Fragment książki Pawła Stacewicza
„Pojęcia jako funkcje decyzyjne. Zagadnienia informatyczne, metodologiczne i filozoficzne”
(Oficyna Wydawnicza PW, rok wyd. 2021)

5.1. Metafora i problem czarnej skrzynki

(…)

Koncepcję *czarnej skrzynki* wprowadzili do nauki cybernetycy i psychologowie behawioryści, zajmujący się odpowiednio: ogólną teorią sterowania i komunikacji w układach przetwarzających informacje oraz teorią zachowania zwierząt i ludzi. Ci drudzy podnosili szczególnie mocno problem zasadności pewnych tradycyjnych metod badania umysłu, takich jak introspekcja i wieńczący ją werbalny opis postrzeganych świadomie stanów umysłowych (Skinner 1938). Metody te uznawali za mało wiarygodne, ponieważ ich wyniki są zawsze obciążone czynnikami subiektywnymi, a przez to niemierzalnymi, takimi jak nastawienia, oczekiwania czy chwilowe nastroje (a mówiąc ogólniej: cały historyczny bagaż świadomych i nieświadomych przeżyć konkretnego człowieka). Metafora czarnej skrzynki miała zredukować umysł do tego, co dobrze opisywalne, mierzalne i zewnętrzne, a więc do układu relacji między dochodzącymi do organizmu bodźcami i obserwowalnymi reakcjami. Przykładowo: zamiast posługiwać się nieprecyzyjnym i subiektywnym opisem przeżycia typu „X odczuwa silny ból” behawiorysta użyłby sformułowań zdających sprawę z wyników konkretnych eksperymentów – sformułowań typu „Przy takim a takim natężeniu bodźca B, zarejestrowano taki a taki zestaw mierzalnych reakcji organizmu X-a”[1]. Intencja behawioralnej redukcji była proeksplanacyjna: wyjaśnienia w kategoriach postrzeganym introspekcyjnie stanów uznawano za słabsze (mniej naukowe) niż wyjaśnienia w kategoriach korelacji między układami mierzalnych bodźców i reakcji. Wyróżniając określone typy bodźców i reakcji, a także postulując pewien matematyczny opis relacji między nimi, np. układ implikacji, behawioryści przekształcali metaforę w model (nie wychodząc jednak poza postulat redukcji opisu do tego, co fizykalne i mierzalne). Cybernetyka zaś stosowała tego

[1] Filozoficznym odpowiednikiem podejścia behawioralnego w psychologii był program behawioryzmu logicznego w filozofii umysłu. Program ten postulował opis zjawisk umysłowych bez używania jakichkolwiek predykatów „mentalnych”, odnoszących się do wewnętrznych (mentalnych czy świadomych) stanów podmiotu. Predykaty te traktowano co najwyżej jako pewne określenia skrótowe, które należy rozwinąć, a tym samym precyzyjnie określić, za pomocą zdań o znaczeniu czysto fizykalnym, dotyczących zatem obserwowalnych i mierzalnych stanów, zdarzeń czy procesów (por. (Ryle 1970)).

rodzaju charakterystyki i modele na poziomie bardziej ogólnym: nie tylko do opisu ludzi czy zwierząt, lecz do opisu wszelkich systemów przetwarzających informacje i wchodzących w interakcję ze swoim środowiskiem (w tym: sztucznych; por. (Ashby 1961)). Warto podkreślić, że przy obydwu podejściach, behawioralnym i cybernetycznym, mianem czarnej skrzynki określano obiekt modelowany (np. umysł lub organizm), a nie sam model. Preferowana przez cybernetyków metoda czarnej skrzynki polegała na dochodzeniu do modelu poprzez rejestrowanie i analizowanie zewnętrznych zachowań obiektu modelowanego[2].

Wraz z rozwojem technik algorytmicznych – przynależnych już do informatyki, a nie cybernetyki – eksplanacyjny potencjał pojęcia i metody czarnej skrzynki zaczął się wyczerpywać. Wyjaśnienia w postaci „zewnętrznych" korelacji między wielkościami rejestrowanymi na wyjściu i wyjściu przestały wystarczać. Metody algorytmiczne, a zwłaszcza symboliczne, znane z badań nad sztuczną inteligencją, **zapewniały cenny poznawczo wgląd w możliwe związki między wejściem a wyjściem**. Przykładowo: jeśli były to metody oparte na logice, pozwalały ujawnić przejrzysty znaczeniowo łańcuch wnioskowań prowadzących od wejściowych przesłanek do wyniku. Mówiąc zaś ogólniej, sam algorytm generowania wyników, poprzez swą intersubiektywnie dostępną strukturę, pozwalał określić (a w pewnych przypadkach: zrozumieć) pewne istotne elementy relacji wejście-wyjście.

