Paweł Grabarczyk

O niearbitralnym kryterium posiadania struktury obliczeniowej.[1]

1. Wstęp

Jest takie ironiczne określenie pracy filozofa - to jest człowiek, który zajmuje się wykazywaniem, że coś, co działa w praktyce, nie działa w teorii. Choć w wielu przypadkach jest ono zupełnie nieadekwatne i nawet nieco krzywdzące, to nie najgorzej pasuje do rozważań nad strukturami obliczeniowymi i kryteriami ich posiadania.

Na pierwszy rzut oka sprawa wydaje się jasna. Nie ulega wątpliwości, że istnieją przedmioty, które obliczenia realizują. Pisząc te słowa dotykam jednego z nich. Możemy spierać się o to, czy wolno nam przypisywać komputerom i robotom stany intencjonalne, wolną wolę albo emocje, ale nie będziemy się raczej spierać co do tego, czy są one w stanie realizować obliczenia. Z drugiej strony, niekontrowersyjne wydaje się także to, że są na świecie obiekty, które różnią się od komputerów pod względem zdolności obliczeniowych - ściany, wiadra, kamienie itd. Gdyby było inaczej, to nie musielibyśmy w ogóle budować komputerów, wystarczyłoby tylko w nowy sposób wykorzystać to, czym już dysponujemy. Wszystko wskazuje zatem na to, że istnieje coś, co przysługuje komputerom, ale nie przysługuje kamieniom a co ja przyjąłem tutaj nazywać strukturą obliczeniową.

Okazuje się jednak, że te zdroworozsądkowe konstatacje postawione zostały pod znakiem zapytania. Niezależnie od siebie, trzej filozofowie przedstawili argumenty za tym, że to, czy dany obiekt realizuje jakieś obliczenia czy nie (i, co za tym idzie, czy posiada strukturę obliczeniową), zależy tylko od pomysłowości obserwatora, który go opisuje. Ian Hinckfuss pokazał, że da się bez trudu przyjąć, że obliczeń dokonuje pewne wypełnione wodą wiadro,2 John Searle, że robi to jego ściana, a Hilary Putnam, nie tracąc czasu na poszukiwanie kolejnego spektakularnego kontrprzykładu, dowiódł, że robi to dowolny obiekt. Dwa z tych argumentów chciałbym poniżej szczegółowo omówić, ale zanim do tego przejdziemy, przedstawię jeszcze kilka uwag natury pojęciowej.

Przez strukturę obliczeniową będę rozumiał w tym artykule taką budowę obiektu, która pozwala na

[1] Za wiele cennych uwag krytycznych wobec tej i kilku poprzednich wersji artykułu chciałbym podziękować: Jakubowi Michalskiemu, Marcinowi Miłkowskiemu i Michałowi Zawidzkiemu.

[2] O propozycji Hinckfusa (która omówiona jest w Copeland, 1996. s. 336) wspominam jedynie ze względów historycznych, mimo pierwszeństwa, nie jest ona szeroko dyskutowana w literaturze, a nie różni się pod żadnymi, istotnymi dla nas względami od przykładu Searle'a

realizację obliczeń.[3] Obliczenia rozumiał zaś będę jako formalizmy, których własności dowodzi teoria obliczalności. Fakt, iż nie nakładam na nie żadnych dodatkowych ograniczeń sprawia, że uniwersalna maszyna Turinga jest jedynie jednym z możliwych przykładów takiego formalizmu.[4] Pojęcie *realizacji obliczenia* zdefiniować można teraz następująco:5

Mówimy, że obiekt fizyczny *realizuje pewne obliczenie*, gdy istnieje takie odwzorowanie *f* ze stanów fizycznych obiektu na stany formalne obliczenia, że gdy system jest w stanie fizycznym *p*, to przechodzi w taki stan fizyczny *q*, że formalny stan *f(p)* przechodzi w formalny stan *f(q)*.6

Ponieważ, jak zobaczymy poniżej, powyższa definicja realizacji obliczeń prowadzi do trywializacji tego pojęcia, to będę tak zdefiniowane pojęcie nazywał w dalszych akapitach „naiwnym pojęciem realizacji". Sposób, w jaki pojmuję arbitralność doprecyzowany zostanie w sekcji 2.

Będę bronił stanowiska, że posiadanie struktury obliczeniowej jest rzeczywistą, niearbitralną własnością, którą można wykryć w obiekcie metodami empirycznymi. Stanowisko to samo w sobie nie jest specjalnie zaskakujące - jest ono odzwierciedleniem potocznych intuicji, o których wspominałem. Problemem nie jest sam podział na obiekty posiadające struktury obliczeniowe i resztę. Jest nim jego nieodporność na kontrprzykłady w stylu wymienionych. Obrona będzie polegała na tym, że pokażę takie kryterium posiadania struktury obliczeniowej, które blokuje podane przykłady a także pewne ich modyfikacje, oraz nie jest za wąskie, czyli pasuje do tych obiektów, o których wiemy, że strukturę obliczeniową posiadają.

W sekcji 2 omówię argumenty Putnama i Searle'a. W sekcji 3 istniejące sposoby na uratowanie pojęcia realizacji obliczeń przed zagrażającą mu trywializacją oraz powody, dla których tak naprawione pojęcie nie nadaje się na interesujące mnie niearbitralne kryterium posiadania struktury obliczeniowej. W sekcji 4 przedstawię alternatywną propozycję takiego kryterium. Sekcja 5 poświęcona jest odpowiedzi na trzy możliwe zarzuty przeciw podanemu kryterium.

2. Argumenty sceptyczne

Prezentację argumentów sceptycznych zacznijmy od przykładu Searle'a – jest obrazowy i

---

[3] Celowo nie precyzuję w tym miejscu tego, czy chodzi tu o wszystkie obliczenia, czy tylko niektóre – będzie to podjęte w dalszej części artykuł.

[4] Wzoruję się w tym rozwiązaniu na Piccininim (Piccinini 2011, 7) i Miłkowskim (Miłkowski 2009, 168).

[5] Definicja ta wzorowana jest na definicji podanej przez Chalmersa (Chalmers 1996).

[6] Niektórym czytelnikom niejasne może się wydawać pojęcie „stanów formalnych obliczenia". Chodzi tu, jak jest w przypadku maszyn Turinga, o stany pewnego hipotetycznego urządzenia, które wykonuje obliczenie. Stany te nazywamy „formalnymi", ponieważ opis taki jak specyfikacja maszyny Turinga, abstrahuje od szczegółów fizycznych opisywanej maszyny. Możemy, rzecz jasna, pójść jeszcze dalej - nie musimy nawet wspominać o żadnych maszynach - ma to być jedynie specyfikacja procedury wykonania pewnego obliczenia, a poszczególne stany formalne rozumieć można jako poszczególne etapy tej procedury, co sprowadzić można do kolejnych zdań specyfikacji. W dalszych częściach tekstu, taki właśnie najprostszy przypadek będę miał na myśli.

pomoże pobudzić intuicje, które przydadzą się przy omawianiu dowodu Putnama. W artykule *Is the Brain a Digital Computer* (Searle 1990)*7* autor stawia następujące prowokacyjne pytanie: czy możemy opisać zwykłą ścianę tak, aby stanowiła realizację jakiegoś konkretnego programu - na przykład używanego przez niego w tamtym czasie edytora tekstu *Wordstar*? Okazuje się, że nie widać żadnych zasadniczych powodów, dla których miałoby to być niewykonalne. Pamiętajmy, że ściana nie jest obiektem tak prostym, jakby to się mogło początkowo wydawać - chropowatość powierzchni ściany nie jest symetryczna, nie jest też ona pokryta jednolitym kolorem - ilość szczegółów będzie rosła wraz z dokładnością opisu, a tę sami przecież dobieramy w zależności od potrzeb. Mając daną taką różnorodność, możemy pozwolić sobie na przyporządkowanie poszczególnych własności ściany poszczególnym elementom danego formalizmu tak, aby wszystko się zgadzało. Zawsze da się wskazać funkcję, która pierwszy zbiór (zbiór stanów ściany) odwzorowuje w drugi (zbiór etapów obliczenia) - ostatecznie może to być po prostu ciąg takich korelacji. Opis ten będzie prawdopodobnie dość karkołomny, ale co z tego? Opis komputera realizującego program tak złożony, jak edytor tekstu też nie będzie prosty - jaką niearbitralną dopuszczalną granicę komplikacji mielibyśmy tu przyjąć?

