej. Siła komputacji pochodzi z faktu, że ani system notacji symboli, ani szczegóły budowy maszyny nie są związane z przeprowadzanym obliczaniem. Całkowicie odmienny sprzęt [*hardware*] używające całkowicie odmiennego oprogramowania może przeprowadzić dokładnie te same obliczenia. Tym, co się liczy są własności formalne, nie zaś fizyczne. Taka abstrakcja od szczegółów fizycznych jest częścią tego, co daje Uniwersalnej Maszynie Turinga moc do przeprowadzania jakiegokolwiek obliczenia.

Nieodróżnialność w sensie Turinga Wymieńmy to, czym okazało się obliczanie: niezależna od realizacji, systematycznie interpretowalna manipulacja symbolami. Tylko w ten sposób okazało się możliwym przeprowadzenie jakiejkolwiek kalkulacji, o jakiej matematycy mogli jedynie marzyć, a także spowodowanie, aby maszyny wykonywały wiele z innych inteligentnych czynności, które wcześniej jedynie ludzie (i zwierzęta) mogli wykonywać. Prawdopodobnie było w tych warunkach całkiem naturalnym wywnioskować, że ponieważ:

(1) nie wiemy w jaki sposób podmioty poznawcze poznają i ponieważ

(2) obliczanie może robić tak wiele rzeczy (wykonać tak liczne czynności), jakie tylko poznający mogą robić,

poznanie jest tylko formą obliczania (C=C). Ostatecznie, zgodnie z CTT obliczanie jest wyjątkowe i najwyraźniej wszechmocne; zgodnie zaś z CTTP, cokolwiek mogą zrobić systemy fizyczne, mogą to także zrobić komputery. Ten sposób myślenia spowodował najpierw, że komputacjonaliści uznali się za psychologów (i spowodował, że psychologowie stali się komputacjonalistami), i, empirycznie mówiąc, w tym czasie miało to sens. Ten rodzaj myślenia uzyskał wsparcie ze strony jednego z najbardziej niedostępnych problemów poznania, mianowicie kwestii świadomości (Harnad 1982, 1991; Nagel 1974, 1986). Wszystkie poczynione przez fizykalistów