W takiej sytuacji określenie „czarna skrzynka" stało się synonimem układu, który generuje poprawne (lub: akceptowalne) wyniki, ma też pewną intersubiektywnie dostępną „zawartość", nie zapewnia jednak (i nie przedstawia) zrozumiałych dla człowieka wyjaśnień uzyskiwanych wyników. Analizując taki układ, możemy uzyskać konkretną wiedzę „jak" (np. jak dochodzi do podjęcia decyzji czy rozwiązania problemu), nie możemy jednak zdobyć odpowiednio czytelnej wiedzy „dlaczego". W przypadku układów tego rodzaju określenie „czarna skrzynka" dotyczy zatem struktury wyjaśnień, a nie struktury relacji wiążącej wejście

[2] Rozumowanie w kategoriach czarnej skrzynki znajdujemy w idei słynnego testu Turinga, który dotyczy stwierdzania hipotetycznej inteligencji maszyn na podstawie oceny werbalnych zachowań tychże (Turing 1950). Zgodnie z tą ideą maszynę należy uznać za inteligentną wówczas, gdy generowane przez nią odpowiedzi na pytania są w dostatecznie dużym stopniu nieodróżnialne od wypowiedzi ludzi. O behawioralnym charakterze testu decyduje fakt, ż porównując maszynę z człowiekiem abstrahuje się z jednej strony, od jakichkolwiek jego przeżyć czy stanów świadomych, z drugiej strony zaś, od znajomości jakichkolwiek stanów wewnętrznych maszyny. Ocenia się tylko i wyłącznie zewnętrzne wyniki tych stanów, czyli obserwowalne zachowania językowe.

z wyjściem. Choć ta druga jest znana, to sama w sobie nie prowadzi do uchwytnego dla człowieka (również konstruktora czy programisty) układu wyjaśnień[3].

Dobrym przykładem czarnych skrzynek w sensie algorytmicznym i eksplanacyjnym zarazem (nie zaś cybernetycznym czy behawioralnym) są *sztuczne sieci neuronowe*. Algorytmy kontrolujące ich sposób działania, w tym uczenia się, są precyzyjnie określone, a co za tym idzie całkowicie przejrzyste co do swojej struktury (przynajmniej dla osób zajmujących się profesjonalnie takimi układami). Ponadto, na każdym etapie działania sieci istnieje możliwość dokładnego odczytu konkretnych wartości wag międzyneuronalnych i sygnałów wyjściowych poszczególnych neuronów. Mimo to, wiedza płynąca ze znajomości wspomnianych algorytmów i odczytów nie przekłada się wprost na wysoko-poziomowe uzasadnienia, które wyjaśniałyby, w języku stosowanym do opisu powierzonego sieci zadania, dlaczego w konkretnym przypadku sieć dochodzi do takiego a nie innego wyniku. Przykładowo: pewna sieć wspomagająca lekarzy może rozpoznawać prawidłowo choroby, czyli generować prawidłowe diagnozy na podstawie przedstawianych jej objawów. Mimo to, ani znajomość jej struktury wewnętrznej (rozkładu i wartości wag), ani znajomość reguł przetwarzania sygnałów przez pojedyncze neurony, ani znajomość algorytmu zmiany wag podczas wcześniejszego treningu sieci, nie pozwalają sformułować wprost, bez użycia specjalnych algorytmów „interpretujących", uzasadnienia diagnozy w języku zrozumiałym dla lekarza czy pacjenta (języku objawów, jednostek chorobowych, związków przyczynowo-skutkowych między objawami itp.).