Aby lepiej uzmysłowić sobie, co Searle ma tutaj na myśli, porzućmy na chwilę jego argumentację i przywołajmy przykład, do którego będziemy się jeszcze w dalszej części tego artykułu odwoływali - załóżmy, że ktoś głosi, że w dowolnej książce zawarta jest dowolna inna książka. Co przez to rozumie? Weźmy jakiś konkretny przykład - *Lalkę* - co ma na myśli ktoś twierdząc, że w pewnym sensie książka ta zawiera w sobie treść dowolnej innej książki, na przykład *Zbrodni i kary*? Chodzi mu po prostu o to, że można stworzyć taki opis *Lalki*, który poszczególnym wyrażeniom przypisze niestandardową treść, a te niestandardowe przypisania ułożą się w nową, zaskakująco spójną całość.8 Ktoś mógłby zaoponować, twierdząc, że jest to jedynie trick, ponieważ ta nowa treść pojawiła się w książce tylko dzięki dodatkowi, jakim jest (zapewne opasły) podręcznik interpretacji, podczas gdy oryginalna treść znajdowała się „w samej *Lalce*" Mówiąc tak zapomniał jednak o tym, że treść oryginalna również nie jest dekodowana bez udziału podręcznika interpretacji - został on jedynie zinternalizowany przez czytelników znających język polski, ale mógłby równie dobrze znajdować się na stole. Różnica przestaje więc być taka łatwa do wskazania.

Wróćmy do Searle'a - zauważa on, że mamy tu do czynienia z pewnym naiwnym przeoczeniem. Ludzie z chęcią przyznają, że twierdzenie, że komputer przemnożył przez siebie dwie liczby, jest jedynie wygodnym sposobem mówienia, bo komputer żadnych liczb nie mnoży -

---

7 Przykład ten znajdzie czytelnik również w: Searle 1999.

8 Tym czytelnikom, którzy uznają wymieniony przykład za typowy dla filozofii dziwaczny eksperyment myślowy, pragnę przypomnieć, że jest to codzienność szyfranta.

jedyne, co robi, to manipulacja zerami i jedynkami.9 Zapominają jednak, że uznanie jednego ze stanów za 0, a drugiego za 1 również jest przyjętą konwencją, a nie czymś, co znaleźć można badając sam przedmiot. Niezależnie od tego, jak długo badalibyśmy naszą maszynę, nie znajdziemy w niej nie tylko semantyki (co pokazywał słynny przykład z chińskim pokojem), ale i składni, konkluduje Searle (Searle 1990, s. 25-27). Ponieważ ściany nie są w żadnym znanym nam sensie lepszym materiałem na komputer, niż inne przedmioty, to wynik Searle'a daje się uogólnić do następującego twierdzenia:

Twierdzenie Searle'a: Dla każdego obiektu o wystarczającej ilości części istnieje taki opis, który sprawia, że urządzenie to jest modelem dowolnego algorytmu. (Copeland 1996, s.339)

Zauważmy, że zastrzeżenie o „wystarczającej ilości części" jest w zasadzie niepotrzebne, ponieważ to, ile obiekt ma części, zależy od tego, jak szczegółowo zechcemy go opisać. Co więcej, jeśli tylko przyzwolimy sobie na pewną ontologiczną elastyczność (a nie wiadomo, co miałoby nas przed tym powstrzymać), to uzyskujemy wynik jeszcze trudniejszy do zaakceptowania – realizację obliczeń znaleźć możemy nie tylko w dowolnym obiekcie, ale i w dowolnym zbiorze obiektów albo dowolnej części dowolnego obiektu.

Dowodu twierdzenia, że dowolny obiekt realizuje dowolne obliczenie, dostarcza Hilary Putnam. Dowód ten stanowi doprecyzowanie intuicji, którą autor ten wyraził już wcześniej w artykule *The Nature of Mental States* (Putnam 1979), a którą Gualtiero Piccinni nazywa kanonicznym sformułowaniem mocnego pankomputacjonizmu (Piccinini 2007, s. 93).10 Oszczędzę tu czytelnikowi przybliżania szczegółów dowodu Putnama - choć nie jest on zawiły, to jego ekspozycja zajmuje sporo miejsca - wymaga on udowodnienia tezy pomocniczej i objaśnienia dwóch tez fizycznych, (które same w sobie nie są niekontrowersyjne).11 Co istotniejsze, krytyka, której został poddany, dotyczy założeń w nim przyjętych, a nie któregoś z dalszych kroków - nie ulega wątpliwości, że sam dowód jest poprawny. Skupmy się więc na przybliżeniu tych założeń i ocenie powodów, dla których wzbudzają kontrowersje.

W wystarczającym dla naszych potrzeb skrócie - Putnam dociera do potrzebnej mu konkluzji w następujący sposób: najpierw zakłada, że każdy obiekt fizyczny da się tak opisać, by wyróżnić w nim pewną liczbę następujących po sobie stanów12 a następnie powołuje się na pojęcie

---

9 Rozumianymi, rzecz jasna, jako wartości logiczne, a więc nie liczby.

10 Przez mocny pankomputacjonizm rozumie się tezę, że dowolny obiekt realizuje dowolne obliczenie. Pankomputacjonizm słaby to twierdzenie, że dowolny obiekt realizuje jakieś obliczenia.

11 Nie bez powodu sam Putnam umieszcza go w dodatku na końcu książki (Putnam 1988)

12 To zapewnione jest dzięki wspomnianym założeniom fizycznym. Zainteresowanych ich objaśnieniem i krytyką odsyłamy do Chrisley 1994

automatów o skończonej liczbie stanów (w oryginale Finite State Automata - w dalszej części artykułu posługiwał się będę akronimem FSA). Należy pamiętać, że podobnie do maszyn Turinga, automaty te są abstraktami - sprowadza się to do tego, że FSA rozumieć należy jako specyfikację automatu. Specyfikacja taka składa się z następujących elementów: opisu stanu początkowego obiektu i opisu repertuaru jego zachowań, który rozumieć należy jako zestawienie dwóch par uporządkowanych - jednej składającej się z wejścia i jakiegoś stanu wewnętrznego i drugiej, składającej się ze stanu wewnętrznego i wyjścia. Najciekawsze jest to, że Putnam pozwala sobie na dość zaskakujące dalsze uproszczenie - twierdzi, że interesują go izolowane (pozbawione wejść - w oryginale *inputless)* automaty. Specyfikacja takiego FSA sprowadza się po prostu do wymienienia ciągu stanów wewnętrznych, które następują po sobie w obiekcie. Po takim przygotowaniu przedpola możemy już dowieść, że dowolny obiekt daje się opisać tak, jakby realizował dowolny FSA.

Wniosek ten osiąga się w ten sposób, że mając do dyspozycji ciąg różnorodnych stanów obiektu fizycznego i ciąg stanów automatu, który wykonuje jakieś obliczenia, zawsze możemy wskazać funkcję, która kolejnym etapom przekształceń w automacie przypisuje kolejne stany fizyczne obiektu. W użytym na potrzeby dowodu przykładzie Putnam korzysta z niezwykle prymitywnego automatu, który zmienia swoje dwa stany (A i B) w następującej sekwencji: ABABABA. Wystarczy teraz w interesującym nas obiekcie wyróżnić jakąkolwiek sekwencję różniących się od siebie stanów – interesuje nas jedynie to, aby dwa kolejne elementy sekwencji nie były takie same. Powiedzmy, że jest to ciąg stanów 1234567. Warto zauważyć, że nie ułatwiamy sobie tutaj zadania i nie zakładamy, że udało się znaleźć sekwencję dwóch stanów występujących naprzemiennie! Wystarczy teraz, że zdefiniujemy A jako alternatywę 1 ∨ 3 ∨ 5 ∨ 7, a B jako alternatywę 2 ∨ 4 ∨ 6. Mówiąc inaczej - zawsze możemy wprowadzić *ad hoc* konwencję mówiącą, że poszczególne stany obiektu reprezentują dane stany automatu. To, czy mamy do czynienia z obiektem realizującym obliczenie, zależy tylko od naszego widzimisię, a tego właśnie mieliśmy dowieść.

Zgodnie z wcześniejszymi zapowiedziami, doprecyzuję teraz sposób, w jaki rozumiem arbitralność, przed którą chcę uratować własność posiadania struktury obliczeniowej. Potocznie „arbitralność” rozumie się często jako „zależność od obserwatora” (tym sformułowaniem posługuje się choćby Searle w Searle 1999). Sformułowanie to wydaje mi się jednakże niewystarczająco precyzyjne, ponieważ nie rozróżnia ono dwóch przypadków – zależności od własności obserwatora i zależności od opisu, czy konwencji, którą posługuje się obserwator. Dla wyjaśnienia tej różnicy posłużmy się takim przykładem: to, że znajdująca się przed oczami czytelnika zapisana kartka papieru jest dla niego artykułem, zależy po pierwsze od tego, że jego wzrok umożliwia mu odróżnianie liter od tła a po drugie od tego, czy stosuje on pewną konwencję, pozwalającą na

odczytanie tych liter. W obu przypadkach, mówiąc o artykule odwołujemy się do własności obserwatora, ale jedynie drugi z nich skłania sceptyka do wyrażenia podejrzenia, że własność bycia artykułem może być całkowicie zależna od obserwatora. Powodem, dla którego odwołanie się do konwencji otwiera sceptykowi tę możliwość jest to, że w odróżnieniu od zwykłych relacyjnych własności (takich jak „bycie zauważalnym dla x"), własności konwencjonalnej nie sposób wykryć empirycznie. Jest to, jak się wydaje, nieusuwalna charakterystyka konwencji, ponieważ konwencje nie są zdeterminowane żadnymi konkretnymi fizykalnymi własnościami obiektów, które są za ich pomocą opisywane.

