„Teologia i Moralność", Volumen 18 (2023), numer 1 (33), s. 271-285
ISSN 1898-2964 eISSN 2450-4602

DOI: 10.14746/TIM.2023.33.1.16

Zachęta do ostrożności

Prompting for caution

DAMIAN SZCZĘCH
Katolicki Uniwersytet Lubelski Jana Pawła II, Wydział Filozofii, Polska
szczech.dam@gmail.com
https://orcid.org/0000-0003-2290-1726

Susan Schneider *Świadome Maszyny. Sztuczna inteligencja i projektowanie umysłów*, przeł. Joanna Bednarek, WN PWN, Warszawa 2021, ss. 249, ISBN: 978-83-012-1952-9

Abstract: The purpose of this review article is to present and comment on the contents of Susan Schenider's book *Artificial You: AI and the Future of Your Mind*. The applied research method is analysis (extracting terms and definitions from the text and reconstructing them) and comparison (cross-referencing the notions used by the author with those used in other works). The conclusions emphasize timeliness of the issues addressed by the author and soundness of her advice for caution when deciding about modifying one's own brain.

Keywords: consciousness; philosophy of mind; identity; artificial intelligence; brain emulation; transhumanism

Abstrakt: Celem artykułu recenzyjnego jest przybliżenie i skomentowanie treści zawartych w książce Susan Schneider *Świadome Maszyny*. *Sztuczna Inteligencja i projektowanie umysłów* wydanej w 2021 roku. Stosowaną metodą badawczą jest analiza (wydobycie z tekstu terminów i definicji oraz ich rekonstrukcja) oraz porównanie (zestawienie pojęć stosowanych przez autorkę z pojęciami używanymi w innych tekstach). Wnioski podkreślają aktualność problematyki podejmowanej przez autorkę oraz trafność jej zalecenia ostrożności podczas decydowania o modyfikacji własnego mózgu.

Słowa kluczowe: świadomość; filozofia umysłu; tożsamość; sztuczna inteligencja; emulacja mózgu; transhumanizm

Wstęp

Susan Schneider (ur. 1968) jest amerykańską filozofką i kognitywistką. Pracuje na Florida Atlantic University. Jest założycielką i dyrektorką *Center for the Future Mind*, które poszukuje głębszego naukowego i filozoficznego zrozumienia nowych technologii oraz przyszłości inteligencji[1]. Jest współredaktorką *The Blackwell Companion to Consciousness* (2007), redaktorką *Science Fiction and Philosophy* (2009) oraz autorką *The Language of Thought: A New Philosophical Direction* (Schneider 2011). Recenzowana praca jest jej najnowszą książką. Oryginał ukazał się w 2019 roku nakładem Princeton University Press.

Omawiany tekst ma być „wstępną eksploracją przestrzeni projektowania umysłu" (s. 214), rodzajem przewodnika czy intelektualnej mapy zagadnień związanych z naturą umysłu, świadomością, tożsamością oraz badaniami nad Sztuczną Inteligencją (SI). Jest to owoc prac autorki nad komputacyjną teorią umysłu oraz wyraz jej antymaterialistycznych poglądów na temat jego natury. Schneider prowadzi rozważania z filozoficznego punktu widzenia, zajmując się rozmaitymi pomysłami dotyczącymi projektowania i modyfikacji umysłu. Choć uważa się za transhumanistkę, zaleca ostrożność przy podejmowaniu decyzji o ulepszaniu mózgu. Pisze: „Byłabym szczęśliwa, gdyby książka stała się choćby najskromniejszym wkładem w udzielenie młodemu pokoleniu pomocy w radzeniu sobie ze wspomnianymi wyzwaniami technologicznymi i filozoficznymi" (s. 224), przy czym „młodym pokoleniem" są raczej jej dzieci, którym dedykowany jest tekst. Cytowany fragment dobrze odzwierciedla charakter książki i koncepcję stojącą za jej stworzeniem.

Książka składa się z: wprowadzenia, ośmiu rozdziałów, podsumowania i załącznika w postaci *Deklaracji transhumanistycznej*, zawierającej główne założenia transhumanizmu, które zostały przedstawione w siedmiu dość ogólnych punktach. Autorka dokonuje przeglądu zagadnień i „oprowadza" po prezentowanych koncepcjach. Podejmuje się dyskusji z niektórymi przytaczanymi stanowiskami (fizykalizmem, teorią wzoru, naturalizmem biologicznym), wskazując ich słabe strony, oraz ocenia je z punktu widzenia swoich poglądów, przy czym prezentuje dość wyważone stanowisko. Podkreśla konieczność prowadzenia społecznego dialogu w takich kwestiach jak ulepszanie umysłów i rozwój SI.

1. W poszukiwaniu inteligencji ludzkiego typu

Dyskusje wokół natury umysłu oraz inteligencji obecnie wydają się nierozstrzygalne (Boden 2020, 160). Autorka pomija poszukiwanie rozwiązania pro-

[1] Opis ze strony: https://www.fau.edu/future-mind/ (dostęp: 27.09.2022).

blemu relacji ciało-umysł i podejmuje spekulacje wokół możliwości zaistnienia świadomości w systemach SI. Zakłada przy tym, że przynajmniej niektóre SI będą przypominały człowieka w swoim działaniu i że dzięki temu będzie możliwe ich badanie pod kątem świadomości, która ma funkcjonować tak, jak świadomość ludzka. Takie założenie nie jest oczywiste i wymaga uzasadnienia. Systemy, które powstaną, mogą być całkowicie odmienne od biologicznych istot (Bostrom 2016, 55), a więc także od ludzi. Testy proponowane przez autorkę w dalszej części tekstu obejmą jedynie SI podobną do nas. Jeżeli emulacja ludzkiego mózgu i stworzenie AGI (*Artificial General Intelligence*, czasem zwana „silną SI") są możliwe, to tym bardziej możliwe będzie stworzenie SI obejmującej mózgi szympansów czy innych zwierząt, które także mogą być świadome.

