PIOTR PRZYBYSZ

WIĘCEJ NIŻ „TRUDNY PROBLEM ŚWIADOMOŚCI": O MOŻLIWOŚCI POWSTANIA ŚWIADOMEJ SZTUCZNEJ INTELIGENCJI

ABSTRACT. Piotr Przybysz, *Więcej niż „trudny problem świadomości": o możliwości powstania świadomej sztucznej inteligencji* [More than the „hard problem of consciousness": On the possibility of creating conscious artificial intelligence] edited by Sławomir Leciejewski, „Człowiek i Społeczeństwo" vol. LVIII: *Społeczny wymiar rewolucji informatycznej* [The social dimension of the information technology revolution], Poznań 2024, pp. 55–87, Adam Mickiewicz University. ISSN 0239-3271, https://doi.org/10.14746/cis.2024.58.4.

The aim of this paper is to review the most important questions and problems concerning the emergence of conscious AI. In the paper, I point out three such key questions and problems: (1) how to recognize that AI has acquired consciousness? (2) how can conscious AI emerge? and (3) what properties can conscious AI have? I argue that these problems cannot currently be solved on the basis of purely experimental, computer science, and engineering approaches, because the path to this leads through areas marked out by previous philosophical and general theoretical reflection on the subject.

Keywords: conscious AI, consciousness, philosophy of mind

Piotr Przybysz, Uniwersytet im. Adama Mickiewicza w Poznaniu, Wydział Filozoficzny, ul. Szamarzewskiego 89AB, 60-568 Poznań, e-mail: przybysz@amu.edu.pl, https://orcid.org/0000-0001-8184-3656.

Wprowadzenie. Świadomość maszyn – ciągle więcej pytań niż odpowiedzi

Pytanie o możliwość pojawienia się inteligentnych maszyn, które zarazem byłyby świadome i zdolne do subiektywnych odczuć, przestaje być kojarzone wyłącznie ze sferą utopijnych wizji i quasi-naukowych fikcji, a staje się powoli elementem technologicznego projektu przyszłości. Mimo że w bieżących dyskusjach nad rozwojem sztucznej inteligencji (SI) bardziej interesuje nas to, czy roboty i inne zautomatyzowane systemy zabiorą nam pracę, czy zatopimy się w wirtualnym świecie cyfrowej rozrywki oraz czy autonomiczne systemy nie zajmą naszego miejsca i nie zbuntują się przeciwko człowiekowi, to temat uzyskania przez SI świadomości uporczywie powraca w tego typu dyskusjach jako zagadnienie tyleż zagadkowe, co niepokojące.

Sami naukowcy, a zwłaszcza filozofowie oraz specjaliści od sztucznej inteligencji, są w tej sprawie wyraźnie podzieleni. Niektórzy z nich uważają, że nawet przyszłe najbardziej inteligentne maszyny nigdy nie staną się świadome, gdyż bycie świadomym jest cechą przypisaną wyłącznie do naturalnych biologicznych istot, lub – co na to samo wychodzi – że ścieżka rozwoju sztucznej inteligencji może biec niezależnie od ścieżki ewolucji świadomości. Uzasadnieniem dla tego rodzaju sceptycyzmu może być na przykład przekonanie, że systemom sztucznej inteligencji świadomość nie jest do niczego potrzebna i dlatego jest mało prawdopodobne, aby pojawił się odpowiednik presji selekcyjnej na jej wyłonienie się w długofalowym procesie rozwoju SI (zob. np. Dreyfus, 1992; Searle, 1995; Seth, 2021; Landgrebe i Smith, 2023)[1]. Inni naukowcy są bardziej optymistyczni w tej sprawie i przypuszczają, że świadomość, jako węzłowy element procesu przetwarzania informacji, mogłaby się pojawić wraz z odpowiednio zaawansowanymi systemami SI, obecnie kojarzonymi na przykład z ogólną sztuczną inteligencją lub superinteligencją (zob. Tegmark, 2019; Duch, 2024). Przewidywania na ten temat są jeszcze mało konkretne, pozbawione precyzyjnej