Dotychczasowe rozważania możemy teraz streścić w następujących sześciu punktach (z których pierwsze trzy precyzują pojęcia własności arbitralnej i własności konwencjonalnej):

1. Własność arbitralna to taka, która przysługuje dowolnemu obiektowi przy jakiejś specyfikacji
2. Własność konwencjonalna to taka, która przysługuje danemu obiektowi przy jakiejś konwencji
3. Konwencja, to taka specyfikacja obiektu, która nie jest zdeterminowana jego własnościami fizykalnymi.
4. Każdą konwencję można dostosować do innego przedmiotu, pod warunkiem, że daje się w nim wyróżnić tyle samo elementów, co w obiekcie wyjściowym
5. Ilość elementów, które dają się w obiekcie wyróżnić zależy tylko od szczegółowości jego specyfikacji.
6. Jeżeli wykrycie, czy jakiś obiekt posiada strukturę obliczeniową wymaga odwołania się do konwencji, to jest to własność arbitralna.

3. Próby obrony pojęcia *realizacji*

W jaki sposób uchronić się możemy przed omówioną w sekcji 2 trywializacją pojęcia realizacji? Przede wszystkim Putnam musi jakoś ustosunkować się do nasuwającej się od początku wątpliwości – zasadności ograniczenia rozważań do dość specyficznego przypadku automatów pozbawionych wejścia.. Czy omawiany dowód daje się rozszerzyć na automaty z wejściem? Putnam wprost przyznaje, że nie jest to proste, ale stara się temu ograniczeniu zaradzić zauważając, że mając już dany konkretny zapis zachowań urządzenia (rozumiemy przez to zapis kolejnych stanów urządzenia pod wpływem napływających bodźców) możemy przebieg ów opisać tak, że pasował będzie do dowolnego FSA.

Tak dalekie uproszczenie FSA, sprowadzające je do pojedynczego, faktycznego przebiegu programu owocuje jeszcze jedną niepożądaną konsekwencją. Jak wskazał Chalmers (Chalmers 1996), aby obiektom, o których Putnam pisze, można było zasadnie przypisać realizację jakiegoś

automatu, powinny one w swojej budowie odzwierciedlać nie tylko faktycznie zrealizowane przez automat kroki, ale i takie, które w danym przebiegu nie zostały zrealizowane, ale byłyby zrealizowane w innych warunkach (czyli przy innych wartościach zmiennych). Taka zauważona przez nas korelacja stałaby się wygodną podstawą do przewidywania kolejnych działań obiektu. Doskonale widać teraz, że rozważanie automatów pozbawionych wejść okazuje się mieć konsekwencje filozoficzne znacznie istotniejsze, niż to by się mogło początkowo wydawać i nie można tego założenia nazwać jedynie nieszkodliwą idealizacją, na czym zapewne Putnamowi by zależało. Urządzenie posiadające wejścia musi mieć wbudowane alternatywne scenariusze zachowań. Rozpoznanie takich scenariuszy daje nam zdolność wypowiadania zdań kontrfaktycznych w rodzaju - „obiekt poszedł w lewo, bo zapaliło się czerwone światło, gdyby jednak zapaliło się zielone, poszedłby w prawo". Takie zdania są zaś podstawą do przewidywania nowych zachowań przedmiotu. W zasadniczy sposób ogranicza to nasze możliwości przypisywania struktury obliczeniowej dowolnym obiektom. Gdy tylko do przypisanej realizacji obliczenia dodamy alternatywne warunki, to, w sytuacji, gdy nie znamy przyszłych wejść, wystawiamy nasze hipotezy na testy, które z pewnością zmuszą nas do odrzucenia przynajmniej niektórych z nich.

Doskonale widać to w przypadku ściany Searle'a: spróbujmy bowiem na chwilkę serio potraktować jego przykład. Jak właściwie rozumieć mamy stwierdzenie, że ściana realizuje program *Wordstar*? Czy znaczy to, że możemy otwierać na niej nowe dokumenty, wprowadzać nowy tekst, kasować stary albo zmieniać formatowanie edytowanego akapitu? Trudno sobie wyobrazić, co miałoby to znaczyć, ponieważ nie dowiedzieliśmy się niczego o wejściach i wyjściach ściany - jedyne, co nam przedstawiono, to możliwość skorelowania pewnych własności ściany z sekwencją stanów wewnętrznych komputera, który na jakieś wejścia reagował. Przykład ten brzmi przekonująco tylko przy założeniu, że ścianie został przypisany pojedynczy przebieg pewnego programu, a nie program.

Co mogłoby zapewnić nam tę brakującą zdolność do wypowiadania się o sytuacjach kontrfaktycznych? Oczywiście związek przyczynowo-skutkowy odpowiada Chalmers i na obiekty, które podejrzewamy o to, że coś obliczają, proponuje nałożyć dodatkowy warunek: ich struktura przyczynowa ma odzwierciedlać strukturę formalną przypisywanego im obliczenia.

Jak należy to rozumieć? Chodzi w tym zastrzeżeniu o to, że nasze przyporządkowanie *f* ma być tak dobrane, że jeśli stan formalny *f(p)* przechodzi w nim w stan formalny *f(q)*, to odpowiadający temu pierwszemu stan fizyczny *p* wywołuje stan fizyczny *q* odpowiadający temu drugiemu.

Czy to rozwiązanie można uznać za poszukiwane przez nas niearbitralne kryterium posiadania struktury obliczeniowej? Zanim odpowiemy sobie na to pytanie, przyjrzymy się jeszcze rozwinięciu tej idei, jakim jest odwołanie się do pojęcia *mechanizmu* (Piccinini 2008, Miłkowski 2009). Głosi

ono, że aby przypisać czemuś własności obliczeniowe w nietrywialnym sensie, obiekt ten musi działać w ściśle określony sposób (na przykład spełniać pewną funkcję) dzięki wzajemnemu oddziaływaniu przyczynowemu jego części. Najważniejsza różnica pomiędzy podejściem mechanicystycznym a przyczynowym polega więc na tym, że zamiast o ciągu stanów, pomiędzy którymi zachodzi związek przyczynowo-skutkowy, mówi się o dynamicznej strukturze oddziałujących ze sobą przyczynowo części.

Na pierwszy rzut oka wydaje się że to rozwijające intuicje Chalmersa rozwiązanie ostatecznie oddala od nas widmo trywializacji pojęcia realizacji obliczenia. Od razu eliminuje ono spektakularne kontrprzykłady w rodzaju ściany, która nawet przy bardzo przychylnej ocenie nie wygląda na mechanizm. Nie cieszmy się jednak przedwcześnie, okazuje się bowiem, że tak zmodyfikowane pojęcie realizacji wcale nie lepiej nadaje się na poszukiwane przez nas niearbitralne kryterium posiadania struktury obliczeniowej. Jest to spowodowane tym, że wykorzystane w charakterze koła ratunkowego pojęcie *mechanizmu* posiada te same wady, co eksplikowane pojęcie *realizacji obliczenia*. Rozumiane jako „obiekt z dającymi się łatwo wyróżnić częściami, które wchodząc ze sobą w interakcje, spełniają pewną funkcję" (Piccinini 2008) przepuszcza zbyt wiele przypadków - pewne przypominające mechanizmy obiekty naturalne, takie jak galaktyki nadal zakwalifikowane mogą być do zbioru obiektów wyposażonych w strukturę obliczeniową. Jaki jednak inny sens słowa „mechanizm" mógłby wchodzić w grę? Z pewnością nie możemy powiedzieć, że dany obiekt jest mechanizmem, gdy funkcja, którą spełnia, jest realizacją jakiegoś obliczenia. Taka definicja nie nadawałaby się na dodatkowe kryterium realizowania obliczenia (z powodu błędnego koła). Na dodatek, jeżeli rzeczywiście jest tak, że każdy obiekt realizuje jakieś obliczenie, to każdy obiekt jest mechanizmem w tym sensie. Z drugiej strony, nie chcąc *a priori* odmawiać własności obliczeniowych obiektom, które swą budową w ogóle nie przypominają naszych stereotypowych przypadków mechanizmów (pomyślmy tu na przykład o jakimś komputerze obcej cywilizacji) możemy ulec presji i zdefiniować mechanizm w jakiś ogólniejszy i bardziej mglisty od wyjściowego sposób,13 co jeszcze bardziej poluzuje nasze, i tak już zbyt luźne, kryterium. Wystarczy wtedy tylko zakasać rękawy i zacząć szukać kolejnych spektakularnych kontrprzykładów.