próby rozwiązania problemu psychofizycznego - problemu jaki wszyscy mamy w wytłumaczeniu w jaki sposób stany umysłowe mogą być stanami fizycznymi - napotykał na niepokonane trudności w jednym punkcie: jakiekolwiek przyczynowe czy funkcjonalne wyjaśnienie systemu fizycznego jest zawsze równie zgodne z mentalną, jak i z nie-mentalną interpretacją (Harnad 1994b); interpretacja mentalna zawsze wydaje się w jakiś sposób niezależna od fizycznej, nawet pomimo to, że są one wzajemnie w sposób oczywisty skorelowane, ponieważ przyczynowość/funkcjonalność systemu zawsze sprawia wrażenie całkowicie zdolnej do wyjaśniania równie dobrze, faktycznie nieodróżnialnie (przyczynowo/funkcjonalnie mówiąc), z pomocą umysłowości czy też bez niej. Rozważmy różnicę pomiędzy pewnym przyczynowym/funkcjonalnym systemem, który przystosowawczo unika zniszczenia tkanki a innym systemem, który robi to samo, ale czyniąc to odczuwa ból/unika bólu: skłania to do mówienia o funkcjonalnej/przyczynowej roli bólu, ale jakąkolwiek funkcjonalną/przyczynową rolę się mu przypisze, można ją równie dobrze przypisać fizycznemu podłożu (*substrate*) bólu, a następnie równie dobrze można odjąć ból i odnosić się tylko do funkcjonalnej/przyczynowej roli jego fizycznego podłoża. I co jest prawdziwe lub nieprawdziwe o bólu, jest prawdziwe o przekonaniu, pożądaniu, faktycznie o wszystkich stanach umysłowych. Wszystkie one mają tego typu odniesienie do ich fizycznego podłoża. Otóż owa niezależność realizacji stanów obliczeniowych pasuje dobrze do omówionej cechy stanów umysłowych: jeśli poznanie jest formą obliczania, nic dziwnego, że mamy problemy z przyrównaniem umysłowości [*the mental*] do fizyczności [*the physical*]: mieliśmy również problemy z przyrównaniem obliczeniowości i fizyczności, ponieważ obliczanie jest niezależne od jego fizycznej realizacji (Pylyshyn 1984). Także komputacjonalizm (C=C) wyciągnął pewne nieodparte wnioski z własności semantycznej interpretowalności: jeżeli już raz odkryjesz, że pewien system jest systematycznie interpretowalny, ciężko jest [później] dostrzec to w sposób zdeinterpretowany (jak to było w przypadku "niedefiniowalnego" "+" w formalnej arytmetyce). To jest jak próba ponownego usłyszenia obcego języka, kiedy już poznałeś sposób w jaki on brzmi, gdy jeszcze język ten był pozbawiony znaczenia dla ciebie (nie jest to łatwe!). Otóż, jeśli ta interpretacja jest mentalistyczna a nie jedynie semantyczna, staje się ona nawet jeszcze bardziej nieodparta, zawsze się potwierdzająca (z definicji, na mocy systematycznej semantycznej interpretowalności); Nazwałem to "hermeneutycznym gabinetem luster" (Harnad 1990b,c, Hayes i in. 1992). Alan Turing (niektórych czytelników kusi, aby przytaczać go jako przykład potwierdzający moją wcześniejszą sugestię, że żaden z ojców komputacjonalizmu błędnie nie uznawał się za psychologa) może być postrzegany, jako ktoś zachęcający nas do wstąpienia w ten hermeneutyczny krąg w swojej obronie słynnego Testu Turinga (Turing 1964). Według Turinga, gdyby była "osoba", której nie moglibyśmy w żaden sposób odróżnić od innego człowieka (powiedzmy przez długie lata), wówczas nie mamy nie-arbitralnej podstawy do wyciągnięcia wniosku, że "osoba" ta nie myśli, jeśli by nas poinformowano, że jest ona maszyną. To może zostać zinterpretowane jako zachęta do ulegnięcia pokusie hermeneutycznej mocy systematycznej interpretowalności – w szczególności dlatego, że aby wykluczyć złudzenie oparte na wyglądzie, test Turinga został sformułowany z wykorzystaniem istoty, z którą komunikujemy się jedynie przy pomocy symboli [*pen-pal*]. można jednak doszukać się u Turinga głębszego sensu, aniżeli wyżej wymienionego i zinterpretować go jako dowodzącego, że jedynie głupiec mógłby się ośmielić próbować rozróżnić pomiędzy czymś funkcjonalnie nieodróżnialnym.

Problem ugruntowania symboli [*Symbol grounding problem*] Zatem widzę Turinga jako orędownika poglądu mówiącego, że to raczej maszyny w ogóle mają funkcjonalne zdolności nieodróżnialne od naszych, aniżeli komputery i obliczanie w szczególności. Są jednakże tacy, którzy interpretują Test Turinga jako wsparcie dla

