Paweł Grabarczyk

O niearbitralnym kryterium posiadania struktury obliczeniowej

1. WSTĘP[1]

Jest takie ironiczne określenie pracy filozofa: to człowiek, który zajmuje się wykazywaniem, że coś, co działa w praktyce, nie działa w teorii. Choć w wielu wypadkach jest ono zupełnie nieadekwatne i nawet nieco krzywdzące, to dobrze 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 dyskutować, czy wolno nam przypisywać komputerom i robotom stany intencjonalne, wolną wolę albo emocje, ale nie będziemy się raczej spierać o to, czy potrafią one realizować obliczenia. Niekontrowersyjne wydaje się także to, że są na świecie przedmioty, które różnią się od komputerów pod względem zdolności obliczeniowych, np. ściany, wiadra czy kamienie. Gdyby było inaczej, 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 czego nie mają kamienie, a co będę tu 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 (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

[1] Za wiele cennych uwag krytycznych chciałbym podziękować Jakubowi Michalskiemu, Marcinowi Miłkowskiemu i Michałowi Zawidzkiemu.

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 przedmiot. Dwa z tych argumentów chciałbym szczegółowo omówić, ale zanim do tego przejdę, przedstawię kilka uwag natury pojęciowej.

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

> Mówimy, że przedmiot fizyczny *realizuje pewne obliczenie*, gdy istnieje takie odwzorowanie *f* ze stanów fizycznych przedmiotu w stany formalne obliczenia, że gdy system znajduje się 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, taka definicja realizacji obliczeń prowadzi do trywializacji tego pojęcia, będę tak zdefiniowane pojęcie nazywał „naiwnym pojęciem realizacji".

Będę bronił stanowiska, zgodnie z którym posiadanie struktury obliczeniowej jest rzeczywistą i niearbitralną własnością, którą można wykryć w obiekcie metodami empirycznymi (sposób, w jaki pojmuję arbitralność, doprecyzowany zostanie w części 2). Stanowisko to samo w sobie nie jest specjalnie zaskakujące, jest ono odzwierciedleniem potocznych intuicji, o których wspominałem. Problemem jest nie tyle sam podział na przedmioty posiadające struktury obliczeniowe i ich pozbawione, ile jego nieodporność na kontrprzykłady w rodzaju wymienionych. Obrona będzie polegała na tym, że wskażę takie kryterium posiadania struktury obliczeniowej, które po pierwsze radzi sobie z podanymi przykładami i ich pewnymi modyfikacjami, a po

---

[2] O omówionej przez Copelanda (1996: 336) propozycji Hinckfussa wspominam jedynie ze względów historycznych; mimo pierwszeństwa nie jest ona szeroko omawiana w literaturze, nie różni się też pod żadnymi istotnymi dla nas względami od przykładu Searle'a.

[3] Celowo nie precyzuję w tym miejscu, czy chodzi tu o wszystkie obliczenia, czy tylko niektóre. Problem ten podejmę w dalszej części artykułu.

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

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

[6] Niektórym czytelnikom niejasne może się wydawać pojęcie „stanów formalnych obliczenia". Chodzi tu, tak jak w wypadku 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 maszyny. Możemy rzecz jasna pójść dalej: nie musimy nawet wspominać o żadnych maszynach, ma to być jedynie opis procedury wykonania pewnego obliczenia, a poszczególne stany formalne rozumieć można jako etapy tej procedury, co da się sprowadzić do kolejnych zdań specyfikacji. W dalszych częściach tekstu będę miał na myśli taki właśnie najprostszy przypadek.

drugie nie jest za wąskie, czyli pasuje do tych przedmiotów, o których wiemy, że mają strukturę obliczeniową.

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

2. ARGUMENTY SCEPTYCZNE

Przedstawienie argumentów sceptycznych zacznijmy od przykładu Searle'a: jest obrazowy i pomoże pobudzić intuicje, które przydadzą się przy omawianiu dowodu Putnama. W artykule *Is the Brain a Digital Computer* (Searle 1990; por. Searle 1999) autor stawia prowokacyjne pytanie: czy możemy opisać zwykłą ścianę tak, aby stanowiła realizację jakiegoś konkretnego programu, na przykład edytora tekstu *Wordstar*? Okazuje się, że nie widać żadnych zasadniczych powodów, dla których miałoby to być niewykonalne. Pamiętajmy, że ściana nie jest przedmiotem tak prostym, jakby to się mogło początkowo wydawać: chropowatość powierzchni ściany nie jest symetryczna, nie jest też ona pokryta jednolitym kolorem, a liczba szczegółów będzie rosła wraz z dokładnością opisu, który 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 nie ma to tu większego znaczenia. Opis komputera realizującego program tak złożony, jak edytor tekstu, też nie będzie prosty. Jaką więc 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 odwoływali w dalszej części artykułu. Załóżmy, że ktoś głosi, iż w każdej książce zawarta jest dowolna inna książka. Co przez to rozumie? Weźmy na przykład *Lalkę*. Co ma na myśli ktoś, twierdząc, że w pewnym sensie książka ta zawiera w sobie treść *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[7]. Ktoś mógłby zaoponować, twierdząc, że jest to jedynie sztuczka, ponieważ ta nowa treść pojawiła się w książce tylko dzięki dodatkowi, którym jest (zapewne opasły) podręcznik interpretacji, podczas

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

gdy oryginalna treść znajdowała się „w samej *Lalce*". Mówiąc tak, zapomniałby jednak o tym, że odkodowanie treści oryginalnej również wymaga podręcznika interpretacji. Został on jedynie zinternalizowany przez czytelników znających język polski, ale mógłby równie dobrze znajdować się przed nimi na stole jako książka, z której korzystają przy lekturze. Różnica przestaje więc być taka łatwa do wskazania.