A zatem, niezależnie od tego, jak użyteczne pojęcie mechanizmu okazuje się być w poszukiwaniach nietrywialnego sensu „realizacji obliczenia", zawiera ono zbyt wiele luk, przez które niechciana arbitralność wedrzeć się może do naszej teorii. Podkreślmy trzy takie newralgiczne aspekty tego pojęcia:

1. Ustalenie granic pomiędzy mechanizmem a jego otoczeniem - jak zauważa Craver

[13] Pojęciowe zawory, od których zależy, czy przyjęta definicja automatu będzie węższa, czy szersza to wyrażenie „spełnia pewną funkcję" i nasze przekonanie o tym, co można, a czego nie można uznać za część przedmiotu.

(rozwijając tezę obecną u Wimsatta), zawsze skazani jesteśmy tutaj na przyjęcie pewnych pragmatycznych ustaleń.14

2. Wydzielenie części w obiekcie - decyzja o tym, jak szczegółowo podzielimy badany przedmiot wydaje się zależeć tylko od nas. Tę deskrypcyjną elastyczność tym łatwiej uzyskać, że nikt nie nakazuje nam wskazywania zbioru części wyróżnionych na tym samym poziomie szczegółowości. Jedna z części może na przykład stanowić połowę całego obiektu, druga kawałek drugiej jego połowy, a trzecia coś wyodrębnionego na poziomie molekularnym.15
3. Wskazanie funkcji, jaki dany obiekt spełnia (jak również funkcji spełnianych przez jego części) także wydaje się być na łasce naszych deskrypcji. Rzeczywiste mechanizmy i ich części nie spełniają swojej funkcji bezbłędnie, przez co część z ich zachowań musimy w opisie pominąć. Zawsze istnieją jednakże konkurencyjne deskrypcje, w których to, co w poprzednich było awarią, jest inną spełnioną poprawnie funkcją.

Co istotniejsze, nawet jeśli uznamy związany z pojęciem mechanizmu poziom arbitralności za dopuszczalny, to nie unikniemy konieczności zmierzenia się z kłopotami, które rodzi integralnie z mechanizmami związane pojęcie przyczynowości, Podkreślmy, że jest to problem, z którym w równym stopniu boryka się rozwiązanie Chalmersa. Na czym dokładnie te kłopoty polegają?

Ogólne założenie wspólne wszystkim, którzy w kontekście problemu realizacji obliczeń powołują się na przyczynowość, można tak oto wyrazić: przyczynowość to coś więcej, niż następstwo zdarzeń. Poszczególni filozofowie mogą co najwyżej spierać się co do tego, na czym to „coś więcej" dokładnie polega. Kłopot w tym, że żadnej teorii przyczynowości nie udało się oddalić klasycznego zarzutu Hume'a: nawet jeśli postulujemy różnicę pomiędzy związkiem przyczynowym a zwykłym następstwem zdarzeń, to różnicy tej nie sposób wykryć empirycznie. Zauważmy, że nie pomoże nam ani odwołanie się do okresów kontrfaktycznych, ani (jak jest w koncepcji interwencjonistycznej) do możliwości manipulacji obiektem.16 Żadna taka teoria nie jest bowiem w stanie wykluczyć, że mamy do czynienia z wielce nieprawdopodobnym (ale możliwym, a to wystarczy) kosmicznym zrządzeniem losu, w którym badany obiekt przypadkowo przeszedł przez dokładnie taki ciąg stanów, jaki sugerowała nam nasza przyczynowa hipoteza.17

---

14 Wimsatt podaje następujący przykład takiego pragmatycznego założenia: chcąc odróżnić obiekt od jego otoczenia, możemy powołać się na kryterium zaproponowane przez Herberta Simona i uznać, że decydująca jest tu częstość oddziaływań – oddziaływania pomiędzy częściami obiektu są częstsze, niż oddziaływania między tymi częściami a otoczeniem. Wimsatt zauważa jednak, że dobór średniego poziomu oddziaływań, który uznamy za istotny, zależy już w dużej mierze od naszej arbitralnej decyzji. (Craver 2007, s. 142)

15 Craver zauważa, że często robimy tak opisując mechanizmy działające w mózgu.

16 Interwencjonizm zakłada, że związek przyczynowy różni się od zwykłego następstwa głównie tym, że w przypadku tego pierwszego jesteśmy w stanie interweniować w ciąg zdarzeń, wywołując oczekiwane skutki.

17 Czytelnikowi, który uzna takie pechowe zrządzenie losu za zbyt mało prawdopodobne by zawracać sobie nim głowę, pragnę przypomnieć, że zaczęliśmy od programu *Wordstar* uruchomionego na ścianie.

Empiryczną niewykrywalność związków przyczynowo-skutkowych możemy też zdyskontować w drugą stronę – każde przypadkowe zdarzenie, nawet jeśli zaszło jeden raz, mogę opisać jako niepowtarzalną konfigurację okoliczności, które powiązane były przyczynowo, ale jako jednorazowe zostały przez nas zakwalifikowane jako przypadek. Raz jeszcze podkreślmy - nawet jeśli nie przekreśla to użyteczności pojęcia związku przyczynowo-skutkowego18 a dzięki temu użyteczności zdefiniowanej za jego pomocą pojęcia realizacji obliczeń, to jest to dokładnie ten rodzaj sceptycyzmu, jaki uderza w ideę empirycznego kryterium, które pomogłoby nam w niearbitralny sposób wyróżnić obiekty zawierające struktury obliczeniowe. Jeżeli obiekty autentycznie dokonujące obliczeń różnią się od tych, które tylko na to wyglądają występowaniem w tych pierwszych związków przyczynowo-skutkowych, to nie odróżnimy pierwszych od drugich, bo nie umiemy empirycznie odróżniać rzeczywistych związków przyczynowo-skutkowych od ich przypadkowych kopii i przypadków od bardzo nietypowych związków przyczynowo-skutkowych. Co mamy w tej sytuacji począć?

Pewne nadzieje wzbudza propozycja Jacka Copelanda (Copeland 1996), który analizując przykład Searle'a stara się ratować pojęcie realizacji nakładając na nie obostrzenia podobne do tych, które proponował Chalmers, ale bez powoływania się na przyczynowość, czy mechanizmy.

Pierwszym etapem procedury, która ma, zdaniem Copelanda, pomóc nam w ustaleniu, czy mamy do czynienia z realizacją obliczenia, powinno być wyróżnienie części obiektu i nadanie im etykiet, które korelujemy z wyrażeniami języka, w którym zapisany jest interesujący nas formalizm. Teraz, by uchronić się przed pankomputacjonizmem, wprowadzamy dwa dodatkowe zastrzeżenia:

1. Korelacja nie może być dana *ex post factum*.
2. Korelacja musi pozwalać na formułowanie okresów kontrfaktycznych.

Warunek (1) rozumieć należy następująco: aby opis był adekwatny, nie możemy mieć do czynienia z sytuacją, w której z góry dane są: jakiś obiekt i jakiś algorytm, a nasze zadanie polega na możliwie najzręczniejszym dopasowaniu jednego do drugiego. Powinno być tak, że mając pewien obiekt, mogę, badając jego własności, wykryć, jakie obliczenie realizuje. Nietrudno dostrzec, że jest to warunek w kontekście naszych poszukiwań niezwykle obiecujący.

Warunek (2) powinien nam już być dobrze znany i oznacza, że mając opis obiektu zgodny z jakimś formalizmem, powinienem potrafić ustalić, jak obiekt zachowałby się, gdyby pewne zmienne parametry tego formalizmu były inne.

Zastanówmy się teraz, czy warunki te w niearbitralny sposób wyróżniają zbiór przedmiotów, o

[18] Na przykład Craver uznając zasadność hume'owskiej krytyki zauważa, że na szczęście nie umniejsza ona funkcji eksplanacyjnej przyczynowości (Craver 2007, s.64)

których powiedzieć można, że realizują obliczenia, a zatem, że posiadają strukturę obliczeniową

Warunek (1) oddać miał intuicję, że jeśli własności obliczeniowe rzeczywiście tkwią w obiekcie, to powinniśmy móc je wykryć bez wcześniejszego przyjmowania hipotez co do obliczeniowego charakteru obiektu – mówiąc obrazowo, powinno się je dać w przedmiocie zauważyć. Nietrudno go jednak nagiąć, choćby za pomocą takiego przykładu: wyobraźmy sobie, że wykonuję procedurę nadawania etykiet, a następnie metodą *brute force*, przeszukuję wszystkie znane mi algorytmy w poszukiwaniu takiego, który najlepiej pasował będzie do mojego opisu fizycznego. Następnie dokonuję mniej lub bardziej drastycznych korekt w etykietowaniu - korzystając z tego, że jak wspomnieliśmy, i tak była to czynność w dużej mierze arbitralna. Ktoś mógłby zauważyć, że jest to posunięcie nie *fair*, ponieważ celowo staram się wprowadzić kuchennymi drzwiami wypędzoną uprzednio arbitralność. Kłopot jednak w tym, że opisana procedura jest całkiem realistycznie brzmiącą taktyką badawczą. Czy mając pewien opis nieznanego obiektu i znajdując dość dobrze pasujący do niego algorytm, nie powinienem wprowadzić korekt, zakładając, że moje pierwotne wyróżnienie elementów było nie w pełni trafne? Czy poszukiwania realizowanego przez obiekt formalizmu nie mógłbym scedować na jakąś maszynę, która z braku lepszego pomysłu stosowałaby metodę *brute force*? Jak moglibyśmy upewnić się o tym, że pechowo nie trafiliśmy po prostu na algorytm, który akurat da się stosunkowo łatwo przedmiotowi przypisać?