W rozdziale 1. Schneider przedstawia swe główne tezy: SI jest obecna niemal wszędzie, a silnej SI możemy się spodziewać już wkrótce. AGI może powstać „już za kilkadziesiąt lat" (s. 20) a z jej rozwojem wiążą się obawy o zmiany, jakie w nas spowoduje. Biorąc pod uwagę ogromne fundusze przeznaczane na badania, autorka stwierdza, że SI ogólnego zastosowania nie jest wyłącznie domeną sci-fi. Warto przy tej okazji wspomnieć, że charakterystycznym żartem towarzyszącym rozmowom o fuzji jądrowej i innych przełomowych projektach jest to, że dana technologia będzie dostępna „już za 30 lat". W przypadku SI występuje podobne zjawisko – perspektywa trzech dekad pojawia się zaskakująco często. Vernor Vinge wprowadził pojęcie „osobliwości technologicznej" oraz jako pierwszy zapowiedział jej nadejście „w ciągu najbliższych 30 lat" (O'Connell 2020, 90; Larson 2021, 46). Po nim wielu autorów przyjmowało podobną perspektywę czasową. Vinge w późniejszej rewizji tekstu doprecyzował, że chodzi o przedział między 2005 a 2030 rokiem i uznał go za rozsądny (Vinge 2003, 2). Raymond Kurzweil (2016) w dużej mierze dzięki swojej książce *Nadchodzi osobliwość* stał się najbardziej znanym popularyzatorem tej idei, bywa nawet z nią utożsamiany.

Jednym z częściej przytaczanych sposobów stworzenia AGI oraz osobliwości jest emulacja działania ludzkiego mózgu. Nawet jeśli założymy, że jest możliwa, to jeszcze przyjdzie nam na nią długo czekać. Obecnie wielkim wyzwaniem jest szczegółowe zeskanowanie całego mózgu oraz odpowiednie przetworzenie powstałej przy tej okazji ogromnej liczby danych. Niedawno sukcesem zakończył się projekt mapowania fragmentu wynoszącego około 0,08% objętości ludzkiego mózgu. Autorzy wskazują, że było to możliwe dzięki znaczącemu postępowi techniki w ostatnich latach i że „mapowanie eksaskalowe", które pozwoli na odwzorowanie całego mysiego mózgu, będzie w naszym zasięgu w ciągu najbliższych dziesięciu lat (Shapson-Coe et al. 2021, 19). Jednak ludzki mózg jest ponad dwa tysiące razy większy od mysiego. Emulacja jego działania nadal znajduje się w sferze *science-fiction*.

Bostrom i Sandberg (2008, 81) oszacowali, że emulacja „powinna być możliwa" przed 2050 rokiem, przy założeniu, że przyjęte modele są wystarczające. Ich raport pochodzi z 2008 roku i nie doczekał się rewizji w świetle nowych osiągnięć nauki i techniki. Eth, Foust i Whale (2013, 149) wskazują rok 2063 jako prawdopodobny dla pojawienia się działającej emulacji. Niemniej jednak osobliwość ma przekroczyć jej możliwości, więc nadejdzie jeszcze później. Niektórzy co prawda przyznają, że prognozowanie jej nadejścia przed końcem bieżącego stulecia jest nierealne (Boden 2020, 168). Inni z kolei zauważają „żenujący rozdźwięk między faktycznymi postępami" w dziedzinie powstania SI a wizjami futurystów (Larson 2021, 49). Postaci, którymi inspiruje się Schneider, wykazują wielki optymizm co do możliwości stworzenia sztucznych umysłów, zakładając, że mózg da się opisać poprzez względnie prosty algorytm lub „program mózgowy" (Kurzweil 2018, 242), dzięki któremu „wymagania obliczeniowe niezbędne dla tego projektu są niemal w zasięgu ręki" (Kurzweil 2018, 262). Obecnie nie sposób stwierdzić, które z tych przewidywań są słuszne albo przynajmniej prawdopodobne.

1.1. Próba określenia pojęcia świadomości

W 2. rozdziale autorka wprowadza problem kluczowy dla całości książki: określa świadomość jako „poczucie bycia sobą". Z takiego ujęcia wynika posiadanie perspektywy pierwszoosobowej i poczucia odrębności, których przejawy ma badać test zaproponowany w dalszej części książki. Odwołuje się do „trudnego problemu świadomości" sformułowanego przez Davida Chalmersa (s. 31) oraz wykorzystuje znany eksperyment myślowy, jakim jest chiński pokój (John Searle) – ma być on ilustracją łagodnej wersji argumentu naturalistów biologicznych. Jest to argument, w którym pewien człowiek (lub sam Searle) zostaje zamknięty w pokoju, do którego wrzucane są kartki z pytaniami pisanymi po chińsku. Searle nie zna chińskiego, ale jest w posiadaniu książki zawierającej instrukcje, dzięki którym odpowiada na otrzymane pytania w taki sposób, że stwarza wrażenie, iż faktycznie umie się posługiwać tym językiem. Człowiek w pokoju nie rozumie pytań ani odpowiedzi, tylko mechanicznie przepisuje symbole z książki. System opisywany przez Searle'a nie ma (według autora argumentu) zdolności zrozumienia lub nie rozumie podejmowanych działań, a zatem nie może mieć świadomości. Co ciekawe, krytycy argumentu Searle'a zarzucają mu stosowanie „filozoficznych sztuczek" polegających na niepoprawnym traktowaniu człowieka jako jedynego elementu systemu, podczas gdy pod uwagę należy brać całość: człowieka, pokój i książkę z odpowiedziami (Kurzweil 2018, 356). Taki system musi w pewnym stopniu rozumieć chiński, ponieważ, zgodnie z założeniem przyjętym przez

Searle'a, udzielane odpowiedzi są przekonujące dla badaczy. Innym rodzajem krytyki jest zarzut umieszczenia systemu (robota) „w próżni" oraz przyjęcia zbyt ogólnikowych (*vague*) definicji znaczenia i wiedzy semantycznej (Okoro 2022). Autorka do takich zarzutów dodaje przekonanie, że system chińskiego pokoju jest zbyt prosty, aby mógł być świadomy (s. 38), a zatem argumentacja Searle'a nie może dotyczyć zaawansowanych, złożonych systemów SI. Ostatecznie Schneider stwierdza, że spór między zwolennikami możliwości powstania inteligencji na podłożu innym niż biologiczne a tzw. naturalistami biologicznymi, głoszącymi, iż inteligencja może powstać jedynie na podłożu biologicznym, nie ma odpowiedniej podbudowy teoretycznej – nie istnieją przekonujące argumenty na rzecz naturalizmu ani przeciwko niemu (s. 38).