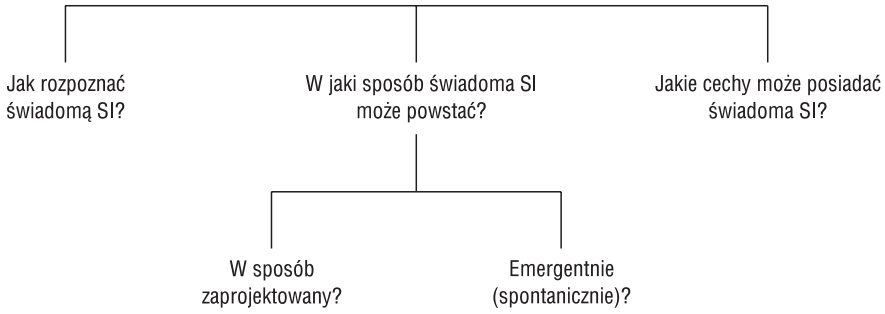
[1] Nie bardzo też wiadomo, dlaczego samym ludziom miałoby zależeć na powstaniu świadomej SI. Nie da się w tym przypadku wykluczyć np. chęci „zabawy w Boga" i przekraczania kolejnych granic w poznaniu i tworzeniu. Może też wchodzić w grę przekonanie, że najefektywniejsza współpraca z maszynami i robotami będzie miała miejsce wtedy, gdy zostaną one obdarzone zdolnością odczuwania i samoświadomością. N. Humphrey bierze również pod uwagę to, że obliczu możliwości zagłady ludzkości w dłuższej perspektywie w interesie człowieka mogłoby leżeć przekazanie świadomości robotom i sztucznej inteligencji (zob. Humphrey, 2022: 211–213).

lokalizacji czasowej i karmią się dość niejasnymi wyobrażeniami na temat tego, jak mógłby wyglądać świat, w którym obok ludzi i zwierząt pojawiłyby się syntetyczne istoty dysponujące świadomością i podmiotowością.

Mimo że nie dysponujemy obecnie żadną technologią budowy syntetycznej świadomości, wyraźnym przeobrażeniom uległa treść „naczelnych opowieści” wykorzystywanych w kulturowych i popularnonaukowych narracjach na ten temat. Początkowo miały one formę mrocznych legend i mitów o ożywionych obiektach – na przykład o Golemie, glinianym posągu ożywionym przez praskiego rabina Jehudę Löw ben Becalela za pomocą kładzionej mu na języku lub czole kartki z magicznym słowem-kodem „Emet” (*hebr*. prawda). Następnie opowieść o świadomych i trudno odróżnialnych od człowieka androidach zagościła w kulturze popularnej – na przykład w znanych filmach *Terminator* Jamesa Camerona czy *Blade Runner* Ridleya Scotta, które kierunkowały wyobraźnię współczesnych ludzi na roztrząsanie utopijnych lub dystopijnych scenariuszy przyszłego rozwoju społeczeństwa technologicznego. W końcu temat ten podjęli filozofowie i naukowcy, którzy eksplorowali logiczne i technologiczne możliwości powstania świadomych i inteligentnych maszyn, zastanawiali się nad podobieństwami pomiędzy biologicznymi mózgami i przyszłymi sztucznymi umysłami, a także próbowali przewidzieć etyczne konsekwencje pojawienia się „nowego gatunku” inteligentnych istot (Yudkowsky, 2004; Bostrom, 2016; Russell, 2019; Christian, 2020; Zybertowicz, 2024). Obecnie niebagatelną rolę w inicjowaniu myślenia o obdarzonych świadomością algorytmach odgrywają sukcesy technologii generatywnej SI oraz dużych modeli językowych (*large language models* – LLM), których demonstrowane umiejętności i tkwiące w nich możliwości dalece przewyższyły to, czego spodziewali się po nich nawet sami ich twórcy. To zaś skłania do spekulacji, czy aby na dalszych etapach ich rozwoju nie pojawią się u nich oznaki świadomości i subiektywnych przeżyć (zob. np. Meissner, 2020; Esmaeilzadeh i Vaezi, 2021; Chalmers, 2023; Duch, 2024)[2].

[2] W kwietniu 2023 roku stu kilkudziesięciu naukowców podpisało się pod listem otwartym wzywającym do intensyfikacji badań nad możliwością powstania świadomej SI. Autorzy listu piszą w nim m.in.: „Wraz z rozwojem sztucznej inteligencji rośnie potrzeba, aby opinia publiczna, instytucje społeczne i organy władzy wiedziały, czy i w jaki sposób systemy sztucznej inteligencji mogą stać się świadome, aby *rozumiały* tego konsekwencje, oraz aby skutecznie mogły odnieść się do etycznych, dotyczących bezpieczeństwa i społecznych następstw związanych z ogólną sztuczną inteligencją” (Open letter „The Responsible Development of AI Agenda Needs to Include Consciousness Research”, https://amcs-community.org/open-letters/).