Naturalnym ratunkiem zdaje się być warunek (2) – mamy do czynienia z czymś, co rzeczywiście oblicza, jeśli nie tylko udało mi się przypisać mu wykonanie jakiegoś algorytmu, ale i przypisanie to pozwala na wskazanie możliwych innych zachowań obiektu. Sens tego warunku jest taki, że nasz opis powinien wykraczać poza czas wykonywania danego obliczenia - jest coś dziwnego - zauważa Copeland - w tym, że opisawszy ścianę tak szczegółowo, nie jesteśmy w stanie powiedzieć niczego o tym, co robiła przed i po wykonaniu przypisanego jej obliczenia. Gdyby nasz sceptyk bronił się, twierdząc, że przed i po wykonaniu obliczenia ściana pozostawała w stanie nieustannej awarii, sam sprowadziłby swoją propozycję do absurdu.

Na pierwszy rzut oka niełatwo odmówić argumentowi Copelanda słuszności. Rozważmy jednak następujący scenariusz: załóżmy, że badamy obiekt O i spełnia on oba wymienione warunki - udaje się nam przypisać mu wykonywanie jakiegoś obliczenia w czasie $t_m...t_n$ (ale nie zrobiliśmy tego po fakcie) i ustalić, jak zachowywałby się w innych sytuacjach. Załóżmy nawet, że któreś z tych hipotetycznych sytuacji później się zdarzyły i O zachowywał się zgodnie z naszymi przewidywaniami. Rozważmy teraz inny przedmiot O', którego budowa i stany są przez pewien czas $t_{m-j}...t_{n+k}$ (chodzi po prostu o to, że odcinek czasu jest dłuższy, niż czas obliczania, dzięki czemu O' spełnia nie tylko warunek (1), ale i (2)) identyczne z budową i stanami naszego wyjściowego obiektu, ale jest to całkowity przypadek (w tym sensie, w jakim rozumieliśmy to przy okazji rozważań o związku przyczynowym). Czy mamy uznać, że O' realizował w czasie $t_m...t_n$

obliczenie, czy nie?

Jeżeli ze względu na przypadkowość (rozumianą jako zwykłe następstwo zdarzeń w czasie) uznamy, że obiekt nie realizował obliczenia, to warunek (2) nie różni się od rozważanego powyżej rozwiązania przyczynowego i podlega tej samej krytyce.

Jeżeli uznamy, że to zależy od wartości j i k, to wypadałoby powiedzieć, gdzie przebiega granica - z jednej strony mamy tu omówiony absurdalny przypadek, w którym wartość j i k to 0, z drugiej równie absurdalny wymóg, aby nasze przewidywania pokrywały wszystkie wcześniejsze i późniejsze zachowania obiektu. Trudno stwierdzić, czym, poza arbitralną decyzją, której chcieliśmy uniknąć, mielibyśmy się przy wyborze tych wartości kierować.

Ostatnią ze strategii radzenia sobie z zagrożeniem pankomputacjonizmu, której przydatność ocenimy, jest powołanie się na reprezentacje. Zamiast szukać cech wyróżniających struktury obliczeniowe, możemy poszukać ich w charakterystyce wykonywanego przez dany system obliczenia. Nawet jeśli jest tak, że każdy obiekt coś tam oblicza, to jedynie pewien ich podzbiór oblicza reprezentacje (Peacocke 1999). Ten tok rozumowania doczekał się nawet swojego sloganu, autorstwa Jerrego Fodora - *no computation without representation* (Fodor 1975).[19]

Zauważmy, że nawet gdyby przyjąć tę propozycję bez zastrzeżeń, to uzyskalibyśmy kryterium zdecydowanie zbyt restrykcyjne. Typowe komputery cyfrowe, od których wyszliśmy jako od paradygmatycznego przypadku obiektów ze strukturą obliczeniową, kryterium tego nie spełniają, ponieważ nie muszą posiadać żadnych reprezentacji. Jeżeli uznamy, że mimo to każdemu z nich można przypisać posiadanie reprezentacji „w jakimś sensie", to istnieje spore ryzyko, że posłużyliśmy się tak luźnym pojęciem posiadania reprezentacji, że przypisać je można wszystkiemu. Wystarczy choćby wspomnieć, że niektóre z teorii reprezentacji fundują tę relację na relacji przyczynowej, co czyni je dla naszych celów bezużyteczne z powodów omówionych wyżej.20 Doskonałą ilustracją obu tych pułapek są następujące dwie definicje reprezentacji:

(DR1) Reprezentacja jest to zakodowana w systemie treść (czy informacja), którą da się odnieść do dowolnych obiektów lub ich własności. (Żegleń 2005, s.

44)

(DR2) Reprezentacja jest to zakodowana w systemie treść (czy jakaś informacja), którą system jest w stanie zinterpretować i odnieść do określonych obiektów lub ich własności. (Żegleń 2005, s. 45).

Pierwsza z tych definicji prowadzi do pojęcia równie arbitralnego, co naiwne pojęcie realizacji, każdą zakodowaną w systemie treść *da się* odnieść do dowolnych obiektów lub ich własności,

---

[19] W dalszej części nie odnoszę się do pojęcia „reprezentacji" w rozumieniu Fodora, ponieważ podlega ono przedstawionej wcześniej krytyce rozwiązania odwołującego się do związku przyczynowo-skutkowego.

[20] Nie mam tu miejsca na to, by szczegółowo omawiać ten problem, ponieważ najistotniejsze jest dla mnie to, że powoływanie się na reprezentacje prowadzi do kryterium, które jest zbyt wąskie.

wystarczy stworzyć odpowiednią specyfikację systemu. Druga wyklucza komputery, o których nie zakładamy, że są w stanie odnosić przetwarzane treści do obiektów ze świata zewnętrznego.

Co to wszystko oznacza dla poszukiwań niearbitralnego kryterium posiadania struktury obliczeniowej? Wydaje się, że jedyne, co udało się nam uzyskać, to pewnego rodzaju "odroczenie wyroku". Pojęcie realizacji obliczeń daje się ochronić przed trywializacją na kilka omówionych powyżej sposobów. Kłopot w tym, że jedyne, co sceptyk musi teraz zrobić, by pozostać w grze, to przeformułować swoje zarzuty tak, aby trafiały w środki, których użyliśmy do ratowania pojęcia realizacji z czym, jak starałem się pokazać, nie będzie miał wielkich kłopotów. Być może powinniśmy zaakceptować ten stan rzeczy i uznać, że mieliśmy, po prostu, zbyt wygórowane wymagania?

4. W stronę niearbitralnego kryterium posiadania struktury obliczeniowej.

Sądzę, że byłoby to przedwczesne złożenie broni i że zabezpieczone przed zarzutem arbitralności kryterium posiadania struktury obliczeniowej da się sformułować. Aby je przedstawić, przywołajmy raz jeszcze sformułowanie, którym Searle streszcza swój zarzut – składnia nie jest własnością wewnętrzną przedmiotu (Searle 1990). Co rozumie się tutaj przez „składnię”?

Powiązanie pomiędzy strukturami obliczeniowymi i składnią nie powinno nikogo zaskakiwać - jednym ze sposobów na wyrażenie postulowanej przez nas różnicy pomiędzy ścianami a obiektami rzeczywiście realizującymi obliczenia jest powiedzenie, że choć wszystkie te fizyczne obiekty podlegają prawom przyrody, a więc rządzą nimi pewne reguły, to tylko o tych ostatnich można powiedzieć, że stosują się do pewnych reguł (Piccinini 2007, s. 94). Opozycja ta przywoływana jest często przy okazji omawiania innego ważnego rozróżnienia - na maszyny Turinga i uniwersalną maszynę Turinga. O ile poszczególne maszyny Turinga wykonują jedynie jakieś konkretne obliczenie, przez co możemy o nich myśleć jako o zbudowanych w taki sposób, by to konkretne obliczenie wykonywać (tablice zawierające kolejne kroki uznamy nie za oprogramowanie, a za część specyfikacji sprzętu), to w przypadku uniwersalnej maszyny Turinga tak nie jest. Ponieważ przeprowadzić ona może obliczenie wykonalne przez dowolną konkretną maszynę Turinga, to jej budowa odzwierciedlać musi reguły dekodowania instrukcji, które pozwalają jej dowolną maszynę emulować.