Mając na uwadze powyższy przykład, możemy objaśnić w sposób ogólny sens przesunięcia znaczeniowego, jakie dokonało się w sposobie rozumienia terminu „czarna skrzynka". W ujęciu tradycyjnym, cybernetyczno-behawioralnym, za czarne skrzynki uznawano układy o nieprzejrzystej poznawczo strukturze wewnętrznej. Wskutek tejże nieprzejrzystości układy takie opisywano tylko i wyłącznie w kategoriach korelacji między danymi wejściowymi (bodźcami) i wyjściowymi (reakcjami) – uznając takie opisy za jedyne wartościowe naukowo wyjaśnienia tego, jak dany układ działa. W ujęciu nowszym, zgodnym ze współczesnymi dokonaniami informatyki, **znaczeniowy punkt ciężkości określenia „czarna skrzynka" przesunął się w kierunku siły i struktury wyjaśnień**. Chodzi jednak o

[3] Warto dopowiedzieć, że kwestia wyjaśnialności i zrozumiałości generowanych przez system wyjaśnień jest względna. Znaczy to, że zależnie od oczekiwań danej grupy osób (tzw. interesariuszy) wymagane są inne typy wyjaśnień. Przykładowo: programiści są zainteresowani czytelnością układu instrukcji programu i znajomością ogólnego algorytmu, który program realizuje (zakłada się przy tym, że znają oni określony język programowania), użytkownicy natomiast oczekują wyjaśnień w kategoriach zależności między obiektami z dziedziny, której dotyczy problem rozwiązywany przez dany system (np. figurami geometrycznymi, jeśli program rozwiązuje problem geometryczny). Por. (Zednik 2019).

wyjaśnienia, które dotyczą nie sposobu działania systemu, lecz generowanych przezeń wyników (por. Creel 2020). Rozpiszmy tę konkluzję dokładniej.

O systemach informatycznych, czyli działających na podstawie pewnych algorytmów, nie można twierdzić zasadnie, że przypominają czarne skrzynki w sensie cybernetyczno-behawioralnym. Z zasady bowiem znamy strukturę wchodzącego w grę algorytmu i odpowiadającego mu programu (nawet jeśli ma ona ogromną złożoność), możemy także manipulować wewnętrznymi parametrami algorytmu i obserwować adekwatne do tych manipulacji zmiany w zachowaniu sterowanego algorytmicznie systemu[4]. Pozostaje to prawdą również w odniesieniu do takich systemów, których oprogramowanie powstaje lub podlega zmianom w drodze uczenia się, a więc dzięki zastosowaniu pewnego programu wyższego poziomu odpowiedzialnego za zmianę parametrów programu właściwego. W przypadku systemów uczących się również znamy właściwości obydwu algorytmów/programów, możemy dokonywać zmian ich wewnętrznych parametrów oraz obserwować efekty tych zmian. Ze strukturalnego i operacyjnego punktu widzenia nie są to zatem żadne nieprzejrzyste dla programisty (czy nawet: użytkownika) czarne skrzynki. Mimo to, z perspektywy oczekiwanych przez użytkownika wyjaśnień podejmowanych decyzji, systemy takie okazują się nieskuteczne, to znaczy ich sposób działania (a niekiedy też: uczenia się) nie zapewnia przejrzystej struktury wyjaśnień. I w tym właśnie sensie, sensie eksplanacyjnym, przyrównanie takich układów do czarnych skrzynek wydaje się zasadne.

Nieprzejrzystość poznawczą w sensie eksplanacyjnym przypisuje się najczęściej informatycznym *systemom uczącym się*, w tym sztucznym sieciom neuronowym (Zednik 2019). Oprogramowanie do automatycznego uczenia się nie jest jednak jedynym czynnikiem, który może prowadzić do efektu nieprzejrzystości. Co więcej, niektóre metody uczenia się, o ile są dostosowane do symbolicznych metod reprezentacji wiedzy, efektu tego nie powodują. Zestawmy w sposób niewyczerpujący różne możliwe powody tej niepożądanej w wielu sytuacjach właściwości.