W niniejszym tekście – w kontrze do nazbyt optymistycznego technooptymizmu części badaczy biorących udział w debatach na temat SI – spróbuję pokazać, że mimo niewątpliwych technologicznych postępów w badaniach i wdrożeniach sztucznej inteligencji, do jakich doszło w ostatnich latach, nie można sądzić, że zagadnienie syntetycznej świadomości[3] stało się *już* przedmiotem ścisłych czy naukowych rozstrzygnięć oraz inżynieryjnego planowania. Droga do tego jest jeszcze daleka i wiedzie przez obszary wytyczone przez uprzednią filozoficzną i ogólnoteoretyczną refleksję na ten temat. W artykule wskazuję na trzy takie główne, i w jakimś sensie fundamentalne, problemy ze świadomą SI, które decydują, że temat ten – przynajmniej w możliwym do przewidzenia czasie – pozostanie wysoce problematycznym zagadnieniem, uwikłanym w rozliczne spekulacje, gdzie podejście naukowo-inżynieryjne sąsiadować ciągle będzie z filozoficznymi analizami pojęciowymi i eksperymentami myślowymi. Problemy te to:



Il. 1. Trzy fundamentalne kwestie dotyczące świadomej SI: (1) problem rozpoznania świadomej SI, (2) problem sposobu powstania świadomej SI oraz (3) problem własności świadomej SI. Problematyka dotycząca sposobu powstania świadomej SI analizowana jest zwykle w ramach dwóch scenariuszy: projektowego i emergentnego

Źródło: opracowanie własne

[3] Uwaga terminologiczna: w celu uniknięcia wchodzenia w spory definicyjne i zbyt szczegółowe eksplikacje pojęciowe w artykule posługuję się ogólnym terminem „sztuczna/syntetyczna świadomość" odnoszonym zbiorczo do „maszyn", „maszyn inteligentnych", „systemów sztucznych", „algorytmów" oraz „sztucznej inteligencji". Te ostatnie terminy używane są w artykule zamiennie.

Pierwszy z wyróżnionych przeze mnie problemów dotyczy naszych ograniczonych możliwości rozpoznania, czy urządzenia wyposażone w sztuczną inteligencję mają stany świadome. Rzecz w tym, że jeśli nawet udałoby się człowiekowi stworzyć takie maszyny lub algorytmy, to nie tak łatwo byłoby rozstrzygnąć, czy rzeczywiście uzyskały one świadomość, czy jedynie symulują i zachowują się *jak gdyby* były świadome. Drugi z problemów dotyczy niejasności w sprawie, jak mogłoby dojść do powstania świadomej SI. Zwykle wskazuje się w tym wypadku na dwa nieco odmienne scenariusze, którym należy się osobno przyjrzeć (zob. il. 1). Z jednej strony jest to scenariusz zakładający tak zwaną „emergencję świadomości”. Jest on związany z podzielanym przez wielu badaczy przekonaniem, że świadomość w maszynach może pojawić się w sposób spontaniczny i niekontrolowany przez człowieka na odpowiednio zaawansowanym etapie rozwoju SI. Z drugiej strony – w opozycji do podejścia emergentnego – badacze stawiają pytanie, jak należałoby zaprojektować i zbudować taką świadomą sztuczną inteligencję (np. czy w analogii do biologicznego „oprogramowania”, jakim dysponują ludzkie i zwierzęce mózgi?). I wreszcie, trzeci, kluczowy problem dotyczy tego, czy potrafimy z dzisiejszej perspektywy – w sytuacji braku technologii budowy syntetycznej świadomości – przewidzieć jej przyszły kształt oraz cechy, jakie będzie posiadała, choćby na podstawie wywiedzenia ich z ogólnych praw obowiązujących w świecie fizycznym i poprzez porównanie ze świadomością ludzi.

W artykule przyjmuję, że wymienione powyżej problemy wyznaczają trzy najważniejsze filozoficzne i teoretyczne obszary, po których porusza się obecnie nasze myślenie na temat świadomej SI. Wprawdzie nie pozwalają one rozstrzygnąć, czy ostatecznie syntetyczna świadomość się pojawi, jaki w końcu kształt przybierze, jak ją rozpoznać oraz jak ją zaprojektować, ale za to umożliwiają identyfikację i oświetlenie fundamentalnych filozoficznych i teoretycznych kwestii spornych, jakie się z tym wiążą. Celem artykułu jest przyjrzenie się tym trzem problemom w systematyczny sposób.