C=C. Dowodzą oni: poznanie jest obliczaniem. Zrealizuj [*implement*] prawidłowy system symboli, system, który zda test Turinga [*pen-pal test*] (w ciągu życia) - i będziesz miał zrealizowaną [*implemented*] myśl. Niestety zwolennicy tego poglądu muszą stawić czoła słynnemu Argumentowi Chińskiego Pokoju Searle'a (1980), w którym wskazuje on, że jakakolwiek osoba może zająć miejsce takiego testowanego komputera, realizując dokładnie ten sam system symboli, bez zrozumienia ani jednego słowa z danych wprowadzanych z klawiatury. Ponieważ obliczanie jest niezależne od realizacji, jest to dowód przeciwko jakiemukolwiek zrozumieniu ze strony komputera realizującego ten sam system symboli. Jednak, jak sugerowałem, argument Searle'a faktycznie nie kwestionuje Testu Turinga (Harnad 1989); on po prostu atakuje czysto symboliczną wersję tego testu, którą nazwałem T2. Pozostawia natomiast nietkniętą wersję robotyczną (T3) - która wymaga nierozróżnialnych w sensie Turinga własności symbolicznych i sensomotorycznych (jak i nie podważa T4: nierozróżnialności symbolicznej, sensomotorycznej i neuromolekularnej). Zatem jedynie czysto symboliczny system ulega argumentowi Searle'a i nie jest trudno dostrzec dlaczego. Nazwałem tę przyczynę "**Problemem Ugruntowania Symboli**" (Harnad 1990a, 1993c): nikt nie wie czym jest poznanie, ale wiemy, że podmioty poznawcze to robią. Systemy symboli mają zauważalną własność bycia zdolnymi do obliczania czegokolwiek, co jest obliczalne (to jest CTT). Pod tym względem, to, co mogą one zrobić i to, co mogą zrobić ludzie wydaje się biec wzdłuż tych samych kanałów. Ale pod jednym podstawowym względem się między sobą różnią: łańcuchy symboli jak "2+2=4" czy "kot jest na macie", generowane przez system symboli, są przykładami nietrywialnego obliczania, jeśli są one systematycznie interpretowalne jako znaczące to, co "2+2=4" i "kot jest na macie" znaczą. Ale to znaczenie, jak wcześniej stwierdzono, nie zawiera się w systemie symboli. Taki system jest po prostu syntaktyczny, manipulujący symbolami pozbawionymi znaczeń na podstawie reguł opartych na kształtach tych symboli, kształty symboli zaś są arbitralne w odniesieniu do tego, co mają zgodnie z interpretacją znaczyć. Szukanie znaczenia w takim systemie jest analogiczne do szukania znaczenia w chińsko/chińskim słowniku, kiedy się w ogóle nie zna chińskiego. Wszystkie słowa są tutaj w pełni zdefiniowane; wszystko jest systematyczne i spójne. Jednakże poszukując pewnego hasła, wszystkim co się znajduje jest ciąg nic nie znaczących symboli w formie definicji i jeżeli poszukuje się z kolei każdego ze słów definiujących, znajduje się więcej takich samych ciągów symboli. Takie poszukiwanie jest nieugruntowane. System taki jest systematycznie interpretowalny dla kogoś, kto zna choć trochę chiński, ale w sobie i dla siebie samego jest on pozbawiony znaczenia i prowadzi jedynie do nieskończonego regresu. Otóż tutaj właśnie pojawia się ta zasadnicza różnica pomiędzy obliczaniem i poznaniem: nie mam żadnego pojęcia, czym są moje myśli, ale jest jedno, co mogą z całą pewnością o nich powiedzieć: są one myślami o czymś, mają one znaczenie, i nie dotyczą one tego, czego dotyczą jedynie dlatego, że są one systematycznie interpretowalne przez ciebie jako dotyczące tego, czego dotyczą. Dotyczą one tego autonomicznie i wprost, bez jakiegokolwiek pośrednictwa. Skutkiem tego problem ugruntowania symboli jest problemem połączenia symboli z tym, czego one dotyczą bez jakiegokolwiek pośrednictwa zewnętrznej interpretacji (Harnad 1992a, 1993a). Rozwiązaniem, które się samo nasuwa jest to, że T2 potrzebuje ugruntowania w T3: własności symboliczne muszą być ugruntowane we własnościach robotycznych. Wiele sceptycznych rzeczy można powiedzieć o robocie, który jest w sensie T3 nierozróżnialny od osoby (w tym to, że może mu brakować myśli), ale nie można powiedzieć, że wewnętrzne symbole tego robota dotyczą przedmiotów, zdarzeń i stanów rzeczy, do których się odnoszą tylko dlatego, że są w taki sposób przeze mnie interpretowane, ponieważ ten robot sam może i faktycznie oddziałuje, autonomicznie i wprost na te przedmioty, zdarzenia i stany rzeczy (jak również one na niego) w sposób, który odpowiada interpretacji. Robot wypowiada "kot" w obecności kota, tak, jak i my to robimy,

"mata" w obecności maty etc. I wszystko to w takiej skali, która jest całkowicie nieodróżnialna od sposobu,w jaki my to robimy, nie tylko w odniesieniu do kotów i mat, ale w odniesieniu do wszystkiego, obecnego i nieobecnego, konkretnego i abstrakcyjnego. Gwarantuje to T3, tak jak T2 gwarantuje to, że symboliczna komunikacja z komputerem będzie systematycznie spójna. Ceną jednak, jaką trzeba zapłacić za ugruntowanie systemu jest to, że nie jest on już jedynie obliczeniowy! Dla robotycznego ugruntowania niezbędne jest przynajmniej przetwarzanie sensomotoryczne [*sensorimotor transduction*], a przetwarzanie nie jest obliczaniem.