1. *Złożoność strukturalna*. Zbyt duża złożoność układu przetwarzającego dane (np. ogromna liczba reguł w systemie eksperckim lub połączeń między neuronami w sieci neuropodobnej) może powodować trudności z wyodrębnieniem tych elementów, które wpłynęły istotnie na wygenerowanie konkretnego wyniku. Ponadto, jeśli układ wchodzi

[4] Z ogólniejszego punktu widzenia zaletą metody algorytmicznej – w odróżnieniu od metod angażujących niewerbalizowaną do końca intuicję – jest właśnie intersubiektywna dostępność i kontrolowalność schematu, według którego postępujemy. Te cechy powodują, że metodę algorytmiczną można łatwo zautomatyzować, np. za pomocą komputerów cyfrowych (Stacewicz 2016).

w dynamiczne interakcje ze swoim środowiskiem, w wyniku czego jego struktura wewnętrzna zmienia się, to identyfikacja elementów istotnych, dokonywana *post factum*, może być w ogóle nieosiągalna (Creel 2020).

2. ***Uczenie się.*** Jeśli system informatyczny doskonali swoje działanie (a niekiedy: jest tworzony od podstaw) na podstawie pewnego algorytmu uczenia się, to algorytm ten może powodować dodawanie do programu sterującego systemem pewnych technicznych, trudno-interpretowalnych parametrów, które mają na celu tylko i wyłącznie efektywne działanie systemu (np. dopasowanie jego sposobu działania do pewnych przykładowych par [dane, oczekiwany wynik]).
Efekt nieprzejrzystości jest tym większy, im szerzej na etapie uczenia się są wykorzystywane wybory losowe. Ich losowy charakter sprawia, że finalna struktura systemu kształtuje się w sposób nieprzewidywalny, a to powoduje, że nawet twórca systemu może nie rozumieć wpływu poszczególnych elementów na generowane przez system wyniki (Zednik 2019).

3. ***Naśladowanie natury.*** Za nieprzejrzystość systemu może odpowiadać fakt, że tworzy się go na wzór pewnych nie dość dobrze poznanych układów naturalnych, jak komórki biologiczne, mózgi, ewoluujące gatunki (itp…) – przyjmując po prostu, że układy takie działają w przyrodzie wystarczająco efektywnie. Ponieważ działanie tego typu układów (w tym: ludzkiego mózgu), a także wpływ ich wewnętrznych zmian na obserwowane zachowania, wciąż jest przedmiotem badań naukowych i daleko nam do pełnego zrozumienia wielu zależności, to systemy konstruowane na podobieństwo układów naturalnych dziedziczą niejako pierwotną nieprzejrzystość tych ostatnich. Obserwacja ta dotyczy, na przykład, pewnych typów sztucznych sieci neuronowych (Hinton 1989).

4. ***Operowanie na danych niepewnych*****.** Jeśli system operuje na danych, które są obarczone pewnymi stopniami niepewności, to procedura dochodzenia do finalnej decyzji, nawet w ramach symbolicznych metod reprezentacji wiedzy, jest bardzo nieintuicyjna. Stopnie (nie)pewności wpływają na decyzje cząstkowe w sposób nieoczywisty – np. zgodnie z arbitralnymi regułami logiki rozmytej (Klir, Yuan 1995). Wyjaśnienie ostatecznej decyzji, wybieranej spośród wielu hipotetycznych rozwiązań o różnych stopniach pewności, zależy od tego, jaki mechanizm propagacji niepewności

(statystyczny, oparty na logice wielowartościowej, oparty na logice rozmytej…) zastosowano. Wobec możliwości wyboru wielu rożnych mechanizmów wyjaśnienie ostatecznej decyzji nie jest jednoznaczne.

5. *Brak procedur translacji*. W przypadku wielu systemów (w tym: sztucznych sieci neuronowych) nie mamy do dyspozycji wiarygodnych metod translacji reguł niskiego poziomu (tworzonych często w drodze uczenia się) na przejrzyste znaczeniowo reguły symboliczne. Sytuacja taka występuje wówczas, jeśli uczenie się przebiega tylko i wyłącznie na poziomie operacyjnym, nie angażuje żadnych struktur symbolicznych wyższego poziomu, a polega na „dopasowywaniu" struktury operacyjnej (np. wag połączeń międzyneuronalnych wewnątrz sieci neuronowej) do decyzji wymuszanych przez nauczyciela lub środowisko[5].

[5] Przykładowo: system uczy się w taki sposób, aby odpowiadać określonymi sygnałami na pewne przykładowe wzorce (np. obrazy czy dźwięki) oraz podobnymi sygnałami na podobne wzorce; po zakończeniu treningu system reaguje prawidłowo, nie dostarcza jednak żadnych wyjaśnień symbolicznych.