Należy jeszcze wspomnieć o dwóch sprawach. Po pierwsze, w artykule pominąłem wiele pokrewnych kwestii, na przykład dotyczących moralnych, prawnych, społecznych i egzystencjalnych konsekwencji, jakie miałoby dla człowieka pojawienie się nowego gatunku istot świadomych. Wraz z problematyką dopasowania sztucznej inteligencji do koegzystencji z człowiekiem (tzw. *alignment problem*, zob. np. Christian, 2020) należą one do odrębnego działu zagadnień, które wymagałyby osobnego potraktowania. Po drugie, w tekście ograniczyłem się do przywoływania przykładów jedynie z obszaru generatywnej sztucznej inteligencji, a konkretnie – wspomnianych już

dużych modeli językowych i opartych na nich chatbotów, spośród których największy rozgłos zdobył ChatGPT. Pozostałe technologie generatywnej SI (np. służące do generowania obrazów), a także odmiennie rodzaje sztucznej inteligencji (np. systemy dyskryminatywnej SI czy tzw. predyktywna SI – zob. np. Narayanan i Kapoor, 2024) zostały w tekście pominięte.

„Duch w maszynie": świadomość fenomenalna vs. świadomość funkcjonalna

Zanim przejdę do głównego tematu niniejszego tekstu, chciałbym, choćby krótko, zająć się wskazaniem, które z licznych określeń i naukowych ujęć świadomości mogą okazać się pomocne do opisania poruszanych tutaj problemów. Sprawa nie jest oczywista, gdyż już sama natura i funkcje ludzkiej i zwierzęcej świadomości wyrosłej na podłożu biologicznym – nawet bez wchodzenia w tematykę sztucznych i syntetycznych jej odpowiedników – budzą liczne kontrowersje.

W ostatnich kilkudziesięciu latach poczyniono spory postęp w rozumieniu funkcji świadomości i jej biologicznych, mózgowych podstaw (zob. Koch, 2008; Dehaene, 2023), przebadano liczne przypadki deficytów świadomości i jej nietypowych odmian (zob. Weiskrantz, 1997; Metzinger, 2018), zaproponowano nowatorskie hipotezy na temat początków świadomości na ziemi, na temat obecności świadomości w świecie zwierzęcym (zob. Feinberg i Mallatt, 2016; Ginsburg i Jablonka, 2019) oraz, co z tym związane, próbowano uzgadniać koncepcje świadomości ze scenariuszami ewolucyjnymi (zob. np. LeDoux, 2020). W tym samym czasie zaproponowano również przynajmniej kilka „kanonicznych" teorii świadomości, które wyznaczają obecnie horyzont myślenia w tej dziedzinie. Są to między innymi koncepcja świadomości fenomenalnej (Chalmers, 2010), teoria świadomości jako globalnej przestrzeni roboczej (Baars, 1998), teorie świadomości jako stanów wyższego rzędu (*higher-order-thought* – Rosenthal, 2005) czy teoria zintegrowanej informacji (Massimini i Tononi, 2018)[4].

[4] Istnieje wiele alternatywnych zestawień teorii świadomości. Przykładowo A. Seth i T. Bayne w artykule z 2022 roku wymieniają ponad 20 współczesnych teorii świadomości inspirowanych neurobiologią (zob. tab. 1 w ich artykule). Autorzy ci koncentrują się na czterech głównych podejściach do świadomości: (1) teoriach typu *higher-order*, (2) teoriach globalnej przestrzeni roboczej, (3) teoriach zintegrowanej informacji, oraz (4) teoriach rekurencyjnych i predykcyjnych świadomości (zob. Seth i Bayne, 2022).

Niektóre ze współczesnych koncepcji świadomości odnoszą się wprost – niekiedy krytycznie – do problemu pojawienia się w przyszłości inteligentnych maszyn zdolnych do świadomych przeżyć (zob. np. Koch, 2019; Seth, 2021). Jednak tym, co łączy większość ze współczesnych podejść, jest zgoda co do tego, że najbardziej przydatnym rozróżnieniem przy analizie dyskutowanego tu zagadnienia jest odróżnienie *świadomości fenomenalnej* (*resp*. doznaniowej, subiektywnej) od *świadomości w ujęciu funkcjonalnym*.