Zastanówmy się teraz, czy nie udałoby się nam wykorzystać tezy mocnego pankomputacjonizmu do naszych potrzeb. Zgódźmy się na to, że gdybyśmy się tylko postarali, moglibyśmy dowolnemu obiektowi przypisać realizację dowolnego obliczenia. Pomysł, który chciałbym teraz rozważyć, sprowadza się do tego, żeby przyjrzeć się dokładniej tej klasie obiektów, przy której nie musimy zbytnio się starać, bo, mówiąc obrazowo, one wykonują lwią część pracy za nas. Byłoby tak, gdybyśmy odkryli, że pewne obiekty zamiast być realizacją jakiegoś konkretnego przebiegu lub zbioru alternatywnych przebiegów, są po prostu (lub raczej

zawierają) realizację zbioru reguł do generowania wszystkich takich przebiegów. Wśród wszystkich obiektów (o których wiemy, że realizują jakieś obliczenie) wyróżniamy więc taką specyficzną klasę obiektów, które zamiast realizować jakiś pojedyncze obliczenie (i nie jest już dla nas istotne, na ile jest to obliczenie skomplikowane i czy zawiera reprezentacje, czy nie) mają po prostu dyspozycję do realizowania dowolnego obliczenia. Jeżeli zdolność do realizacji dowolnego obliczenia to zdaniem Czytelnika zbyt wiele, wystarczy uświadomić sobie, że te przedmioty, co do których nie mamy wątpliwości, że realizują obliczenia, czyli komputery, zdolność tę posiadają.

Mówiąc, że obiekt wykonuje część pracy za nas, mam na myśli tylko to, że mając zadane pewne warunki wyjściowe, resztę przebiegu daje się po prostu „przepowiedzieć" z obiektu, ponieważ jest zdeterminowana jego wewnętrzną budową.21 Przez „reguły składniowe" rozumiem zaś raczej coś na kształt Carnapowskich reguł formacji i transformacji, niż to, o czym mówi tradycyjne językoznawstwo. Zinternalizowane reguły transformacji to na przykład wytrawione w krzemie ścieżki, które, otrzymując pewien sygnał, przekazują go dalej w zmienionej formie (blokując lub przepuszczając jego części, a więc modyfikując go). Reguły formacji polegałyby zaś na tym, że nie wszystkie sygnały system jest w stanie odebrać - sam kształt receptorów wejścia blokuje sygnały niewłaściwe, co odpowiada odrzucaniu wadliwie skonstruowanych wyrażeń.

Idąc dalej tym tropem, spróbujmy to mówienie strukturze składniowej (a więc językowej) potraktować zupełnie dosłownie i korelację pomiędzy specyfikacją badanego obiektu i opisem jakiegoś obliczenia nazwać po prostu przekładem. Będzie to przekład języka użytego do opisu budowy jakiegoś obiektu na język, w którym zapisane jest dane obliczenie. Proponuję nazywać to przekładem, a nie korelacją, czy izomorfizmem, ponieważ chodzi nie tylko o wskazanie odpowiedniości pomiędzy dwoma tekstami, ale stworzenie podręcznika, który pokaże nam, jak jeden język przekładać na drugi, co sprowadza się do tego, że będziemy mogli tworzyć hipotezy mówiące nam, co musiałoby się z obserwowanym obiektem stać, aby można było o nim powiedzieć, że realizuje dowolne inne obliczenie.

Wróćmy raz jeszcze do naszego przykładu z *Lalką* i *Zbrodnią i karą*. Liczę na to, że każdego przykład ten uderza raczej jako rodzaj filozoficznej sztuczki, niż rzeczywisty problem. Na czym sztuczka ta polega? Polega, jak się wydaje, na tym, że twierdzi się, że treść *Zbrodni i kary* wyczytana jest w *Lalce*, podczas gdy wprowadzona jest ona kuchennymi drzwiami w dostarczonym podręczniku przekładu. Na zupełnie powierzchownym poziomie naszą podejrzliwość powinna zaś wzbudzać zaskakująca dysproporcja pomiędzy objętością przekładanego tekstu a owym podręcznikiem - będzie on znacznie dłuższy niż *Zbrodnia i kara* i *Lalka* razem wzięte.

Ponieważ będziemy z tej obserwacji dalej korzystać, podręczniki przekładu, które są krótsze od

---

[21] Dokładnie tak, jak wynik jakiejś operacji arytmetycznej daje się przepowiedzieć z budowy procesora, który by ją wykonywał.

sumy przekładów, które generują, nazwijmy *efektywnymi (*nie trzeba chyba dodawać, że wszystkie rzeczywiste podręczniki przekładów są efektywne)*.22* Mając te wątpliwości w pamięci, wyobraźmy sobie teraz następującą sytuację: załóżmy, że mamy do czynienia z przedmiotem, o którym praktycznie nic nie wiemy - nie wiemy, czy jest to egzemplarz gatunku naturalnego, czy artefakt, czy przejawia jakieś symptomy realizacji obliczeń czy nie itd. Naszym celem jest wykrycie, czy obiekt ten zawiera strukturę obliczeniową czy nie. Sądzę, że zadanie to moglibyśmy wykonać za pomocą empirycznej procedury składającej się z następujących etapów:23

(1) Analizujemy budowę przedmiotu i wprowadzamy etykiety na oznaczenie wszystkich zaobserwowanych stanów obiektu oraz potencjalnych stanów, co do których widzimy, że jego budowa je umożliwia. Zbiór takich etykiet nazwijmy L.

Stosujemy przy tym wszystkie rozsądne strategie, które są nam dostępne – szukamy powiązań przyczynowych, mechanizmów itd. Istotne jest jednak to, że cały czas traktujemy ten podział jako hipotezę, co do której możemy się całkowicie mylić – być może całość albo część powstałego zbioru etykiet stanowi tylko naszą projekcję i nie odzwierciedla rzeczywistych części przedmiotu.

(2) Tworzymy fizykalny opis F_1, składający się ze zdań (f_1,f_2...f_n) zbudowanych z etykiet ze zbioru L taki, że F_1 jest opisem rzeczywistego ciągu stanów, w których przedmiot znajdował się w danym czasie.

(3) Szukamy takiego opisu obliczenia C_1, rozumianego jako ciąg zdań (c_1,c_2...c_n), że potrafimy skorelować go z F_1 za pomocą przyporządkowania T_1 (korelującego poszczególne zdania, a więc stany obiektu i etapy obliczenia ze sobą).

---

22 Pojęcie *efektywności* podręcznika przekładu nasuwa skojarzenia z miarą złożoności Kołmogorowa, tym bardziej że o podręczniku przekładu można równe dobrze myśleć jak o programie do generowania przekładów. Nasz oparty na tricku, nieefektywny przekład byłby wtedy odpowiednikiem losowego ciągu znaków, który Kołmogorow zdefiniował jako taki ciąg, do którego wygenerowania trzeba użyć programu dłuższego od niego. Powiązanie z przypadkowością jest dość naturalne, ponieważ problem z przykładami w stylu Searle'a polega właśnie na tym, że ściana, czy (jak w naszym przykładzie) ta, czy inna książka, dobrane zostały zupełnie przypadkowo - równie dobrze można było dobrać inne, ponieważ całą pracę i tak wykonuje interpretator. Wolę mówić o nieefektywnym podręczniku przekładu, zamiast o programie, który go generuje, ponieważ nie muszę się wtedy przejmować pewnymi trudnościami, choćby tą, że każdy bardzo krótki ciąg jest w sensie Kołmogorowa losowy. Innym skojarzeniem, na którego analizę nie ma tutaj miejsca jest stała Chaitina. Przystępne wyjaśnienie tego zagadnienia w kontekście komputacjonizmu znajdzie czytelnik w Dębowski 2004

23 Nie przesądzam tu o tym, że jest to jedyna możliwa procedura, która pozwala na empiryczne wykrycie struktury obliczeniowej. Mogą istnieć inne, być może prostsze, metody ustalania, czy powstały podręcznik przekładu spełnia warunki opisane poniżej. Nie jest to istotne, ponieważ podane w zakończeniu sekcji 4 niearbitralne kryterium posiadania struktury obliczeniowej odwołuje się tylko do odpowiedniego podręcznika przekładu, a nie procedury opisanej w punktach 1-5.

Pary C_1, T_1 możemy zupełnie dobrze poszukiwać metodą *brute force*. Zauważmy, że w punkcie tym wykorzystujemy tezę pankomputacjonizmu na naszą korzyść – ponieważ nie nałożyliśmy żadnych dodatkowych warunków na nasze przyporządkowanie T_1 (nie wymagamy, by wydobywało ono strukturę przyczynową obiektu, by obiekt był mechanizmem, by korelować reprezentacje itd.), mamy stuprocentową pewność, że coś znajdziemy. Podobnie, jak w przypadku punktu (1), traktujemy to odkrycie ze sporą dozą nieufności – może rzeczywiście badany obiekt wykonał w tym czasie C_1, być może tylko mu to przypisaliśmy.

(4) Korzystając z T_1, zaczynamy tworzyć podręcznik przekładu f-zdań na c-zdania.