W rozdziale 3. autorka rozważa możliwość powstania świadomych SI i postuluje rozwój „inżynierii świadomości". Twierdzi, że musimy rozumieć, jak budować świadomość, bo moglibyśmy stworzyć ją przypadkiem, co prowadziłoby do powstania swoistej postaci niewolnictwa. Oczywiście możemy chcieć stworzyć ją celowo, uważając, że świadoma SI będzie bezpieczniejsza dla człowieka (s. 63). Swoje rozważania kontynuuje w kolejnym rozdziale, w którym postuluje wypracowanie szeregu testów pozwalających rozpoznać, czy mamy do czynienia ze świadomą SI; jej zdaniem bowiem, zbudowanie jednego uniwersalnego testu jest niemożliwe. Zaproponowany przez nią test ma badać występowanie domniemanych przejawów świadomości fenomenalnej, tj. „odbierania świata od wewnątrz" (s. 79), i opiera się na dość nieoczywistych założeniach. Zagadnienie świadomości fenomenalnej jest „prawdziwym filozoficznym grzęzawiskiem" (Boden 2020, 138), choć Schneider zdaje się je traktować jako ogólnie zrozumiałe. Większość badaczy i praktyków skupia się na świadomości funkcjonalnej (którą Schneider nazywa maszynową), ignorując przy tym świadomość fenomenalną (Boden 2020, 139, 144). Próba rozwiązania twardego problemu świadomości jest dość odważnym posunięciem i zasługuje na pochwałę za dociekliwość i wytrwałość. Jednak wymaga solidnego teoretycznego opracowania, którego tutaj zabrakło. Autorka nie przedstawia, jak rozumie subiektywne doświadczenia, które chce badać poprzez proponowane testy. Przyjmuje wiele założeń (m.in. to, że ludzie potrafią sobie wyobrazić oddzielenie umysłu od ciała czy też zamianę ciał), których nie wspiera żadnymi argumentami. Proponowane przez nią badanie opiera się głównie na specyficznych konstrukcjach semantycznych, co sprawia, że jest ono jedynie nieco zmodyfikowaną wersją testu Turinga, która ma badać zachowania wskazujące na występowanie świadomości u maszyn (s. 86). Schneider twierdzi, że nieprzejście testu przez daną SI nie oznacza, iż nie jest ona świadoma, natomiast przejście testu miałoby się wiązać z koniecznością objęcia takiego systemu prawną ochroną (s. 87). Drugim rodzajem badania ma być zaproponowany przez autorkę „test implantu" polegający na wykorzystaniu hipotetycznego

implantu mózgowego, który miałby zastąpić część mózgu odpowiadającą za świadomość. Skuteczne (tj. niepowodujące utraty samoświadomości u pacjenta) zastosowanie takiej technologii wskazywałoby, że możliwe jest stworzenie świadomych SI opartych na mikrochipach wykorzystanych w tym implancie. Powodzenie takiej procedury będzie również argumentem za możliwością istnienia maszyn o świadomości podobnej do zwierząt, które nie posługują się językiem takim jak ludzie i nie mogłyby przejść innych wersji testu.

2. Rozumienie pojęć „osoba" i „tożsamość"

W następnej części Schneider podejmuje rozważania nad pojęciem osoby. Słusznie podkreśla, że pojęcie to jest rozumiane w różnych filozofiach rozmaicie. Ma to duże znaczenie dla kwestii ulepszania umysłu oraz przenoszenia go na inne podłoże. Na przykład, jeśli rację mają fizykaliści, to w rezultacie zmiany nośnika umysłu dana osoba przestawałaby istnieć. Z kolei transhumaniści przyjmują „teorię wzoru", wedle której umysł jest to „program działający na sprzęcie mózgu" (s. 117). Zgodnie z takim rozumieniem o byciu tym oto konkretnym człowiekiem stanowi konkretna konfiguracja psychiczna. Warto dodać, że wiąże się z tym także przyjmowana przez nich teza o niezależności od substratu, zgodnie z którą tożsamość danej rzeczy (osoby) nie zależy od sposobu fizycznej realizacji tej rzeczy. Transhumaniści wnioskują, że odtworzenie takiej konfiguracji na innym podłożu (nośniku) oznaczałoby odtworzenie danej osoby. Według nich „stany mentalne mogą występować w wielu rodzajach fizycznych substratów" (Bostrom 2003, 244). Zatem możliwe ma być powstanie świadomego systemu na podłożu niebiologicznym. Co ciekawe, jeśli prawdą jest, że „do generowania subiektywnych doświadczeń wystarczy odpowiednio szczegółowe odtworzenie struktury procesów obliczeniowych ludzkiego mózgu" (Bostrom 2003, 244), to dokładne zrozumienie działania ludzkiego mózgu może nie być konieczne – wystarczy odtworzyć jego strukturę w przestrzeni wirtualnej. Autorka przy tym nie czyni uwagi, że analogia umysłu jako programu komputerowego wprowadza wiele niuansów i dodatkowe komplikacje – w rzeczywistości dany program działa tylko na urządzeniach konkretnego typu, z myślą o których został stworzony. Aby można było go uruchomić na urządzeniach innego typu, konieczne jest użycie tzw. emulatorów, czyli programów „udających" inną maszynę niż ta, na której są uruchomione. Stąd pojawił się postulat stworzenia emulacji ludzkiego mózgu, uruchomienie bowiem algorytmu stworzonego na bazie skanu biologicznego mózgu będzie wymagało systemu „udającego" połączenia nerwowe i środowisko, z którym mózg się komunikuje. Ludzki „[m]ózg jest najmniej poznaną i najbardziej skomplikowaną strukturą we wszechświecie", co gorsza „wciąż

nie powstała jedna naukowa koncepcja funkcjonowania tego narządu" (Krajczyńska 2010). Emulacja wymaga stworzenia systemu, który będzie bardziej złożony (skomplikowany) niż sam mózg.