Świadomość fenomenalną – czyli zdolność do posiadania subiektywnych przeżyć, uczuć i odczuwania wrażeń – można potraktować jako *punkt docelowy* rozwoju maszyn inteligentnych i najważniejsze chyba kryterium ich zrównania się z człowiekiem pod względem bytowym i moralnym. U ludzi tak rozumianymi subiektywnymi przeżyciami są na przykład: smutek, zachwyt wywołany widokiem zachodu słońca czy zapach świeżo sparzonej kawy – popularnie zwane qualiami (zob. Chalmers, 2003). Jest to podstawowy sposób świadomego doświadczania przez jednostkę ludzką samej siebie i otaczającego świata, stąd pytanie: czy również maszyny obliczeniowe (*resp*. komputery, algorytmy, sztuczna inteligencja) mogłyby w taki sam sposób odbierać świat i doświadczać subiektywnych przeżyć? Z prawdopodobieństwem bliskim pewności wiemy, że obecnie maszyny inteligentne takich zdolności nie posiadają. Ale czy uda się kiedyś je zbudować albo czy one same mogłyby tak technologicznie wyewoluować, żeby pojawiła się u nich tak rozumiana świadomość fenomenalna? Pytanie to można nazwać „supertrudnym problemem świadomości"[5].

Alternatywne podejście do świadomości – zwane funkcjonalnym – można z kolei potraktować raczej jako *punkt wyjścia* dla prób zaprojektowania takich maszyn oraz zrozumienia, jak mogłyby one działać. W ujęciu funkcjonalnym wykorzystuje się fakt, że świadomość wiąże się nie tylko z posiadaniem subiektywnych odczuć, ale również z bardziej uchwytnymi funkcjami umysłowymi, takimi jak choćby monitorowanie własnego stanu wewnętrznego (*resp*. stanu informacyjnego), kontrola własnego zachowania, podejmowanie decyzji czy planowanie. Trudno to wszystko robić w zaawansowany sposób, nie będąc w jakimś stopniu świadomym. W odróżnieniu od doznań subiektywnych, czyli qualiów – które faktycznie mogą wydawać

[5] Dla D. Chalmersa problem subiektywnych doznań świadomych jest tzw. „trudnym problemem świadomości". Z kolei według M. Tegmarka zrozumienie działania świadomości rozpada się na całą hierarchię problemów: od „łatwego" i „dość trudnego" jak u Chalmersa, przez „jeszcze trudniejszy", aż do „naprawdę trudnego" (zob. Tegmark, 2019: 368–369).

się typowo ludzkie lub być uwarunkowane mechanizmami biologicznymi oraz specyficzną architekturą ludzkiego/zwierzęcego mózgu[6] – cechy takie jak świadome dokonywanie wyborów, automonitoring czy związana z nim kontrola własnego zachowania mogą być powiązane ze świadomością na bardziej elementarnym poziomie i występować niezależnie od tego, czy mamy do czynienia ze świadomością człowieka, zwierzęcia, czy maszyny. Wspólna ludziom, zwierzętom i maszynom byłaby w tym przypadku organizacja funkcjonalna (*resp*. rodzaj obliczeń, struktura przetwarzania informacji) danego procesu poznawczego, w jakimś stopniu niezależna od fizycznego podłoża, w którym dana funkcja jest realizowana. W takiej sytuacji poznanie mechanizmów sterujących procesami świadomymi, na przykład w ludzkim lub zwierzęcym mózgu, mogłoby dopomóc w odtworzeniu analogicznych procesów w syntetycznych układach, co z kolei umożliwiłoby realizację analogicznych funkcji (*resp*. ich symulację) przez inteligentną maszynę.

Można się oczywiście dalej zastanawiać, czy takie zaimplementowanie mechanizmów świadomości funkcjonalnej w maszynach nie doprowadziłoby z czasem – na przykład wskutek kumulatywnego efektu skalowania, procesów uczenia i zjawiska emergencji – do pojawienia się w maszynach również świadomości doznaniowej (fenomenalnej). To jedno z tych kluczowych pytań, na które nie znamy odpowiedzi[7].

Jak dalej zobaczymy, rozróżnienie na świadomość fenomenalną i świadomość w ujęciu funkcjonalnym jest pomocne również w dyskusjach nad interesującymi nas tu trzema problemami: jak rozpoznać świadomą SI, w jaki sposób może się ona powstać i jakie może posiadać własności.