Poznanie: wirtualne vs. realne A może jest? Przywołajmy powtórnie wprowadzone uprzednio dwie wersje tezy Turinga: Pierwsza z nich, CTT, miała czysto formalny charakter oraz druga fizyczna, CTTP, a ta ostatnia występowała w wersjach słabej i mocnej. Nie ma problemu ze słabą CTT, ponieważ stwierdza ona jedynie że każdy fizyczny system jest formalnie równoważny maszynie Turinga, a to byłoby prawdą równie dobrze o przetworniku, jak i o samolocie, piecu czy Układzie Słonecznym. Każde z nich może być symulowane stan-po-stanie przez komputer w takim przybliżeniu, w jakim chcemy. Ale nikt nie pomyliłby symulacji z realną rzeczą (być może z wyjątkiem osoby w symulacji wirtualnej rzeczywistości, która, jak sobie czytelnik, mam nadzieję, zdaje sprawę, zupełnie nie pasuje do omawianej kwestii, która nie dotyczy złudzeń obserwatora/interpretatora, ale realnego, niezależnego od interpretatora świata). Symulacja komputerowa optycznego przetwornika nie przetwarza realnego światła (co wymagałoby realnego przetwornika), przetwarza natomiast światło wirtualne, tj. światło symulowane przez komputer, i takie przetwarzanie jest samo w sobie wirtualne. To samo jest prawdą o wirtualnym samolocie, który realnie nie lata, ale po prostu symuluje lot w symulowanym samolocie. Na tej samej zasadzie wirtualny piec nie wytwarza realnego ciepła a wirtualny Układ Słoneczny nie ma realnego ruchu planetarnego czy grawitacji. Wszystkie te przypadki są całkowicie nieproblematyczne w ramach słabej CTTP. Nikogo nie kusi zaproponowanie "C=F" - tezy komputacjonalistów według której latanie jest tylko formą obliczania. Obliczanie jest tylko niezależną od realizacji, systematycznie interpretowalną manipulacją symbolami, niezależnie od tego czy te symbole są interpretowane jako samolot, piec, przetwornik, osoba z Testu Turinga, czy nawet jako robot. Wirtualny samolot w rzeczywistości nie lata, wirtualny piec w rzeczywistości nie grzeje, wirtualny przetwornik nie przetwarza; są one jedynie systemami symboli, które są systematycznie interpretowalne jako, odpowiednio, latanie, ogrzewanie i przetwarzanie. Wszystko to powinno być całkiem oczywiste. Odrobinę mniej oczywisty jest równie niezbity fakt, że wirtualna osoba z Testu Turinga nie myśli (czy rozumie lub pojmuje) - ponieważ jest on jedynie systemem symboli, systematycznie interpretowalnym jak gdyby był myślącym (rozumiejącym, pojmującym) Jeszcze mniej oczywistym (szczególnie z punktu widzenia poglądu, o którym już wspominano, dotyczącym ugruntowania robotycznego (T3) systemów symboli (T2)), ale wciąż niezbitym, i to z dokładnie z tych samych powodów jest to, że wirtualny robot nie myśli (nie bardziej niż się porusza, widzi czy dokonuje sensomotorycznego przetwarzania): wirtualny robot w wirtualnym świecie jest po prostu systemem symboli systematycznie interpretowalnym jak gdyby myślał (poruszał się, widział, przetwarzał). Prawdziwość słabej CTTP gwarantowałaby, że takie symulacje są możliwe, i silnie sugerowałaby, że takie komputerowe modelowanie jest doskonałym sposobem dochodzenia do zrozumienia fizycznych systemów (włączając w to samoloty, systemy słoneczne i piece - Searle nazwał to "słabą AI"), ale nie wspierałaby C=C. Dlaczego? ponieważ nawet gdyby symulowany robot T3 ujmowałby (tj. symulowałby) każdą stosowną własność poznania, wciąż byłby tylko nieugruntowanym systemem symboli. Aby go ugruntować, trzeba by było zbudować realnego T3 robota - a, mam nadzieję,