Rozdział 6. jest krytyką teorii wzoru. Autorka odwołuje się do eksperymentu myślowego z powieści *Mindscan*, w której główny bohater chce przenieść swój umysł do robota, czego skutkiem jest jedynie stworzenie kopii protagonisty. Taka kopia ma identyczne wspomnienia i osobowość co biologiczny człowiek oraz, podobnie jak on, rości sobie prawo do bycia oryginalnym sobą – kwestia tożsamości każdego z nich jest problematyczna. Sprawa dodatkowo się komplikuje, gdy zaczynają powstawać kolejne egzemplarze tego człowieka. Autorka nazywa to „problemem podwojenia" (s. 125) i wskazuje, że wynika on bezpośrednio z przyjętej teorii wzoru. Jako możliwe rozwiązanie proponuje uwzględnienie wymogu „ciągłości czasoprzestrzennej" (s. 128), co powoduje uznanie odrębnych tożsamości oryginału i kopii. Stwierdza, że w takim ujęciu sporządzenie duplikatu umysłu nie może być nazwane udoskonaleniem, jak próbują to robić fuzjooptymiści. Swoje stanowisko nazywa „zmodyfikowaną teorią wzoru" (s. 130) i zauważa, że nadal wiąże się z nim problem modyfikacji – trudność w określeniu, co i w jakim stopniu można zmienić (ulepszyć), by nadal pozostać sobą (s. 135). W związku z tym proponuje stanowisko „metafizycznej pokory" i traktowanie postulatu radykalnego ulepszania z pewnym sceptycyzmem. Postuluje podjęcie publicznej dyskusji i edukację społeczeństwa, by to potrafiło podjąć świadomą decyzję w kwestii hipotetycznego ulepszania. Z tekstu nie wynika, czy Schneider dopuszcza możliwość, że gdy (jeśli) ulepszenia staną się dostępne, to wszelkie wątpliwości zostaną rozstrzygnięte poprzez obserwację praktycznych skutków zastosowania takiej technologii, czy też zakłada, iż metafizyczne problemy nawet wtedy pozostaną nierozwiązane.

Autorka wydaje się nie zauważać innego ontologicznego problemu, wiążącego się z krytykowaną przez nią teorią wzoru. Mianowicie tego, że transhumaniści nie odróżniają tożsamości numerycznej od jakościowej – jeśli przyjmiemy takie rozróżnienie, to oczywiste się staje, że stworzenie wiernej kopii danej rzeczy nie wpływa na rzecz będącą wzorem. Kopia jest jakościowo tożsama (identyczna) z oryginałem, lecz ma numerycznie odrębną tożsamość. Teoria wzoru jest absurdalnym ujęciem problemu tożsamości, co postaram się zilustrować przykładem: egzemplarze tej samej książki opuszczające drukarnię mają identyczną treść, ale są numerycznie odrębne. Pojęcie tożsamości stosowane przez transhumanistów każe nam jednak uznać, że wszystkie egzemplarze są tą samą książką, ponieważ wszystkie mają ten sam wzór (tę samą treść). Proponowana przez autorkę modyfikacja w postaci wprowadzenia wymogu ciągłości czasoprzestrzennej rozwiązuje problem odrębności (tak samo jak uznanie pojęcia tożsamości numerycznej), lecz nie rozwiązuje pro-

blemu modyfikacji – nie dostarcza odpowiedzi na pytanie: Jakie zmiany są dopuszczalne. Pojęcie tożsamości jakościowej wiąże się z podobnym problemem – odpowiada na pytanie o dopuszczalność zmian (dozwolone są tylko nieistotne zmiany), ale nie dostarcza kryteriów oceny ich istotności.

Stosowane przez autorkę pojęcie „przeniesienia umysłu" to często nadużywana przez transhumanistów metafora. Z cytowanego na stronach 120-121 opisu przykładowej procedury (skanowania struktury mózgu i jej cyfrowego odtwarzania) jasno wynika, że nie zachodzi tutaj żadne przeniesienie. Biologiczny mózg (umysł) pozostaje tam, gdzie był – w cyfrowym świecie pojawia się jedynie jego odwzorowanie, kopia, reprodukcja. Oczywiste jest, że cyfrowa wersja nie jest tym samym, co biologiczny oryginał. Jednak transhumaniści twierdzą co innego. Autorka wydaje się nie zauważać tej kwestii – traktuje pojęcia przenoszenia i kopiowania umysłu jako synonimy. Wydaje się, że bardziej adekwatnym terminem na określenie takiej procedury byłaby „konwersja". Jest to pojęcie również zaczerpnięte z nauk informatycznych i oznacza zmianę formy danych (umysłu) bez utraty ich treści (tożsamości). Podejrzewam, że w koncepcji *mind uploading* powinno właśnie iść o konwersję (przekształcenie) umysłu, a nie o utworzenie duplikatu, którego użyteczność jest dyskusyjna.

3. Krytyka założeń przyjmowanych przez transhumanistów

W siódmym rozdziale, odwołując się do rozważań na temat istnienia innych cywilizacji we wszechświecie, Schneider przedstawia problem kontroli – jako dość prymitywne i mniej inteligentne istoty nie będziemy w stanie zapanować nad znacznie bardziej rozwiniętymi od nas superinteligencjami. Z tego powodu autorka krytykuje program SETI, próbujący nawiązać kontakt z ewentualną cywilizacją pozaziemską. Określa takie działania jako lekkomyślne (s. 154) oraz jako „igranie z ogniem" i zaleca „intelektualną pokorę" (s. 156). Zaraz potem jednak snuje przypuszczenie, że superinteligencje z kosmosu nie podejmą prób kontaktu z nami. Następnie streszcza przedstawioną przez Nicka Bostroma (2016) teorię dotyczącą możliwych ich postaci, po czym snuje rozważania na temat istnienia pozaziemskich superinteligencji, których konstrukcja miałaby być inspirowana istotami biologicznymi, wydaje się, że ta część jest inspirowana innym tekstem Bostroma (2003) na temat możliwości stworzenia przez wysoce rozwiniętą obcą cywilizację symulacji komputerowej obejmującej cały świat, co jednak nie zostało wskazane w treści ani w bibliografii.