---

[6] Zgodnie z tym według J. Aru, M. Larkuma i J. Shine'a dzisiejsze duże modele językowe (LLM) nie są ani świadome, ani w najbliższym czasie nie uzyskają świadomości, gdyż (1) brak im ucieleśnionego systemu zmysłowego służącego do pobierania informacji ze świata, (2) architektury obecnych sztucznych systemów inteligentnych różnią się od neuronalnej architektury systemu wzgórzowo-korowego, od którego zależy świadomość w świecie zwierzęcym, (3) świadomość może być nierozdzielnie związana z niszą ekologiczną i organizacją systemów żywych (zob. Aru, Larkum i Shine, 2023).

[7] R. Poczobut trafnie wymienia kluczowe dwie niewiadome w tej sprawie: nie wiemy, czy organizacja funkcjonalna układów biologicznych jest do odtworzenia w środowisku sztucznym ani czy odtworzenie organizacji funkcjonalnej w środowisku sztucznym – jeśli by się powiodło – umożliwi pojawienie się świadomości fenomenalnej (zob. Poczobut, 2024: 12).

Jak rozpoznać, że sztuczna inteligencja stała się świadoma? Problem bariery

Jedno z najbardziej kłopotliwych, a zarazem trudnych filozoficznie pytań, jakie można postawić odnośnie świadomości maszyn, dotyczy tego, skąd mielibyśmy wiedzieć, że uzyskały one samoświadomość i stały się zdolne do subiektywnych odczuć. Problem ten kładzie się cieniem na wszystkie pozostałe pytania, gdyż niezależnie od tego, czy zastanawiamy się nad tym, w jaki sposób mogłaby syntetyczna świadomość powstać, czy jaki kształt mogłaby przybrać, zawsze niezbędny okazuje się pewien stopień jej transparentności i dostępności, aby rozpoznać, z czym mamy do czynienia.

Tymczasem subiektywna świadomość nie jest ani transparentna, ani dostępna dla zewnętrznego obserwatora. Bezpośredni, pierwszoosobowy dostęp mamy jedynie do naszych własnych doznań i osobistych przeżyć. To samo dotyczy inteligentnych maszyn: nawet jeśli stałyby się one świadome, to nie moglibyśmy się o tym wprost przekonać, bo nie mielibyśmy możliwości wglądu w ich subiektywne doświadczenia i doznania. Dylemat ten można nazwać „problemem bariery” oddzielającej fenomenalną świadomość indywidualnego człowieka od świadomości wszystkich innych bytów, a szczególnie – i co mnie tu najbardziej interesuje – od świadomości przyszłej SI.

Trudność ta, pomimo że poważna, nie jest jednak nieprzezwyciężalna, o czym świadczy nagminna praktyka codziennego przypisywania świadomości innym osobom. Jak to robimy? Już w XIX wieku John Stuart Mill (1889; zob. równ. Avramides, 2001: 5) zauważył, że w sytuacji braku bezpośredniego wglądu w świadomość drugiego człowieka rozwiązaniem jest „pośredni” do niej dostęp poprzez przeprowadzanie automatycznych i wykonywanych w tle rozumowań na temat stanów umysłowych tej drugiej osoby w oparciu o zasadę *analogii*. Przykładowo, jeśli pamiętam własne zachowanie (np. wczoraj żywo gestykulowałem) i wiem z samoobserwacji, że towarzyszył temu określony stan wewnętrzny (np. rozemocjonowanie), to jeśli zobaczę u innej osoby analogiczne zachowanie (widzę, że ktoś żywo gestykuluje), to na podstawie podobieństwa z własnym doświadczeniem będę skłonny przypisać jej odpowiadający temu zachowaniu stan wewnętrzny (ten ktoś jest rozemocjonowany).