oczywistym jest, że nie byłoby to równoznaczne z prostym przyłączeniem sensomotorycznych przetworników do komputera wykonującego symulację (nie bardziej niż zbudowanie realnego samolotu czy pieca byłoby równoznaczne z prostym przyłączeniem sensomotorycznych przetworników do ich odpowiednich symulacji). Ale nawet jeśli [to] jest [równoznaczne], byłby to ugruntowany system symboli tylko jako całość - samolot, piec, T3 robot - to byłoby latanie, ogrzewanie i myślenie. Czysto symboliczny moduł nie byłby lataniem, ogrzewaniem ani myśleniem. Przeto C=C byłaby fałszywa.

Zróżnicowane równości i zależność od realizacji A co z mocną CTTP, zgodnie z którą samolot jest komputerem? Otóż albo jest ona błędna, albo jest nieinteresująca. Jedyną wartością poinformowania naukowców-kognitywistów, że poznanie jest raczej obliczaniem, niż jakimś innym fizycznym procesem, byłoby poinformowanie, że obliczanie jest pewnym naturalnym rodzajem pewnego rodzaju. Jeśli każdy fizyczny proces jest faktycznie obliczaniem, to kognitywiści mogliby równie dobrze zignorować wiadomość komputacjonalistów i dalej poszukiwać tego, czym mogą się okazać prawidłowe fizyczne procesy. Ale ja myślę, że mocna CTTP jest błędna, a nie pusta, ponieważ nie uwzględnia ona powszechnie ważnej niezależności od realizacji, która odróżnia obliczanie jako pewien naturalny rodzaj: latanie czy ogrzewanie, w odróżnieniu od obliczania, nie są niezależne od realizacji. Stosowną własnością niezmienną, jakie posiadają wszystkie rzeczy, które latają jest to, że wszystkie przestrzegają tych samych zbiorów rozmaitych równości, nie zaś to, że realizują ten sam system symboli (Harnad 1993a). Testem, mówiąc inaczej, jest próba ogrzania domu czy dostania się do Seattle z pomocą czegoś, co realizuje prawidłowy system symboli ale przestrzega złego zbioru różnorodnych równości. Tyle o mocnej CTTP. A co z C=C? Słaba wersja, zgodnie z którą poznanie może być symulowane - w takim przybliżeniu, w jakim się chce, przez obliczanie - jest w rzeczywistości wariantem słabej CTTP: Dlaczego niedualista miałby oczekiwać, że poznanie różniłoby się pod tym względem od jakichkolwiek innych fizycznych procesów (lot, ogrzewanie, grawitacja), wszystkich w podobny sposób symulowalnych przez obliczanie? Ale Mocne Stanowisko Obliczeniowe, według którego każda realizacja prawidłowego systemu symboli poznawałaby, jest także błędne w dokładnie ten sam sposób, w jaki mocna CTTP jest błędna (lub pusta), lub, jeśli zakłada mocną C=C, odrzucając jednocześnie mocną CTTP wówczas jest po prostu wykrętem dotyczącym nieobserwowalności poznania (własności, która oddziela poznanie od lotu, ogrzewania, ruchu, przetwarzania a nawet grawitacji). Co do poznania, definiowanego przez ostensję (z powodu braku naukowej teorii poznania), jest ono obserwowalne tylko przez umysł poznającego. Ta własność - druga strona problemu psychofizycznego - i znana skądinąd jako problem innych umysłów (Harnad 1991) wciągnęła mimowolnie mocnych komputacjonalistów w Hermeneutyczne koło. Miejmy nadzieję, że refleksja nad argumentem Searle'a i problemem ugruntowania symboli oraz w szczególności nad potencjalnymi empirycznymi drogami do dalszych rozwiązań (Andrews, Harnad 1987, Harnad i in. 1991, 1994), może pomóc mocnym komputacjonalistom jeszcze raz zacząć na nowo. Pierwszym krokiem może być próba próby deinterpretacji systemu symboli w arbitralne esy-floresy, jakimi system ten faktycznie jest (ale, podobnie jak oduczenie się języka, którego się już raz nauczyło, nie jest to łatwe!).

Przekład: *Piotr Konderak* (1999)