Rozdział 8. rozpoczyna się od przedstawienia historii Kim Suozzi – dziewczyny chorej na raka mózgu, która poddała się zabiegowi krioprezerwacji (zamrożenia) głowy (s. 176) z nadzieją, że w przyszłości możliwe będzie jej

wyleczenie i przywrócenie do życia w postaci fizycznej lub cyfrowej. Wychodząc od tej historii, autorka krytykuje transhumanistyczne rozumienie umysłu, wskazując, że teoria wzoru jest nieużyteczna przy rozstrzyganiu dylematu psychofizycznego: W jaki sposób subiektywne wrażenia i stany umysłu wiążą się z fizycznym ukształtowaniem mózgu? Stanowisko transhumanistów nie prowadzi do uzyskania odpowiedzi na to pytanie. Pojmowanie umysłu jako specyficznego rodzaju programu komputerowego wikła się w kolejne metafory i problemy podobne do tych związanych z teorią wzoru. Zwolennicy teorii programu twierdzą m.in., że zachowanie odpowiedniej liczby „kopii zapasowych" może zapewnić człowiekowi nieśmiertelność (s. 184). Co prawda ani autorka, ani pozostali transhumaniści nie wskazują takiego źródła inspiracji, ale ten sposób postrzegania dążenia do nieśmiertelności jest zaskakująco podobny do tego, który prezentuje w książkach Joanna Rowling (2006). W jej powieściach pojawiają się wzmianki o horkruksach – przedmiotach, w których czarodziej mógł zakląć fragment swojej duszy, co zapewniało mu nieśmiertelność. Zniszczenie jednego z nich nie powodowało śmierci ich twórcy, jeśli istniał którykolwiek z pozostałych. Ciekawym spostrzeżeniem Schneider jest to, że teoria duszy zapewniałaby znacznie lepsze wyjaśnienie transferu umysłów (s. 129) niż wysiłki zwolenników teorii wzoru. Jeśliby bowiem przyjąć, że to właśnie dusza odpowiada za świadomość oraz że da się ją przenieść, to cała teoria byłaby znacznie mniej skomplikowana. Autorka jasno wskazuje, że w jej ocenie postrzeganie umysłu jako programu jest błędne (s. 184). Następnie przedstawia kilka filozoficznych stanowisk dotyczących dylematu psychofizycznego (panpsychizmu, dualizmu własności i substancji, fizykalizmu, idealizmu) i przechodzi do szczegółowego omówienia problemów związanych z teorią programu. Podkreśla, że pomysł przenoszenia umysłu jest oparty na „błędnych podstawach pojęciowych" (s. 211).

4. Problemy związane z błędnym rozumieniem pojęć „inteligencja" i „świadomość"

Schneider twierdzi, że to nieznajomość filozoficznych zagadnień dotyczących natury umysłu powoduje problemy wyznaczające główny temat książki. Można ująć je w formie pytań: „Czym jest świadomość i jak można ją określić?" oraz „Jakich konsekwencji można się spodziewać w wyniku stworzenia świadomej SI?". Za warte rozważań uznaje dwa aspekty związane z pytaniem o konsekwencje: scenariusze związane ze stworzeniem świadomych maszyn oraz scenariusze dotyczące radykalnych ulepszeń mózgu. Pytanie: „Czy stworzenie świadomej SI jest możliwe?" uważa za zasadniczo nierozstrzygalne. Głoszone przez nią rychłe nadejście silnej SI może w tym kontekście ozna-

czać, że taka SI może nie mieć świadomości, a mimo to nadal przewyższać człowieka w niemal wszystkich aspektach, albo że pytanie o możliwość stworzenia świadomej SI będzie dało się rozstrzygnąć dopiero wtedy, gdy stwierdzimy jej istnienie.

Warto zaznaczyć, że przypisanie maszynom inteligencji na ludzkim poziomie miałoby poważne konsekwencje praktyczne (Boden 2020, 158). Mogłoby na przykład doprowadzić do całkowitej zmiany postrzegania ich roli w ludzkim społeczeństwie, a nawet do powstania praw robotów (Gellers 2021, 163--164) analogicznych do praw zwierząt czy praw natury. Ciekawym założeniem przyjętym przez autorkę jest to, że SI ma (lub może mieć) jakiś rodzaj umysłu, skoro może być świadoma. Schneider nie wyjaśnia natomiast, dlaczego posiadanie umysłu jest warunkiem koniecznym świadomości ani dlaczego jej rozwinięcie (wytworzenie) jest równoznaczne ze staniem się osobą, co prowadzi do zyskania ochrony prawnej. Innym wątkiem wartym rozwinięcia jest związek świadomości z wolnością i moralnością – czy świadoma maszyna musi być wolna, a tym samym musi stać się podmiotem moralności? Kolejnym problemem byłoby określenie, czy świadoma SI jest w jakimś sensie żywa (jeśli bycie świadomym wymaga bycia żywym) i czy SI zyskuje jakieś prawa z samego faktu bycia żywą. Niestety, autorka nie rozważa takich konsekwencji. Transhumanizm zazwyczaj nie podejmuje analiz rzeczywistych (pozatechnicznych) skutków wprowadzania rozwiązań technicznych do życia społecznego. Zamiast tego skupia się na analizie ilościowej, opartej na danych dotyczących dotychczasowego tempa rozwoju technologii spodziewanych przyszłych sukcesów, co prowadzi do przeszacowania wartości i faktycznego oddziaływania proponowanych rozwiązań. Niemal zupełnie pomija się tutaj analizę jakościową. Autorka, jako konsekwentna transhumanistka, również unika rozważań na temat skutków takich ulepszeń w życiu osobistym i społecznym.