Searle zauważa, że mamy tu do czynienia z pewnym naiwnym przeoczeniem. Ludzie z chęcią przyznają, że twierdzenie, iż komputer przemnożył przez siebie dwie liczby, jest tylko wygodnym sposobem mówienia, ponieważ komputer żadnych liczb nie mnoży, a jedynie manipuluje zerami i jedynkami[8]. 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 ani semantyki (co pokazywał słynny przykład z chińskim pokojem), ani składni (Searle 1990: 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 przedmiotu o wystarczającej liczbie części istnieje taki opis, który sprawia, że urządzenie to jest modelem dowolnego algorytmu (Copeland 1996: 339).

Zauważmy, że zastrzeżenie o „wystarczającej liczbie części" jest w zasadzie niepotrzebne, ponieważ to, ile przedmiot ma części, zależy od tego, jak szczegółowo zechcemy go opisać. Co więcej, jeśli tylko przyzwolimy 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, lecz także w dowolnym zbiorze przedmiotów albo dowolnej części dowolnego przedmiotu.

Dowodu twierdzenia, że dowolny przedmiot realizuje dowolne obliczenie, dostarcza Hilary Putnam. Dowód ten stanowi doprecyzowanie intuicji, którą wyraził już wcześniej w artykule *The Nature of Mental States* (Putnam 1979), a którą Gualtiero Piccinini (2007: 93) nazywa kanonicznym sformułowaniem mocnego pankomputacjonizmu[9]. Nie będę przybliżał szczegółów argumentacji Putnama. Choć dowód nie jest zawiły, to jego przedstawienie zajmuje dużo miejsca, wymaga udowodnienia tezy pomocniczej i objaśnienia dwóch tez fizycznych (które same w sobie nie są niekontrowersyjne)[10]. Co istotniejsze, krytyka, której został poddany, dotyczy założeń w nim przyjętych, a nie któregoś z dalszych kroków. Nie ulega bowiem wątpliwości, że

---

[8] Rozumianymi rzecz jasna jako wartości logiczne, a więc nie jako liczby.

[9] Przez mocny pankomputacjonizm rozumie się tezę, zgodnie z którą dowolny przedmiot realizuje dowolne obliczenie. Pankomputacjonizm słaby to twierdzenie, że dowolny przedmiot realizuje jakieś obliczenia.

[10] Nie bez powodu sam Putnam (1998) umieszcza go w dodatku na końcu książki.

sam dowód jest poprawny. Skupmy się więc na przybliżeniu tych założeń i ocenie powodów, dla których wzbudzają kontrowersje.

Putnam dociera do potrzebnej mu konkluzji, zakładając najpierw, że każdy przedmiot fizyczny da się tak opisać, by wyróżnić w nim pewną liczbę następujących po sobie stanów[11], a następnie powołuje się na pojęcie automatów o skończonej liczbie stanów (ang. *Finite State Automata*, dalej FSA). Należy pamiętać, że podobnie do maszyn Turinga automaty te są abstraktami. Oznacza to, że FSA należy rozumieć jako specyfikację automatu, na którą składa się opis stanu początkowego przedmiotu oraz opis repertuaru jego zachowań. Specyfikacja jest więc zestawieniem dwóch par uporządkowanych, pierwsza składa się z wejścia i jakiegoś stanu wewnętrznego, a druga ze stanu wewnętrznego i wyjścia. Najciekawsze, że Putnam pozwala sobie na dość zaskakujące dalsze uproszczenie. Twierdzi, że interesują go automaty izolowane (pozbawione wejść, w oryginale *inputless*). Specyfikacja takiego FSA sprowadza się po prostu do wymienienia ciągu stanów wewnętrznych, które w obiekcie następują po sobie. Po takim przygotowaniu możemy już dowieść, że dowolny przedmiot daje się opisać tak, jakby realizował dowolny FSA.

Mając do dyspozycji ciąg różnorodnych stanów przedmiotu fizycznego i ciąg stanów automatu, który wykonuje jakieś obliczenia, zawsze możemy wskazać funkcję, która kolejnym etapom przekształceń w automacie przypisuje kolejne stany fizyczne przedmiotu. W użytym na potrzeby dowodu przykładzie Putnam korzysta z niezwykle prymitywnego automatu, który zmienia swoje dwa stany (A i B) w sekwencji ABABABA. Wystarczy teraz w interesującym nas obiekcie wyróżnić jakąkolwiek sekwencję różniących się od siebie stanów, ważne jest jedynie, aby dwa kolejne elementy sekwencji nie były takie same. Powiedzmy, że jest to ciąg stanów 1234567. Nie musimy ułatwiać sobie zadania i nie zakładamy, że udało się znaleźć sekwencję dwóch stanów występujących naprzemiennie. Wystarczy, że zdefiniujemy A jako alternatywę $1 \vee 3 \vee 5 \vee 7$, a B jako alternatywę $2 \vee 4 \vee 6$. Mówiąc inaczej, zawsze możemy wprowadzić *ad hoc* konwencję, zgodnie z którą poszczególne stany przedmiotu reprezentują dane stany automatu. To, czy mamy do czynienia z przedmiotem realizującym obliczenie, zależy tylko od naszej decyzji, a tego właśnie mieliśmy dowieść.