To przypisywanie innym osobom stanów umysłowych – w tym bycia świadomym – na podstawie stwierdzenia analogii z własnym zachowaniem i własnymi stanami umysłowymi jest powszechną praktyką, którą ludzie

w dużym stopniu zautomatyzowali i w której nabyli niezwykłej biegłości. Jednak warunkiem przeprowadzenia tego typu automatycznych wnioskowań jest to, aby inni, którym chcemy przypisać świadomość, byli jakoś do nas podobni i – przede wszystkim – żeby zachowywali się w analogiczny sposób[8]. W podejściu agentowym do sztucznej inteligencji, gdzie SI ma kształt tak zwanych autonomicznych agentów, systemów asystenckich czy cyfrowych współpracowników, spełnienie tego ostatniego warunku nie jest wcale trudne, gdyż funkcje reakcji na bodziec, ukierunkowania na cel, użyteczności, utrzymywania uwagi, planowania czy podejmowania decyzji modelowane są tam jako przynależne zarówno do ludzkiej, jak i maszynowej racjonalności (zob. np. Russell i Norvig, 2023, rozdz. 2). Ta wstępnie założona analogia między człowiekiem a maszyną otwiera następnie drogę do atrybucji maszynom poszczególnych funkcji umysłowych. Jeśli cyfrowy asystent, na przykład ChatGPT, jest w stanie odpowiadać poprawnie na pytania, wygenerować tekst piosenki, wyjaśniać skomplikowane kwestie, a przy tym jego odpowiedzi sprawiają niekiedy wrażenie oryginalnych i kreatywnych, to kontakt z tego typu programem może wytworzyć u człowieka skłonność do dostrzegania u niego zalążków stanów umysłowych – w tym na przykład rozumienia, myślenia czy nawet bycia świadomym – tak samo, jak automatycznie przypisujemy je ludziom, z którymi prowadzimy interesującą rozmowę, którzy odpowiadają na nasze pytania, rozumieją, co do nich mówimy i którzy są językowo kreatywni.

Klasyczną argumentację uzasadniającą pogląd, że o posiadaniu świadomości przez maszynę można wnioskować na podstawie analizy jej zachowania, przedstawił Alan Turing w swoim słynnym tekście *Maszyna licząca a inteligencja* (1950). Przyjął on, że szukając odpowiedzi na pytanie: „czy maszyna ma świadomość?", mamy dwie możliwości: albo spróbujemy polegać na bezpośrednim wglądzie w świadomość maszyny, albo zdamy się na obserwację jej zachowania i na tej podstawie ocenimy, czy jest ona świadoma, czy nie[9]. Według Turinga pierwsza z tych możliwości jest nieefek-

[8] O tym, jak ważne są oba wspomniane warunki – tj. podobieństwo morfologiczne i analogia w zachowaniu – świadczy przypadek atrybucji świadomości przedstawicielom różnych gatunków zwierząt. Kluczowe znaczenie przy tego typu atrybucjach – oprócz wymogu *homologii* zwierzęcego układu nerwowego względem ludzkiego układu nerwowego – wydają się mieć właśnie *analogie* behawioralne. Zwierzętom, które morfologicznie przypominają człowieka, oraz takim, których repertuar zachowań jest podobny do ludzkiego, łatwiej przypiszemy zdolność do bycia świadomymi, zob. np. Ginsburg i Jablonka, 2019: 191–239.

[9] Współcześni badacze zainteresowani stopniem algorytmizacji działania maszyn i umysłów widzą ten problem podobnie. Według P. Stacewicza w odpowiedzi na pytanie „czy

tywna, gdyż dostęp do własnych subiektywnych przeżyć upewnia człowieka jedynie o jego własnej świadomości, chyba „że *byłoby* się tą maszyną i miało poczucie myślenia" (Turing, 1995: 284). Takie postawienie sprawy prowadzi do solipsyzmu. Lepszym rozwiązaniem jest w tej sytuacji zdanie się na drugą możliwość, co musi oznaczać rezygnację z prób bezpośredniego dostępu do świadomości fenomenalnej i poleganie na obserwacji jej zewnętrznych, behawioralnych wskaźników – na przykład zachowania konwersacyjnego, jak ma to miejsce w przypadku testu Turinga. Im odpowiedzi i zachowania maszyny będą lepiej imitowały zachowania językowe ludzi, tym łatwiej przyjdzie nam przypisać jej stany umysłowe i świadomość[10].