4.1. Istotność kwestii świadomości SI

Powszechnie stosowane pojęcie SI odnosi się do racjonalności intelektualnej, pomijając inteligencję emocjonalną i społeczną (Boden 2020, 167). Schneider nie definiuje samej SI – czy jest to program (ciąg instrukcji, szczegółowy plan działania), algorytm (abstrakcyjny opis sposobu działania), rodzaj urządzenia czy też „magiczne coś z komputera". Często traktuje SI-program oraz SI-urządzenie (robot) jako zamienne pojęcia. Fragment ze strony 23.: „W dodatku SI można ściągnąć i zainstalować na wielu urządzeniach w wielu miejscach jednocześnie", wskazuje, że ma na myśli raczej rodzaj programu, jednak nie wyraża tego *explicite*. Nie jest także jasne, czy posiadanie fizycznych komponentów uważa za element konieczny powstania świadomej maszyny.

Schneider wskazuje też na rosnącą popularność idei łączenia ludzi z SI jako sposobu na uniknięcie zastąpienia pracowników przez algorytmy oraz jako szansy na uzyskanie nieśmiertelności i superinteligencji. Podkreśla, że ryzyko powodowane przez takie połączenie jest ogromne. Jeśli okazałoby się, że stworzenie świadomej SI jest niemożliwe, to scalenie ludzkich mózgów z urządzeniami elektronicznymi mogłoby doprowadzić do powstania „zombie" (pozbawionej świadomości imitacji człowieka), co byłoby równoznaczne ze śmiercią człowieka poddanego takiemu zabiegowi (s. 17). Nie wyjaśnia natomiast, dlaczego łączenie ludzi z SI miałoby się odbywać wyłącznie fizycznie – wykorzystując narzędzia oparte na SI (np. programy do tłumaczeń), w pewnym sensie łączymy się z nią. Nie musimy dokonywać fizycznej modyfikacji naszych ciał, aby stać się bardziej wydajnymi w pracy. Należy tutaj zaznaczyć, że często podkreślana autonomia maszyn napędzanych SI wcale nie musi świadczyć o ich inteligencji – termostat w żelazku, samozamykacz drzwi czy wyłącznik nadprądowy również działają autonomicznie, jednak w żadnym razie nie są inteligentne. Nie do końca jasny jest także cel łączenia ludzkich mózgów z niebiologicznymi komponentami. Wielu transhumanistów uważa elektronikę za lepszą od tkanki biologicznej, a zatem traktuje jej wszczepianie w ciała jako rodzaj ulepszenia. Podkreśla się, że sygnały elektroniczne są transmitowane znacznie szybciej od biochemicznych (Lovelock 2019, 81). Zatem zastąpienie części lub całości ludzkiego mózgu jego elektronicznym odpowiednikiem mogłoby być korzystne. Nie jest jednak pewne, czy szybsze myślenie (przetwarzanie sygnałów) mogłoby sprawić, że będę bardziej inteligentny, czy też będę tak samo głupi, tylko szybciej. Pomocne byłoby tu określenie, czym jest inteligencja, niestety, autorka nie wskazała, jaką definicję stosuje. Skupia się przede wszystkim na możliwym wpływie implantów na utratę bądź zachowanie świadomości.

Kwestia określenia świadomości SI jest kluczowa także ze względów etycznych i prawnych. Autorka nie rozwija natomiast idei, że posiadanie świadomości jest równoznaczne z byciem osobą, a więc z byciem podmiotem praw, choć wydaje się to zakładać, twierdząc, że gdyby udało się przypadkiem stworzyć świadomy system, to jego wykorzystanie do pracy mogłoby być równoznaczne z niewolnictwem, a więc byłoby przestępstwem (s. 62). Założenie to ujawnia się także przy konstrukcji testów. Schneider proponuje takie testy na określenie świadomości, które w przypadkach wątpliwych orzekałyby na korzyść maszyn. Mimo rozmaitych zastrzeżeń jest bardzo optymistycznie nastawiona co do spodziewanego kierunku rozwoju: „[w]ciąż mam nadzieję, że nowe technologie zapewnią nam znaczne przedłużenie życia, pomogą w rozwiązaniu problemów niedoboru zasobów i chorób [sic!], a nawet udoskonalą nasze życie umysłowe, jeśli tego zechcemy" (s. 27). Proponuje stanowisko metafizycznej pokory, które „zakłada, że postęp możliwy będzie dzięki publicznej

dyskusji, nie zaś teoretyzowaniu" (s. 138). Niestety, nie wyjaśnia pojęcia „metafizyczna pokora" i nie osadza go we właściwym kontekście filozoficznym – zostało ono potraktowane jako potoczne i ogólnie zrozumiałe. A przecież wydaje się, że publiczna dyskusja musi być oparta na dobrze uzasadnionych teoriach dotyczących świadomości i SI. Schneider zresztą uważa, że problemy teoretyczne oraz brak konkretnych definicji sprawiają, iż „fuzjooptymiści i transhumaniści nie mają dobrych argumentów na poparcie tezy, że powinniśmy się ulepszać" (s. 140), a przez to nie odpowiadają na pytanie: „Dlaczego powinniśmy się ulepszać?". Zarzut jest słuszny, ale autorka popełnia ten sam błąd, który zarzuca transhumanistom – nie przedstawia satysfakcjonujących definicji świadomości, inteligencji, SI, AGI. Polski tytuł książki sugeruje, że tekst zawiera jakąś koncepcję, jednak argumentacja Schneider opiera się na stwierdzeniu, że nie dysponujemy odpowiednimi określeniami świadomości oraz inteligencji i próbujemy znaleźć pewne ich wskaźniki, tj. przejawy pierwszoosobowej perspektywy. Autorka wskazuje, że jest wiele koncepcji osoby, transhumaniści arbitralnie wybierają jedną, a ich wybór wiąże się z szeregiem problemów. Radzi powstrzymanie się od podejmowania decyzji na podstawie tak ubogiej metafizyki. Rada wydaje się rozsądna, ale sama autorka nie wykłada swej metafizyki ani nie pokazuje, jaka metafizyka byłaby wystarczająca, by podejmować decyzje co do ulepszania lub nieulepszania się (czy nawet ludzkości jako całości). Mimo to dobrze się stało, że została uwypuklona rola filozofii, a zwłaszcza metafizyki, w dyskusji nad tym zagadnieniem.