Doprecyzujmy 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 1999). Sformułowanie to wydaje się niewystarczająco dokładne, ponieważ nie rozróżnia dwóch przypadków: zależności od własności obserwatora i zależności od opisu czy konwencji, którą posługuje się obserwator. W celu wyjaśnienia tej różnicy posłużmy się przykładem: to, że znajdująca się przed oczami Czytelnika zapisana kartka papieru jest dla niego artykułem, zależy po pierwsze od

[11] Zapewniają to wspomniane założenia fizyczne. Ich objaśnienie i krytykę można znaleźć w Chrisley 1994.

tego, czy 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 wypadkach, 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. Odwołanie się do konwencji otwiera sceptykowi tę możliwość, ponieważ 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 przedmiotów, które są za ich pomocą opisywane.

Dotychczasowe rozważania możemy podsumować w sześciu punktach, z których pierwsze trzy precyzują pojęcia własności arbitralnej i własności konwencjonalnej:

1. Własność arbitralna to taka, która przysługuje dowolnemu przedmiotowi przy jakiejś specyfikacji.
2. Własność konwencjonalna to taka, która przysługuje danemu przedmiotowi przy jakiejś konwencji.
3. Konwencja to taka specyfikacja przedmiotu, 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. Liczba elementów, które można wyróżnić w obiekcie, zależy tylko od szczegółowości jego specyfikacji.
6. Jeżeli wykrycie, czy jakiś przedmiot posiada strukturę obliczeniową, wymaga odwołania się do konwencji, to jest to własność arbitralna.

3. PRÓBY OBRONY POJĘCIA *REALIZACJI*

W jaki sposób możemy się uchronić przed omówioną w części 2 trywializacją pojęcia realizacji? Przede wszystkim Putnam musi jakoś ustosunkować się do nasuwającej się od początku wątpliwości co do zasadności ograniczenia rozważań do dość szczególnego wypadku 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. Zauważa, że mając już dany konkretny zapis zachowań urządzenia (czyli zapis kolejnych stanów urządzenia pod wpływem napływających bodźców), możemy przebieg ów opisać tak, by pasował do dowolnego FSA.

Takie uproszczenie FSA, sprowadzające je do pojedynczego faktycznego przebiegu programu, owocuje jeszcze jedną niepożądaną konsekwencją. Jak pokazał Chalmers (1996), aby przedmiotom, o których pisze Putnam, można było zasadnie przypisać realizację jakiegoś automatu, powinny w swojej budowie odzwierciedlać nie tylko faktycznie zrealizowane przez automat kroki, lecz także takie, które w da-

nym 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 przedmiotu. Doskonale widać teraz, że rozważanie automatów pozbawionych wejść ma znacznie istotniejsze konsekwencje filozoficzne, 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 „przedmiot poszedł w lewo, ponieważ 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 przedmiotom. Jeśli do przypisanej realizacji obliczenia dodamy alternatywne warunki, to w sytuacji, w której nie znamy przyszłych wejść, wystawiamy nasze hipotezy na testy, które z pewnością zmuszą nas do odrzucenia przynajmniej niektórych z nich.

Doskonale widać to w wypadku ściany Searle'a: spróbujmy bowiem potraktować jego przykład poważnie. Jak właściwie mamy rozumieć 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. Przedstawiono nam jedyne 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 sam program.

Co mogłoby zapewnić nam tę brakującą zdolność do wypowiadania się o sytuacjach kontrfaktycznych? Oczywiście związek przyczynowo-skutkowy — odpowiada Chalmers. Proponuje nałożyć dodatkowy warunek na przedmioty, o których przypuszczamy, że coś obliczają: ich struktura przyczynowa powinna odzwierciedlać strukturę formalną przypisywanego im obliczenia. 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 na to pytanie, przyjrzymy się rozwinięciu tej idei, jakim jest odwołanie się do pojęcia *mechanizmu* (Piccinini 2008, Miłkowski 2009). Głosi ono, że w celu przypisania czemuś własności obliczeniowych w nietrywialnym sensie, przedmiot ten musi działać w ściśle określony sposób (na przykład spełniać pewną funkcję) dzięki wzajemnemu oddziaływaniu przyczynowemu jego części. Najważniejsza różnica między podejściem mechanicystycznym a przyczynowym polega więc na tym, że zamiast o ciągu stanów,

między którymi zachodzi związek przyczynowo-skutkowy, mówi się o dynamicznej strukturze oddziałujących ze sobą przyczynowo części.

Na pierwszy rzut oka wydaje się, że to rozwiązanie, rozwijające intuicje Chalmersa, ostatecznie oddala widmo trywializacji pojęcia realizacji obliczenia. Od razu eliminuje spektakularne kontrprzykłady w rodzaju ściany, która nawet przy bardzo przychylnej ocenie nie wygląda na mechanizm. Okazuje się jednak, że tak zmodyfikowane pojęcie realizacji wcale nie lepiej nadaje się na poszukiwane przez nas niearbitralne kryterium posiadania struktury obliczeniowej. Pojęcie *mechanizmu* jest bowiem obarczone tymi samymi wadami co wyjaśniane pojęcie *realizacji obliczenia*. Rozumiane jako „przedmiot z dającymi się łatwo wyróżnić częściami, które wchodząc ze sobą w interakcje, spełniają pewną funkcję" (Piccinini 2008) dopuszcza zbyt wiele przypadków. Pewne przypominające mechanizmy przedmioty naturalne, takie jak galaktyki, nadal zakwalifikowane mogą być do zbioru przedmiotów wyposażonych w strukturę obliczeniową. Jaki jednak inny sens słowa „mechanizm" mógłby wchodzić w grę? Z pewnością nie możemy powiedzieć, że dany przedmiot 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 każdy przedmiot realizuje jakieś obliczenie, to w tym sensie każdy przedmiot jest mechanizmem. Nie chcąc jednak *a priori* odmawiać własności obliczeniowych przedmiotom, 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[12], co jeszcze bardziej rozluźni nasze i tak już zbyt szerokie kryterium. Wystarczy wtedy tylko zakasać rękawy i zacząć szukać kolejnych spektakularnych kontrprzykładów.