Początkowo, gdy dysponujemy jedynie pojedynczym przyporządkowaniem T_1, nasz podręcznik przekładu sprowadza się do listy szczegółowych równoważności korelującej ze sobą f-zdania i c-zdania. Korzystając z tej obserwacji, wprowadźmy w tym miejscu definicję, która za chwilkę nam się przyda:

Podręcznik przekładu nazwiemy *zamkniętym* wtw, gdy nie musimy już do niego dopisywać żadnych nowych szczegółowych równoważności.

Oznacza to, że dla każdej równoważności, którą weźmiemy pod uwagę, podręcznik albo ją już zawiera, albo też pozwala na jej wygenerowanie (będzie tak choćby wtedy, gdy podręcznik zawiera już jakąś ogólną formułę, której ta równoważność jest podstawieniem).

Zauważmy, że gdybyśmy zakończyli naszą pracę na tym etapie, to uzyskany w ten sposób podręcznik nie byłby efektywny. Każda lista szczegółowych równoważności będzie dłuższa niż suma przekładów, które generuje. Jest tak dlatego, że zawiera ona całe przekładane zdania (więc nie będzie od sumy tych zdań krótsza) i dodaje do nich jakiś symbol wskazujący na równoważność danego zdania z odpowiednim korelatem. Nawet jeśli będzie to jakiś jednoliterowy symbol, na przykład znak identyczności, to wydłuży to podręcznik o ten jeden znak, czyniąc go dłuższym, niż suma jego przekładów.24 Z tego powodu nie możemy zatrzymać się na punkcie (4).

(5) Wybieramy sobie jakieś inne obliczenie C_2 (takie, o którym skądinąd wiemy, że jest wykonalne) i próbujemy skonstruować przyporządkowanie T_2 sekwencji C_2 na sekwencję F_2 (utworzoną z etykiet zaczerpniętych z L).

Jest to najistotniejszy punkt naszej procedury. Zdaję sobie też sprawę, że może być na

[24] Tak właśnie wyglądałoby przyporządkowanie w przypadku ściany Searle'a.

pierwszy rzut oka dla czytelnika nieco niejasny – co właściwie nam z tego przyjdzie? Po pierwsze, chciałbym, aby uwadze czytelnika nie umknęło to, że odwróciliśmy tu kolejność badania – zamiast brać jakiś opis fizykalny i szukać dla niego obliczeniowego korelatu, wzięliśmy opis jakiegoś obliczenia i szukamy dla niego korelatu fizycznego. Po drugie, zauważmy, że podobnie, jak w punkcie (4), jesteśmy skazani na sukces, ponieważ celowo posłużyliśmy się naiwnym pojęciem realizacji. Punkt (5) rozumieć najlepiej jest następująco – zapytujemy w nim o pewien hipotetyczny scenariusz – wybieramy sobie jakieś obliczenie i pytamy, jaki ciąg stanów obiektu (z wykrytego przez nas repertuaru stanów L, który możemy przy tej okazji poszerzyć) miałby mu odpowiadać? Choć wiemy, że jakiś na pewno będzie mu odpowiadał (bo taki jest, jak już wiemy, urok naiwnego pojęcia realizacji), nie wiemy, jaką cenę przyjdzie nam za to zapłacić – jak wiele nowych etykiet będziemy musieli dodać, czy możemy ułatwić sobie zadanie, korzystając z niektórych przyporządkowań obecnych już w T_1 czy nie? Dopiero te pytania są dla nas interesujące, ponieważ to właśnie ta różnica w ekonomii opisu jest cechą, która pozostaje niewrażliwa na nasze zabiegi interpretacyjne.

(6) Powtarzamy punkt (5) w celu przekształcenia listy szczegółowych równoważności w efektywny i zamknięty podręcznik przekładu.

Robimy to, stosując dwie strategie:

Po pierwsze, wykorzystujemy wszystkie regularności, które dostrzeżemy i ujmujemy je w ogólnych zdaniach, które pozwolą nam na zastąpienie szczegółowych korelacji (które będą z tych ogólniejszych zdań wyprowadzalne).

Po drugie, poszukujemy w obiekcie „wbudowanych sekwencji" - takich sekwencji stanów, które są zdeterminowane samą jego budową. Jeżeli tylko zdarzy się nam znaleźć taką stałą sekwencję f_1-f_2...f_k,25 którą przełożyć możemy na ciąg zdań c_1...c_k, to zastępujemy odpowiedni ciąg prostych równoważności, korelując pojedyncze zdanie f_1 z całym blokiem c_1...c_k. Nietrudno zauważyć, że regularności te bardzo szybko pozwolą nam na wprowadzenie skrótów do naszego podręcznika przekładu, przekształcając go, tym samym, w podręcznik efektywny i zamknięty. Pozwala nam to na podanie następującej propozycji niearbitralnego i nietrywialnego kryterium posiadania struktury obliczeniowej:

Jeżeli można stworzyć efektywny i zamknięty podręcznik przekładu, który pozwala na skorelowanie dowolnego wykonalnego obliczenia funkcji z aktualnymi lub potencjalnymi stanami

25 Zapis ten rozumieć należy następująco: gdy obiekt znajduje się w stanie opisanym jako f_1, to następnie zawsze znajduje się w ciągu stanów f_2...f_k.

fizycznymi danego obiektu, to obiekt ten zawiera strukturę obliczeniową.

5. Zakończenie.

Zajmijmy się teraz trzema wątpliwościami, które mogły się w trakcie lektury poprzednich sekcji czytelnikowi nasunąć. Po pierwsze, moglibyśmy stworzyć efektywny i zamknięty podręcznik przekładu, gdybyśmy tylko celowo dobrali w punkcie (5) naszej procedury jakieś bardzo proste obliczenie a następnie testowali jedynie jego nieznacznie się różniące warianty. Wtedy od razu wykrylibyśmy regularności i radośnie uczcili to, tworząc odpowiednie uogólnienia, uzyskując dzięki temu efektywny podręcznik. Podręcznik ten mógłby zostać również uznany za zamknięty - wystarczy tylko, że tak dobierzemy kolejne obliczenia, że nie pojawiają się w nich żadne nowe zdania (zmienia się na przykład jedynie kolejność zdań). Czy dobierając taki specyficzny zestaw obliczeń sceptyk nie mógłby kolejny raz zatryumfować nad nami, przemieniając na naszych oczach ścianę w komputer? Oczywiście nie - przedstawilibyśmy mu wtedy kilka innych obliczeń i poprosili o pokazanie, że owa ściana nadal przechodzi podane kryterium. Pamiętajmy, że chcąc zdyskredytować nasze kryterium, sceptyk musi twierdzić, że udało mu się pokazać, że stworzony efektywny i zamknięty podręcznik przekładu pozwala przypisać ścianie realizację dowolnego obliczenia, nie może się więc przed naszym żądaniem wykręcić. Aby choćby przez chwilkę wyglądało na to, że jakiś zupełnie losowo dobrany przedmiot przechodzi nasze kryterium, zestaw korelowanych obliczeń musi być odpowiednio spreparowany. Im bardziej spreparowany zestaw, tym łatwiej go zdemaskować, wskazując na odbiegające od wykorzystanego schematu obliczenie i prosząc o jego przetestowanie.

Po drugie, mogłoby też tak się zdarzyć, że my sami ulegliśmy złudzeniu, że jakiś obiekt zawiera struktury obliczeniowe, bo przypadkowo tak dobraliśmy hipotetyczne obliczenia z punktu (5), że udało się nam stworzyć zamknięty i efektywny podręcznik przekładu. Być może, gdybyśmy popróbowali jeszcze kilka razy, okazałoby się, że się myliliśmy. Cóż, nie można oczywiście tego wykluczyć, można jedynie minimalizować prawdopodobieństwo takiej pomyłki, dbając o różnorodność dobieranych obliczeń. Jest to jednak los, który podane kryterium dzieli ze wszystkimi empirycznymi hipotezami - zawsze może być tak, że hipoteza potwierdzała się, bo trafiliśmy na specyficzny zestaw obserwacji, a nasza metodologia nie ostrzegła nas przed jego specyficznością. Jest to cena, którą trzeba płacić, gdy chce się zastosować indukcyjne uogólnienia. Nie należy jednak tego mankamentu mylić z o wiele od niego groźniejszą (bo prowadzącą do trywializacji) arbitralnością, której chcieliśmy uniknąć. Omówiona niepewność wcale do niej nie prowadzi - nawet jeśli nie możemy być pewni, czy dobraliśmy odpowiedni zbiór obliczeń, to nie mamy wątpliwości, że badany obiekt albo ma poszukiwaną własność, albo jej nie ma i nie zależy to od naszych zabiegów interpretacyjnych. Gdy po czasie dowiemy się, że mieliśmy do czynienia jedynie

z pechowym doborem próbek, skorygujemy swoje zdanie, pozostawiając samo kryterium w mocy.