Proponowany w rozdziale 4. test na świadomość SI jest dość naiwny. Autorka twierdzi, że konieczne jest opracowanie szeregu testów na świadomość SI, ponieważ stworzenie uniwersalnej wersji jest niemożliwe (s. 74). Wyróżnia świadomość fenomenalną i funkcjonalną (maszynową), wskazując, że ta druga jest typowa dla SI niemających subiektywnych doświadczeń (s. 77). Opisuje świadomość fenomenalną jako „odbieranie świata od wewnątrz" i formułuje odważne stwierdzenie, że wiąże się z tym możliwość wyobrażenia sobie „przynajmniej z grubsza": oddzielenia umysłu od ciała, życia pośmiertnego, reinkarnacji oraz eksterioryzacji (s. 79). Na podstawie tych możliwych wyobrażeń konstruuje test mający określić, czy badana SI jest świadoma. Pytania są jednakże zbyt abstrakcyjne i wielu ludzi (w tym ja) będzie z pewnością miało problem z wyobrażeniem sobie reinkarnacji czy życia po śmierci. Test opiera się na konstrukcjach semantycznych charakterystycznych dla tzw. kultury zachodniej – problem z odpowiedzią na takie pytania mieliby przedstawiciele kultur, w których te pojęcia nie występują. Schneider postuluje także zadawanie pytań filozoficznych, co również mija się z celem, ponieważ odpowiedź na wiele z nich wymaga filozoficznego treningu, by w ogóle móc zrozumieć problem. Niewerbalna wersja testu opiera się na obserwacji zachowań, które również są uwarunkowane kulturowo. Ponadto autorka nie-

jawnie zakłada możliwość umieszczenia SI w fizycznym ciele, aby dało się obserwować jej zachowania. Z kolei „test implantu" opiera się na bardzo optymistycznym założeniu co do możliwego rozwoju technologii. Tak skonstruowane testy na świadomość SI badają tylko świadomość ludzkiego rodzaju. Nie obejmą świadomości podobnej do zwierzęcej, a tym bardziej nie obejmą świadomości zupełnie innego typu. Jeśli maszyny mają nas przewyższyć we wszystkim, to całkiem możliwe, że osiągną wyższe, pełniejsze niż my poziomy świadomości, których istnienia nie będziemy w stanie stwierdzić. Należy jednak podkreślić, że sama autorka przedstawia swoje pomysły jako propozycję rozwiązania, a nie jedyny słuszny kierunek postępowania. Zaproponowane przez nią testy mogą być dobrym punktem wyjścia do dyskusji nad badaniem świadomości u maszyn, a nawet u zwierząt – Schneider tego nie wskazuje, ale do badania zwierzęcej świadomości również potrzebny jest jakiś rodzaj testów. Jeśli propozycje podobne do jej pomysłów okazałyby się skuteczne, to mogłyby także znaleźć zastosowanie w badaniu maszyn napędzanych SI.

Podsumowanie

Praca ma charakter popularnonaukowy, na co wskazują liczne odwołania do popkultury i narracja prowadzona pierwszoosobowo z bezpośrednim zwracaniem się do czytelnika. Tekst został napisany językiem publicystycznym z elementami terminów technicznych. Książka realizuje postawione przez autorkę cele. Zawiera przegląd ważnych stanowisk i problemów związanych ze świadomością i naturą umysłu. Pojawia się też wiele spekulacji będących ilustracjami przedstawianych zagadnień, a czasem przybierających formę eksperymentów myślowych. Rozważania ograniczają się do badania możliwości i nie wykazują typowego dla transhumanistów i „futurologów" myślenia życzeniowego.

Książka jest napisana na ogół prostym i bardzo przystępnym językiem, choć w kilku miejscach pojawia się żargon, np. „udoskonalone technologicznie minikolumny neuronalne" (s. 144). Tekst zawiera także uproszczenia, przez co niektóre (na szczęście nieliczne) fragmenty stają się niezrozumiałe dla czytelnika nieobeznanego z fizyką. Na przykład czytamy: „Idea osobliwości została zaczerpnięta z matematyki i fizyki, zwłaszcza z pojęcia czarnej dziury. Czarne dziury to «osobliwe» obiekty czasoprzestrzenne, czyli takie, w których normalne prawa fizyki się załamują. Technologiczna osobliwość ma, analogicznie, spowodować niekontrolowany postęp technologiczny i przełomowe zmiany cywilizacyjne" (s. 21).

Autorka wielokrotnie zwraca się bezpośrednio do czytelnika. Polski przekład dość często zmienia narrację na bezosobową, co jest odstępstwem

od oryginalnego charakteru tekstu. Jest to widoczne nawet w tytule – oryginalny brzmi *Artificial You: AI and the Future of Your Mind* [Sztuczni wy: SI i przyszłość waszego umysłu], co dobrze koresponduje z zamysłem autorki, a co znikło z tłumaczenia. Angielski tekst jest także bardziej jednoznaczny w ocenie omawianych działań związanych z ulepszaniem. Niektóre odwołania bibliograficzne są niepełne. W niektórych miejscach pojawiają się dosłowne tłumaczenia przysłów i powiedzeń, które są zupełnie niezrozumiałe dla polskiego czytelnika, np. „Trzeba pozwolić, by zakwitło tysiąc kwiatów" (s. 74).