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

1. Ustalenie granic między mechanizmem a jego otoczeniem. Jak zauważa Craver (rozwijając tezę obecną u Wimsatta), zawsze skazani jesteśmy na przyjęcie pewnych pragmatycznych ustaleń[13].

---

[12] 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.

[13] Wimsatt podaje przykład takiego pragmatycznego założenia: chcąc odróżnić przedmiot od jego otoczenia, możemy powołać się na kryterium zaproponowane przez Herberta Simona i uznać, że decydująca jest tu częstość oddziaływań — oddziaływania między częściami przedmiotu 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: 142).

2. Wydzielenie części w obiekcie. Decyzja o tym, jak szczegółowo podzielimy badany przedmiot, wydaje się zależeć tylko od nas. Tę elastyczność opisu 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 przedmiotu, druga kawałek drugiej jego połowy, a trzecia coś wyodrębnionego na poziomie molekularnym[14].

3. Także wskazanie funkcji, które spełnia dany przedmiot (jak również funkcji spełnianych przez jego części), wydaje się zależne od nas. Rzeczywiste mechanizmy i ich części nie spełniają swojej funkcji bezbłędnie, przez co część z ich zachowań musimy w opisie pominąć. Zawsze istnieją jednak konkurencyjne charakterystyki, w których to, co w poprzednich było awarią, jest inną spełnianą 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 związane integralnie z mechanizmami pojęcie przyczynowości. Podkreślmy, że jest to problem, z którym boryka się również rozwiązanie Chalmersa. Na czym dokładnie polegają te kłopoty?

Ogólne założenie wspólne wszystkim, którzy w kontekście problemu realizacji obliczeń powołują się na przyczynowość, można wyrazić tak: przyczynowość to coś więcej niż następstwo zdarzeń. Poszczególni filozofowie mogą co najwyżej spierać się co do tego, na czym dokładnie to „coś więcej" polega. Kłopot w tym, że żadnej teorii przyczynowości nie udało się oddalić klasycznego zarzutu Hume'a: nawet jeśli postulujemy różnicę między 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 przedmiotem[15]. 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 przedmiot przypadkowo przeszedł przez dokładnie taki ciąg stanów, jaki sugerowała nam nasza przyczynowa hipoteza[16]. Empiryczną niewykrywalność związków przyczynowo-skutkowych możemy też wykorzystać 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-skutko-

[14] Craver (2007: 10) zauważa, że często tak robimy, opisując mechanizmy działające w mózgu.

[15] Interwencjonizm zakłada, że związek przyczynowy różni się od zwykłego następstwa głównie tym, że w wypadku tego pierwszego jesteśmy w stanie oddziaływać na ciąg zdarzeń, by wywoływać oczekiwane skutki.

[16] Czytelnikowi, który uzna takie pechowe zrządzenie losu za zbyt mało prawdopodobne, pragnę przypomnieć, że zaczęliśmy od programu *Wordstar* uruchomionego na ścianie.

wego[17], a dzięki temu użyteczności zdefiniowanej za jego pomocą pojęcia realizacji obliczeń, to jest to dokładnie taki rodzaj sceptycyzmu, jaki uderza w ideę empirycznego kryterium, które pomogłoby nam w niearbitralny sposób wyróżnić przedmioty zawierające struktury obliczeniowe. Jeżeli przedmioty rzeczywiście dokonujące obliczeń różnią się od tych, które tylko pozornie je wykonują, występowaniem w tych pierwszych związków przyczynowo-skutkowych, to nie odróżnimy pierwszych od drugich, ponieważ nie umiemy empirycznie odróżniać rzeczywistych związków przyczynowo-skutkowych od ich przypadkowych kopii, a przypadków od bardzo nietypowych związków przyczynowo-skutkowych.

Pewne nadzieje na wyjście z impasu budzi propozycja Copelanda (1996), który analizując przykład Searle'a, stara się ocalić pojęcie realizacji przez nałożenie na nie obostrzeń podobnych do tych, które proponował Chalmers, ale bez powoływania się na przyczynowość czy mechanizmy.

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

1. Korelacja nie może być dana *ex post*.

2. Korelacja musi pozwalać na formułowanie okresów kontrfaktycznych.

Zgodnie z warunkiem (1) opis nie jest adekwatny, jeśli nasze zadanie polega na możliwie najzręczniejszym dopasowaniu do siebie z góry danych przedmiotu i algorytmu. Powinniśmy natomiast wykrywać za pomocą badania własności przedmiotu, jakie realizuje on obliczenie. Nietrudno dostrzec, że w kontekście naszych poszukiwań jest to niezwykle obiecujący warunek.

Warunek (2) jest już nam dobrze znany i oznacza, że mając opis przedmiotu zgodny z jakimś formalizmem, powinniśmy potrafić ustalić, jak przedmiot zachowałby się, gdyby pewne zmienne parametry tego formalizmu były inne.