Co istotne, podane kryterium broni się przed arbitralnością również w jej ekstremalnej, hume'owskiej wersji. Załóżmy bowiem, jak to robiliśmy w sekcji 2, że odkrywamy, iż mamy do czynienia z kosmicznym zrządzeniem losu i że badany obiekt przypadkowo pozwalał na wyróżnienie takiego zbioru stanów, który pozwolił mu w tym czasie przejść nasze kryterium. Co wtedy? Okazuje się, że w przypadku naszego kryterium nie ma żadnych wątpliwości co do tego, że w tym czasie był to po prostu obiekt zawierający struktury obliczeniowe - jest tak dlatego, że kryterium to pozwala nam na wykrycie poszukiwanej własności, ale nie przesądza o tym, co właściwie jest tą własnością. Jest nią po prostu taka budowa obiektu (niezależnie od szczegółów tej budowy), która pozwala na przypisanie mu dowolnego obliczenia (za pomocą efektywnego i zamkniętego podręcznika przekładu). Jeżeli dany obiekt taką budową się przez jakiś czas charakteryzował, to nie ma powodów do tego, by twierdzić, że nie miał w tym czasie struktury obliczeniowej. Powstanie w wyniku dziwacznego przypadku nie jest przecież samo w sobie dyskredytujące - gdyby, w wyniku zupełnie nieprawdopodobnego zrządzenia losu, gdzieś w kosmosie powstał obiekt atom w atom identyczny z moim komputerem, to nie widzę powodu, dla którego nie miałbym powiedzieć, że powstał tam komputer.

Po trzecie, moglibyśmy zapytać, czy fakt, że każdy obiekt fizyczny podlega prawom przyrody, nie sprawia, że wykazuje on wystarczającą ilość regularności do tego, by skonstruować efektywny i zamknięty podręcznik przekładu korelujący jego stany z dowolnym obliczeniem?

Aby oddalić ten zarzut, zacznijmy od zupełnie oczywistej uwagi: nikt, łącznie z Searlem, nie sądził nigdy, że na ścianie da się uruchomić edytor tekstu. Nikt nie miał też wątpliwości, że w przypadku niektórych przedmiotów, aby przypisać im zdolności obliczeniowe, musimy uciekać się do tricków w stylu definicji przez alternatywę z dowodu Putnama. Omawiana procedura pozwala na zdemaskowanie tych tricków - nawet jeśli uda się nam, dzięki wykryciu naturalnych regularności obiektu, znaleźć jakąś zręczną korelację z tym, czy innym obliczeniem, to wszystkie te dopasowane na zasadzie szczęśliwego trafu regularności staną się kulą u nogi, gdy tylko zastosujemy punkt (5) procedury z sekcji 4 - wybierzemy jakieś inne obliczenie i spróbujemy dokonać nowego, hipotetycznego przypisania. Nagle będziemy musieli wprowadzić *ad hoc* nowe definicje przez alternatywę albo przypisać tym samym regularnościom zupełnie inne korelaty obliczeniowe. Nasze możliwości będą bowiem ograniczone przez ten podzbiór naturalnych regularności i ich konfigurację, które akurat możemy w przedmiocie znaleźć. Gdyby się zaś tak zdarzyło, że jakiś obiekt, którego w ogóle nie podejrzewalibyśmy o zdolności obliczeniowe, zawiera podzbiór naturalnych regularności w takiej konfiguracji, że można by mu za pomocą efektywnego i zamkniętego podręcznika przekładu przypisać dowolne obliczenie, to powinniśmy po prostu przyjąć, że wbrew naszym oczekiwaniom, ma on strukturę obliczeniową. Jak zauważył Copeland,

gdyby okazało się, że chińskie jedwabniki mają tak skonstruowany system trawienny, że realizują obliczenia, to inżynierowie natychmiast wykorzystaliby ten fakt, a nie uznali za kompromitujący kontrprzykład (Copeland 1996). Raz jeszcze podkreślmy, że głównym problemem, który należało rozwiązać nie jest niejasność, czy nieostrość podziału na obiekty posiadające strukturę obliczeniową i resztę, ale nieodporność tego podziału na pewien typ argumentacji sceptyckiej. Przedstawione kryterium problem ten rozwiązuje, ponieważ, po pierwsze, nie odsiewa niekontrowersyjnych przypadków, takich jak komputery a po drugie, blokuje argumenty sceptyków, ponieważ wszystkie te argumenty oparte są na możliwości wprowadzania dodatkowych konwencji, a nie da się tego zrobić bez wydłużania podręcznika przekładu.[26]

Podsumowując - wykorzystując nieintuicyjną tezę mocnego pankomputacjonizmu, jesteśmy w stanie sformułować poszukiwane kryterium posiadania struktury obliczeniowej. Zamiast szukać mniej lub bardziej pomysłowych obostrzeń dla pojęcia realizacji, czy jakichś szczególnych obliczeń (czy to przez ich komplikację, czy to przez obecność w nich reprezentacji), warto zwrócić uwagę na to, że sama ta procedura przypisywania obiektowi różnych obliczeń może wyglądać rozmaicie. W pewnych przypadkach będzie wlokącym się rejestrem mapowań jednych stanów na drugie, a w innych, dzięki pewnym regularnościom, jedynie zbiorem reguł, które pozwolą nam na generowanie takiej korelacji, jaką tylko zechcemy. Nawet jeśli w przypadku jakiegoś konkretnego obliczenia nie możemy mieć pewności co do tego, czy obiekt faktycznie je realizuje, czy też jedynie tak się nam go szczęśliwie udało opisać, to ujawniająca się w opisanej procedurze uderzająca elastyczność niektórych obiektów jest autentyczną cechą, która im przysługuje. Jest to o tyle istotne, że wydaje się to być cecha, z którą można wiązać w filozofii umysłu spore nadzieje. Powiązanie pomiędzy posiadaniem struktur obliczeniowych a posiadaniem zdolności poznawczych jest bowiem ideą tak znaną, że nie trzeba jej chyba nikomu przedstawiać. Wystarczy zauważyć, że przedstawione przeze mnie kryterium posiadania struktur obliczeniowych doskonale nadaje się do obrony klasycznej wersji komputacjonizmu głoszącej, że umysł to uniwersalna maszyna Turinga.[27]

---

[26] W przypadku komputerów efektywnym i skończonym podręcznikiem przekładu jest podręcznik programowania w języku maszynowym.

[27] Kryterium to nie stanowi jednakże po prostu eksplanacji tej tezy – jest ono ogólniejsze, ponieważ nie zakłada się w nim nic na temat samego sposobu wykonywania obliczenia. Na rozwinięcie tych uwag nie mam w tym artykule wystarczającej ilości miejsca, wypada jednak wspomnieć o tym, pankomputacjonizm nie jest jedynym problemem klasycznego komputacjonizmu jako stanowiska w filozofii umysłu.

Literatura cytowana.

Chalmers D., (1996). *Does a Rock Implement Every Finite-State Automaton?* „Synthese" 108:309-33.

Chalmers D., (2010), *Świadomy umysł*, przeł. Marcin Miłkowski, PWN 2010

Chrisley L. R., (1994). *Why Everything Doesn't Realize Every Computation?* „Minds and Machines", 4, s. 403-420.

Copeland J., (1996). *What is Computation?* „Synthese" 108:335-359.

Craver F. C., (2007). *Explaining the Brain. Mechanisms and the Mosaic Unity of Neuroscience*, Oxford Clarendon Press 2007.

Dębowski, J., (2004) *Pułapki komputacjonizmu*, „Filozofia Nauki" nr 1(45), s. 29-50

Fodor A. J., (1975). *The Language of Thought*. Crowell, 1975

Miłkowski M., (2009). O tzw. metaforze komputerowej, „Analiza i Egzystencja", 9 / 2009, s. 163-185

Peacocke, C. 1999. *Computation as Involving Content*: *A Response to Egan*, "Mind and Language 14", s. 195 –202.

Piccinini G., (2007). *Computational Modelling vs. Computational Explanation: Is Everything a Turing Machine, And Does It Matter to the Philosophy of Mind?*, „Australasian Journal of Philosophy" Vol. 85, No. 1, s. 93-115

Piccinini G., (2008). *Computation without Representation*, „Philosophical Studies", 137.2 (2008)

Piccinini G., (2011). *Information processing, computation, and cognition*. „Journal of Biological Physics", 37.1, s. 1-38

Putnam H., (1979). *The Nature of Mental States*. [w:] Putnam, H. *Mind, Language and Reality. Philosophical Papers, Volume 2*, Cambridge University Press, s. 429-440

Putnam H., (1988). *Representation and Reality*, Cambridge MA: MIT Press.

Searle R. J., (1990). *Is the Brain a Digital Computer?* „Proceedings and Addresses of the American Philosophical Association", Vol. 64, Nr 3. s. 21-37

Searle R. J., (1999). *Umysł na nowo odkryty*, przeł. T. Baszniak, PIW Warszawa 1999.

Żegleń, M. U., (2005) *System poznawczy jako system reprezentacyjny*, „Filozofia Nauki", nr 4(52), s. 37-58