Natomiast zalecenia ostrożności przy podejmowaniu decyzji o ulepszaniu czy tworzeniu świadomej SI są warte wzięcia pod uwagę, ponieważ przedstawione obawy związane z konsekwencjami czy to budowania świadomej SI, czy sprzęgania człowieka z silną SI są dobrze uzasadnione. Również postulowany przez Schneider dialog publiczny na ten temat wydaje się potrzebny – konsekwencje bowiem będą dotyczyć nie pojedynczych ludzi, ale wszystkich, także przyszłych pokoleń. Jednakże podjęcie takiego dialogu wymaga dużego wysiłku edukacyjnego, ponieważ zagadnienia związane z filozofią oraz nowymi technologiami nie należą do tzw. wiedzy powszechnej.

Nie ma wątpliwości, że warto zainteresować się polem analiz, w którym mieści się książka Schneider. Dobrego argumentu dostarczają aktualne dyskusje na temat SI. Oto w czerwcu 2022 roku pracownik koncernu Google, Blake Lemoine stwierdził, że wykorzystujący SI chatbot o nazwie LaMDA zyskał świadomość i zdolność odczuwania emocji. Opublikował nawet wywiad z systemem SI, co miało być dowodem na takie tezy. Google zaprzeczył, że system jest świadomy, a Lemoine ostatecznie został zwolniony z powodu publikacji firmowych dokumentów. Książka jest więc dobrą propozycją dla osób zainteresowanych omawianą problematyką, zwłaszcza jeśli poszukują tekstu przeglądowego, który nie zagłębia się w zbyt szczegółowe analizy. Praca stanowi wstęp do głównych zagadnień z tego pola i podkreśla ich znaczenie dla naszego życia dziś oraz w bliskiej przyszłości. Jest to także interesująca pozycja dla czytelników niemających wyrobionej opinii na temat modyfikacji mózgu – tekst będzie dla nich propozycją pewnej ostrożnej, lecz nieradykalnej postawy. Wydaje się jednak, że w celu uchwycenia myśli i stylu filozofowania Schneider warto konfrontować polskie tłumaczenie z angielskim oryginałem.

BIBLIOGRAFIA

Boden, Margaret. 2020. *Sztuczna Inteligencja. Jej natura i przyszłość*, tłum. Tomasz Sieczkowski. Łódź: Wydawnictwo UŁ.

Bostrom, Nick. 2016. *Superinteligencja. Scenariusze, strategie, zagrożenia*, tłum. Dorota Konowrocka-Sawa. Gliwice: Helion.

Bostrom, Nick i Anders Sandberg. 2008. *Whole Brain Emulation. A Roadmap*. Dostęp: 05.10.2022. https://www.fhi.ox.ac.uk/brain-emulation-roadmap-report.pdf.
Bostrom, Nick. 2003. Are We Living in a Computer Simulation?. *Philosophical Quarterly*, 53, 243--255.
Eth, Daniel, Juan-Carlos Foust i Brandon Whale. 2013. The Prospects of Whole Brain Emulation within the next Half-Century. *Journal of Artificial General Intelligence*, 4(3), 130-152. DOI: 10.2478/jagi-2013-0008
Gellers, Joshua. 2021. *Rights for Robots. Artificial Intelligence, Animal and Environmental Law*. Abington, New York: Routledge.
Krajczyńska, Ewelina. 2010. *Mózg najbardziej skomplikowaną strukturą we wszechświecie*. Dostęp: 01.10.2022. https://naukawpolsce.pl/aktualnosci/news%2C370904%2Cmozg-najbardziej-skomplikowana-struktura-we-wszechswiecie.html.
Kurzweil, Ray. 2018. *Jak stworzyć umysł. Sekrety ludzkich myśli ujawnione*, tłum. Katarzyna Zielińska. Białystok: Studio Astropsychologii.
Kurzweil, Ray. 2016. *Nadchodzi Osobliwość*, tłum. Eliza Chodkowska i Anna Nowosielska. Warszawa: Kurhaus Publishing.
Larson, Erik. 2021. *The Myth of Artificial Intelligence. Why Computers Can't Think the Way We Do*. Cambridge, London: The Belknap Press of Harvard University Press. DOI: 10.2307/j.ctv322v43j.
Lovelock, James. 2019. *Novacene. The Coming Age of Hyperintelligence*. Penguin Books.
Okoro, Angel. *Does the Chinese Room Thought Experiment disprove true AI and mind uploading?*. Dostęp: 27.09.2022. https://carboncopies.org/does-the-chinese-room-thought-experiment-disprove-true-ai-and-mind-uploading/.
O'Connell, Mark. 2020. *Być maszyną. Przygody wśród cyborgów, utopistów, hakerów i futurystów w ich skromnych staraniach, by rozwiązać problem śmierci*, tłum. Aleksandra Małecka. Warszawa: Wydawnictwo Krytyki Politycznej.
Rowling, Joanne. 2006. *Harry Potter i Książę Półkrwi*, tłum. Andrzej Polkowski. Poznań: Media Rodzina.
Schneider, Susan. 2021. *Świadome Maszyny. Sztuczna inteligencja i projektowanie umysłów*, tłum. Joanna Bednarek. Warszawa: WN PWN.
Schneider, Susan. 2011. *The language of thought: A New Philosophical Direction*. MIT Press.
Science Fiction and Philosophy. From Time Travel to Superintelligence, red. Susan Schneider. 2009. Wiley-Blackwell.
Shapson-Coe, Alexander et al. 2021. *A connectomic study of a petascale fragment of human cerebral cortex*. Dostęp: 30.09.2022. DOI: 10.1101/2021.05.29.446289 https://www.biorxiv.org/content/10.1101/2021.05.29.446289v4.full.pdf.
The Blackwell Companion to Consciousness. red. Max Velmans i Susan Schneider. 2007. Malden, Oxford, Carlton: Blackwell Publishing.
Vinge, Vernor. (2003). *Technological Singularity*. Dostęp: 30.09.2022. http://cmm.cenart.gob.mx/delanda/textos/tech_sing.pdf.

Damian Szczęch – student V roku filozofii na KUL, interesuje się filozofią nauki i techniki, zwłaszcza w zakresie badań wpływu techniki na funkcjonowanie człowieka oraz wpływu osiągnięć technonauki na postrzeganie człowieka i jego miejsca w świecie.