Zastanówmy się teraz, czy warunki te w niearbitralny sposób wyróżniają zbiór przedmiotów, o których powiedzieć można, że realizują obliczenia, a zatem, że posiadają strukturę obliczeniową

Warunek (1) oddać miał intuicję, że jeśli własności obliczeniowe rzeczywiście tkwią w przedmiocie, to powinniśmy móc je wykryć bez wcześniejszego przyjmowania hipotez co do obliczeniowego charakteru przedmiotu. Mówiąc obrazowo, powinno być możliwe zauważenie ich w przedmiocie. Nietrudno jednak nagiąć ten warunek, choćby za pomocą takiego przykładu: wyobraźmy sobie, że wykonuję procedurę nadawania etykiet, a następnie przeszukuję wszystkie znane mi algorytmy metodą *brute force* w poszukiwaniu takiego, który najlepiej będzie pasował do mojego

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

opisu fizycznego. Następnie dokonuję mniej lub bardziej drastycznych korekt w etykietowaniu, korzystając z tego, że i tak była to czynność w dużej mierze arbitralna. Ktoś mógłby zauważyć, że jest to posunięcie nieuczciwe, 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 przedmiotu i znalazłszy 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 przedmiot formalizmu nie mógłbym scedować na jakąś maszynę, która z braku lepszego pomysłu stosowałaby metodę *brute force*? Jak moglibyśmy się upewnić, że pechowo nie trafiliśmy po prostu na algorytm, który akurat da się stosunkowo łatwo przypisać przedmiotowi?

Naturalnym ratunkiem zdaje się warunek (2): mamy do czynienia z czymś, co rzeczywiście oblicza, jeśli nie tylko udało nam się przypisać mu wykonanie jakiegoś algorytmu, lecz także przypisanie to pozwala wskazać inne możliwe zachowania przedmiotu. 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 wykonaniem przypisanego jej obliczenia i po nim. Gdyby nasz sceptyk bronił się, twierdząc, że przed wykonaniem obliczenia i po nim ściana pozostawała w stanie nieustannej awarii, sam sprowadziłby swoją propozycję do absurdu.

Na pierwszy rzut oka niełatwo odmówić argumentowi Copelanda słuszności. Rozważmy jednak scenariusz, w którym badamy przedmiot O spełniający oba wymienione warunki. Udaje się nam przypisać mu wykonywanie jakiegoś obliczenia w czasie $t_m \ldots t_n$ (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 $t_{m-j} \ldots t_{n+k}$ (odcinek dłuższy niż czas obliczania, dzięki czemu O' spełnia nie tylko warunek (1), lecz także (2)) identyczne z budową i stanami naszego wyjściowego przedmiotu, ale jest to całkowity przypadek (w tym sensie, w jakim rozumieliśmy to przy okazji rozważań o związku przyczynowym). Czy powinniśmy uznać, że O' realizował w czasie $t_m \ldots t_n$ obliczenie, czy nie?

Jeżeli ze względu na przypadkowość (rozumianą jako zwykłe następstwo zdarzeń w czasie) uznamy, że przedmiot nie realizował obliczenia, to warunek (2) nie różni się od rozważanego wcześniej rozwiązania przyczynowego i podlega tej samej krytyce. Jeżeli uznamy, że to zależy od wartości *j* i *k*, wypadałoby powiedzieć, gdzie przebiega granica. Z jednej strony mamy absurdalny wypadek, w którym wartość *j* i *k* to 0, z drugiej równie absurdalny wymóg, aby nasze przewidywania obejmowały wszystkie wcześniejsze i późniejsze zachowania przedmiotu. Trudno stwierdzić, czym poza arbitralną decyzją, której chcieliśmy uniknąć, mielibyśmy się kierować przy wyborze tych wartości.

Przyjrzyjmy się jeszcze strategii radzenia sobie z zagrożeniem pankomputacjonizmu związanej z powoływaniem się na reprezentacje. Zamiast szukać cech wyróżniających struktury obliczeniowe, możemy poszukać ich w charakterystyce wykonywanego przez dany system obliczenia. Nawet jeśli każdy przedmiot coś oblicza, to jedynie pewien ich podzbiór oblicza reprezentacje (Peacocke 1999). Ten tok rozumowania doczekał się nawet swojego sloganu autorstwa Jerrego Fodora (1975): „no computation without representation”[18].

Zauważmy, że nawet gdyby przyjąć bez zastrzeżeń tę propozycję, uzyskalibyśmy kryterium zdecydowanie zbyt restrykcyjne. Typowe komputery cyfrowe, od których wyszliśmy jako od paradygmatycznego przypadku przedmiotów ze strukturą obliczeniową, kryterium tego nie spełniają, ponieważ nie muszą dysponować żadnymi reprezentacjami. Jeżeli uznamy, że mimo to każdemu z nich można przypisać posiadanie reprezentacji „w jakimś sensie”, to istnieje spore ryzyko, że posłużyliśmy się tak szerokim pojęciem posiadania reprezentacji, że przypisać je można wszystkiemu. Wystarczy wspomnieć, że niektóre teorie reprezentacji analizują tę relację w kategoriach związku przyczynowego, co czyni je bezużytecznymi dla naszych celów z wcześniej przedstawionych już powodów[19]. Doskonałą ilustracją obu pułapek są następujące definicje reprezentacji:

(DR1) Reprezentacja jest to zakodowana w systemie treść (czy informacja), którą da się odnieść do dowolnych przedmiotów lub ich własności (Żegleń 2005: 44).

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

DR1 prowadzi do pojęcia równie arbitralnego, co naiwne pojęcie realizacji. Każdą zakodowaną w systemie treść *da się* bowiem odnieść do dowolnych przedmiotów lub ich własności. Wystarczy stworzyć odpowiednią specyfikację systemu. DR2 wyklucza komputery, o których nie zakładamy, że są w stanie odnosić przetwarzane treści do przedmiotów ze świata zewnętrznego.

Co to wszystko oznacza przy poszukiwaniach niearbitralnego kryterium posiadania struktury obliczeniowej? Wydaje się, że jedyne, co udało się nam uzyskać, to pewnego rodzaju „odroczenie”. Pojęcie realizacji obliczeń można obronić przed trywializacją na kilka omówionych 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. Jak starałem się pokazać,

[18] W dalszej części nie odnoszę się do pojęcia „reprezentacji” w rozumieniu Fodora, ponieważ podlega ono tej samej krytyce co rozwiązania odwołujące się do związku przyczynowo-skutkowego.

[19] Nie będę tu szczegółowo omawiać tego problemu, ponieważ najistotniejsze jest dla mnie, że powoływanie się na reprezentacje prowadzi do kryterium, które jest zbyt wąskie.

nie będzie miał z tym wielkich problemów. Być może powinniśmy więc zaakceptować ten stan rzeczy i uznać, że mieliśmy po prostu zbyt wygórowane wymagania?

4. W STRONĘ NIEARBITRALNEGO KRYTERIUM POSIADANIA STRUKTURY OBLICZENIOWEJ

Sądzę, że byłaby to przedwczesna kapitulacja i że można sformułować kryterium posiadania struktury obliczeniowej odporne na zarzut arbitralności. Aby je przedstawić, przywołajmy raz jeszcze sformułowanie, którym Searle (1990) streszcza swój zarzut: składnia nie jest własnością wewnętrzną przedmiotu. Co rozumie się tutaj przez „składnię"?

Powiązanie między strukturami obliczeniowymi i składnią nie powinno nikogo zaskakiwać: jednym ze sposobów wyrażenia postulowanej przez nas różnicy między ścianami a przedmiotami rzeczywiście realizującymi obliczenia jest powiedzenie, że choć wszystkie przedmioty fizyczne podlegają prawom przyrody, a więc rządzą nimi pewne reguły, to tylko te, które realizują obliczenia, stosują się do pewnych reguł (Piccinini 2007: 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), o tyle w wypadku uniwersalnej maszyny Turinga tak nie jest. Ponieważ może przeprowadzić obliczenie wykonalne przez dowolną konkretną maszynę Turinga, jej budowa odzwierciedlać musi reguły dekodowania instrukcji, które pozwalają jej emulować dowolną maszynę.

Zastanówmy się teraz, czy nie udałoby się nam wykorzystać tezy mocnego pankomputacjonizmu do naszych potrzeb. Zgódźmy się, że gdybyśmy się tylko postarali, moglibyśmy dowolnemu przedmiotowi przypisać realizację dowolnego obliczenia. Pomysł, który chciałbym teraz rozważyć, sprowadza się do dokładniejszego przejrzenia tej klasie przedmiotów, przy której nie musimy zbytnio się starać, ponieważ mówiąc obrazowo, wykonują za nas lwią część pracy. To przedmioty, które zamiast być realizacją jakiegoś konkretnego przebiegu lub zbioru alternatywnych przebiegów, są po prostu realizacją (lub raczej zawierają realizację) zbioru reguł generowania wszystkich takich przebiegów. Wśród wszystkich przedmiotów (o których wiemy, że realizują jakieś obliczenie) wyróżniamy więc taką szczególną klasę przedmiotów, które zamiast realizować jakiś pojedyncze obliczenie (i nie jest już dla nas istotne, jak bardzo skomplikowane jest to obliczenie, ani czy zawiera reprezentacje) mają po prostu dyspozycję do realizowania dowolnego obliczenia. Jeżeli zdolność do realizacji dowolnego obliczenia to zdaniem Czytelnika zbyt wiele, wystarczy uświadomić sobie, że zdolność tę posiadają komputery, czyli przedmioty, co do których nie mamy wątpliwości, że realizują obliczenia.

Mówiąc, że przedmiot wykonuje część pracy za nas, mam na myśli tylko to, że mając zadane pewne warunki wyjściowe, resztę przebiegu można po prostu „przepowiedzieć" z przedmiotu, ponieważ jest zdeterminowany jego wewnętrzną budową[20]. 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 go modyfikując). 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 zupełnie dosłownie potraktować mówienie o strukturze składniowej (a więc językowej), a korelację między specyfikacją badanego przedmiotu i opisem jakiegoś obliczenia nazwać po prostu przekładem. Będzie to przekład języka użytego do opisu budowy jakiegoś przedmiotu na język, w którym zapisane jest dane obliczenie. Proponuję nazywać to przekładem, a nie korelacją czy izomorfizmem, ponieważ chodzi nie tylko o wskazanie odpowiedniości między dwoma tekstami, lecz także o stworzenie podręcznika, który pokaże, jak jeden język przekładać na drugi, co sprowadza się do tego, że będziemy mogli tworzyć hipotezy na temat tego, co musiałoby się stać z obserwowanym przedmiotem, aby można było o nim powiedzieć, że realizuje dowolne inne obliczenie.

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

Ponieważ dalej będziemy z tej obserwacji korzystać, podręczniki przekładu, które są krótsze od sumy generowanych przez siebie przekładów, nazwijmy *efektywnymi* (nie trzeba chyba dodawać, że wszystkie rzeczywiste podręczniki przekładów są efektywne)[21]. Pamiętając o tych wątpliwościach, wyobraźmy sobie, że mamy do

[20] Dokładnie tak, jak wynik jakiejś operacji arytmetycznej można przepowiedzieć z budowy procesora, który by ją wykonywał.

[21] Pojęcie *efektywności* podręcznika przekładu nasuwa skojarzenia z miarą złożoności Kołmogorowa, tym bardziej że o podręczniku przekładu można równie dobrze myśleć jak o programie do generowania przekładów. Nasz oparty na sztuczce i 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 lub (jak w naszym przykładzie) ta czy inna książka zostały dobrane zupełnie przypadkowo. Równie dobrze

czynienia z przedmiotem, o którym praktycznie nic nie wiemy: nie wiemy, czy jest to egzemplarz rodzaju naturalnego, czy artefakt, czy przejawia jakieś symptomy realizacji obliczeń, czy nie itd. Naszym celem jest wykrycie, czy przedmiot ten zawiera strukturę obliczeniową. Sądzę, że zadanie to moglibyśmy wykonać za pomocą empirycznej procedury składającej się z następujących etapów[22]:

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

Stosujemy przy tym wszystkie rozsądne strategie, które są nam dostępne, czyli szukamy powiązań przyczynowych, mechanizmów itd. Istotne jest jednak, 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 i będący 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 C_1 z F_1 za pomocą przyporządkowania T_1 (korelującego poszczególne zdania, a więc stany przedmiotu i etapy obliczenia).

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 przedmiotu ani by korelowało reprezentacje, ani by przedmiot był mechanizmem itd.), mamy pewność, że coś znajdziemy. Podobnie jak w wypadku punktu (1) traktujemy to odkrycie ze sporą dozą nieufności: może rzeczywiście badany przedmiot wykonał w tym czasie C_1, może tylko mu to przypisaliśmy.

---

można było wybrać 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 znajduje się w Dębowski 2004.

[22] Nie przesądzam tu, czy 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 opisywane niżej warunki. Nie jest to istotne, ponieważ podane w zakończeniu części 4 niearbitralne kryterium posiadania struktury obliczeniowej odwołuje się tylko do odpowiedniego podręcznika przekładu, a nie procedury opisanej w punktach 1-5.

(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 chwilę nam się przyda:

> Podręcznik przekładu nazwiemy *zamkniętym* wtedy i tylko wtedy, 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 pracę na tym etapie, uzyskany w ten sposób podręcznik nie byłby efektywny. Każda lista szczegółowych równoważności będzie dłuższa niż suma przekładów, które generuje. Zawiera bowiem 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[23]. Z tego powodu nie możemy zatrzymać się na punkcie (4).

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

Jest to najistotniejszy punkt procedury. Zdaję sobie sprawę, że może być na pierwszy rzut oka nieco niejasny. W związku z tym, po pierwsze, chciałbym podkreślić, że odwróciliśmy tu kolejność badania. Zamiast rozważać 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) najlepiej rozumieć jako zapytanie o pewien hipotetyczny scenariusz — wybieramy jakieś obliczenie i pytamy, jaki ciąg stanów przedmiotu (z wykrytego przez nas repertuaru stanów L, który możemy przy tej okazji poszerzyć) miałby mu odpowiadać? Choć wiemy, że jakiś ciąg na pewno będzie mu odpowiadał (ponieważ 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? Dopiero te pytania są dla nas interesujące, ponieważ

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

to właśnie 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 w ogólnych zdaniach pozwalających na zastąpienie szczegółowych korelacji (które będą z tych ogólniejszych zdań wyprowadzalne). Po drugie, poszukujemy w obiekcie „wbudowanych sekwencji", czyli takich sekwencji stanów, które są zdeterminowane samą jego budową. Jeżeli tylko zdarzy się znaleźć taką stałą sekwencję f_1-f_2…f_k[24], 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 na wprowadzenie skrótów do naszego podręcznika przekładu, przekształcając go tym samym w podręcznik efektywny i zamknięty. Pozwala 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 fizycznymi danego przedmiotu, to przedmiot ten zawiera strukturę obliczeniową.

5. ZAKOŃCZENIE

Przejdźmy do trzech wątpliwości, które mogły nasunąć się w trakcie lektury poprzednich części. 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 różniące się warianty. Wtedy od razu wykrylibyśmy regularności, stworzylibyśmy odpowiednie uogólnienia i uzyskalibyśmy dzięki temu efektywny podręcznik. Podręcznik ten mógłby zostać również uznany za zamknięty, wystarczy tylko, że tak dobierzemy kolejne obliczenia, by nie pojawiały 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 poprosilibyśmy o pokazanie, że owa ściana nadal spełnia podane kryterium. Pamiętajmy, że chcąc zdyskredytować nasze kryterium, sceptyk musi twierdzić, że udało mu się pokazać, iż stworzony efektywny i zamknięty pod-

[24] Zapis ten rozumieć należy następująco: gdy przedmiot znajduje się w stanie opisanym jako f_1, to następnie zawsze znajduje się w ciągu stanów f_2…f_k.

ręcznik przekładu pozwala przypisać ścianie realizację dowolnego obliczenia, nie może się więc od naszego żądania wykręcić. Aby choćby przez chwilę wyglądało na to, że jakiś zupełnie losowo dobrany przedmiot spełnia nasze kryterium, zestaw korelowanych obliczeń musi być odpowiednio spreparowany. Im bardziej spreparowany zestaw, tym łatwiej go podważyć, wskazując obliczenie odbiegające od wykorzystanego schematu i prosząc o jego przetestowanie.

Po drugie, mogłoby się też zdarzyć, że sami uleglibyśmy złudzeniu, iż jakiś przedmiot zawiera struktury obliczeniowe, ponieważ 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. 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 hipotezami empirycznymi. Zawsze może się zdarzyć, że potwierdzimy hipotezę, ponieważ trafiliśmy na szczególny zestaw obserwacji, a nasza metodologia nie ostrzegła nas przed jego szczególnością. Jest to cena, którą trzeba płacić, gdy chce się zastosować uogólnienia indukcyjne. 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 przedmiot albo ma poszukiwaną własność, albo jej nie ma i nie zależy to od naszych zabiegów interpretacyjnych. Gdy 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 skrajnej, humowskiej wersji. Załóżmy bowiem, jak to robiliśmy w części 2, że odkrywamy, iż mamy do czynienia z kosmicznym zrządzeniem losu i że badany przedmiot przypadkowo pozwalał na wyróżnienie takiego zbioru stanów, który umożliwił mu w tym czasie spełnić nasze kryterium. Okazuje się, że w wypadku naszego kryterium nie ma żadnych wątpliwości co do tego, że w tym czasie był to po prostu przedmiot zawierający struktury obliczeniowe. Jest tak dlatego, że kryterium to pozwala na wykrycie poszukiwanej własności, ale nie przesądza o tym, co właściwie jest tą własnością. Jest nią po prostu taka budowa przedmiotu (niezależnie od jej szczegółów), która pozwala na przypisanie mu dowolnego obliczenia za pomocą efektywnego i zamkniętego podręcznika przekładu. Jeżeli dany przedmiot taką budową się przez jakiś czas charakteryzował, to nie ma powodów, 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 przedmiot 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 to, że każdy przedmiot fizyczny podlega prawom przyrody, nie sprawia, że charakteryzuje się wystarczającą liczbą regularno-

ści, by możliwe było skonstruowanie efektywnego i zamkniętego podręcznika przekładu korelującego jego stany z dowolnym obliczeniem. Aby odeprzeć 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 wypadku niektórych przedmiotów, aby przypisać im zdolności obliczeniowe, musimy uciekać się do sztuczek takich jak definicja przez alternatywę z dowodu Putnama. Omawiana procedura pozwala na zdemaskowanie tych chwytów. Nawet jeśli uda się nam dzięki wykryciu naturalnych regularności przedmiotu 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 części 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 przedmiot, 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 je za kompromitujący kontrprzykład (Copeland 1996).

Raz jeszcze podkreślmy, że głównym problemem, który należało rozwiązać nie jest niejasność czy nieostrość podziału na przedmioty posiadające strukturę obliczeniową i ich pozbawione, lecz 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, jako że ich rozumowania opierają się na możliwości wprowadzania dodatkowych konwencji, a nie da się tego zrobić bez wydłużania podręcznika przekładu[25].

Podsumowując, jeśli wykorzystamy 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 procedura przypisywania przedmiotowi rozmaitych obliczeń może różnie wyglądać. W pewnych wypadkach będzie ona żmudnym rejestrem mapowań jednych stanów na drugie, a w innych, dzięki pewnym regularnościom, jedynie zbiorem reguł, które po-

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

zwolą generować taką korelację, jaką tylko zechcemy. Nawet jeśli w wypadku jakiegoś konkretnego obliczenia nie możemy mieć pewności, czy przedmiot faktycznie je realizuje, czy też jedynie tak się nam go szczęśliwie udało opisać, to ujawniająca się w opisanej procedurze uderzająca elastyczność niektórych przedmiotów jest autentyczną cechą, która im przysługuje. Jest to istotne, ponieważ z cechą tą można wiązać w filozofii umysłu spore nadzieje. Związek między posiadaniem struktur obliczeniowych a posiadaniem zdolności poznawczych jest bowiem ideą tak znaną, że nie trzeba jej tu 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[26].

BIBLIOGRAFIA

Chalmers D. (1996), *Does a Rock Implement Every Finite-State Automaton?*, „Synthese" 108(3), 309-333.

Chalmers D. (2010), *Świadomy umysł*, Warszawa: Wydawnictwo Naukowe PWN.

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

Copeland B. J. (1996), *What is Computation?*, „Synthese" 108(3), 335-359.

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

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

Fodor A. J. (1975), *The Language of Thought*, Cambridge (MA): Harvard University Press.

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

Peacocke C. (1999), *Computation as Involving Content. A Response to Egan*, „Mind and Language" 14(2), 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" 85(1), 93-115.

Piccinini G. (2008), *Computation without Representation*, „Philosophical Studies" 137(2), 205-241.

Piccinini G. (2011), *Information Processing, Computation, and Cognition*, „Journal of Biological Physics" 37(1), 1-38

Putnam H. (1979), *The Nature of Mental States* [w:] *Mind, Language and Reality. Philosophical Papers, Volume 2*, Cambridge: Cambridge University Press, 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" 64(3), 21-37.

Searle R. J. (1999), *Umysł na nowo odkryty*, Warszawa: Państwowy Instytut Wydawniczy.

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

[26] Kryterium to nie stanowi jednak po prostu wyjaśnienia tej tezy. Jest ogólniejsze, ponieważ nie zakłada się w nim niczego na temat samego sposobu wykonywania obliczenia. Rozwinięcie tych uwag wykracza poza ramy artykułu, wypada jednak wspomnieć o tym, że pankomputacjonizm nie jest jedynym problemem klasycznego komputacjonizmu jako stanowiska w filozofii umysłu.