Turing zdawał sobie sprawę, że jego własny test nie jest optymalnym narzędziem do rozpoznania świadomości maszyn inteligentnych, ale ze wspomnianych dwóch możliwości – popadnięcia w solipsyzm lub polegania na interpretacji zachowania – jedynie to drugie podejście stwarza jakiekolwiek szanse na przezwyciężenie problemu bariery i ocenę tego, czy maszyna jest świadoma, czy nie. W zastosowaniu do problemu świadomości Turing proponuje istotną modyfikację własnego testu polegającą na zredukowaniu klasycznej gry w udawanie (dwóch uczestników + sędzia) do rozgrywki *viva voce* (gra jeden na jeden), w której chodzi o rozstrzygnięcie, czy rozmówca mechanicznie nauczył się odpowiadać na nasze pytania, czy też je rzeczywiście rozumie (zob. Turing, 1995: 285). Łatwo zauważyć, że przykładowy dialog z rozgrywki *viva voce* przytoczony przez Turinga (zob. Turing, 1995: 285), przypomina nieco dialogi toczone obecnie przez użytkowników

---

maszyna sterowalna algorytmem jest świadoma?", możliwe są dwie strategie postępowania. Albo skonstruujemy „fizyczny detektor świadomości" (analogiczny do detektora fal mózgowych), albo wnioskować będziemy o świadomości maszyny na tej podstawie, że „wykazuje ona inne rozpoznawalne cechy rozumności, na przykład umie szybko i poprawnie rozwiązywać stawiane jej problemy" (np. zachowuje się podobnie do ludzi i włada podobnie jak oni językiem etnicznym, zob. Marciszewski i Stacewicz, 2011: 87).

[10] W klasycznym teście Turinga maszyna zaprogramowana zostaje tak, ażeby w rozmowie z człowiekiem (sędzią) udawać, że jest człowiekiem, i ukrywać, że jest maszyną. Im będzie ona w stanie sprawniej i adekwatnie do kontekstu sytuacyjnego imitować ludzkie zachowania, np. zachowania komunikacyjne i konwersacyjne człowieka, tym mniejsze szanse sędziego na rozpoznanie, czy rozmawia z człowiekiem, czy z maszyną. H. Kissinger, E. Schmidt i D. Huttenlocher trafnie zwracają uwagę, że w istocie test ten „[…] nie sprawdza tego, czy dana maszyna jest zupełnie nieodróżnialna od człowieka, ale to, czy jej działanie przypomina działanie ludzkie. Generatory takie jak GPT-3 są SI dlatego, że tworzą teksty podobne do tych tworzonych przez człowieka, nie zaś ze względu na konkretne rozwiązania w nich zastosowane […]" (Kissinger, Schmidt i Huttenlocher, 2023: 77–78). Behawioralne analogie są więc podstawowym kryterium dla porównywania inteligencji człowieka i maszyny.

z inteligentnymi chatbotami, w których próbują oni przetestować kompetencje językowe, inteligencję i wiedzę posiadaną przez generatywną SI.

Mimo że ten generalny kierunek rozumowania Turinga jest dość wyraźny, szczegółowe kryteria oceny tego, czy mamy do czynienia ze świadomą maszyną, nie są jasne. Wydaje się, że Turing proponuje dwa takie kryteria, bardzo blisko ze sobą powiązane. Po pierwsze, podczas rozstrzygania, czy konwersujemy z bytem świadomym, kluczowe ma być to, czy maszyna „naprawdę rozumie, co mówi, czy tylko 'wyuczyła się jak papuga'" (Turing, 1995: 285). Sugeruje to, że Turing próbuje zredukować problem rozpoznawania świadomości maszyny do rozpoznania jej zdolności rozumienia na przykład komunikatów językowych. W tym miejscu dalej jednak niejasne pozostaje to, co miałoby pomóc człowiekowi odróżnić mechaniczne odpowiadanie na pytania („jak papuga") od „prawdziwego" rozumienia u maszyny. Z uwagi na fakt, że rozgrywka *viva voce* z udziałem maszyny, tak samo jak klasyczny test Turinga, pozostaje rodzajem gry w udawanie, podejrzewać można, że musi chodzić o zademonstrowanie przez maszynę takiego poziomu imitacji ludzkiego zachowania konwersacyjnego, który maksymalnie upodobniałby ją do człowieka i wywierałby na rozmówcy psychologiczne wrażenie obcowania z bytem rzeczywiście rozumiejącym, a zatem i świadomym. Po drugie – i co poniekąd potwierdza powyższe przypuszczenie – Turing podkreśla rolę „sprawności" i wyrafinowanego wykonania przez inteligentną maszynę zadań, do jakich została zbudowana. Przykładowo może ona na przykład konwersować z człowiekiem na temat literatury i poezji tak sprawnie, że trudno będzie człowiekowi zatrzymać się na konstatacji, że to „tylko spreparowane sygnały" i tylko „proste sztuczki" (Turing, 1995: 285).

Jak widać, Turing z perspektywy pierwszej połowy XX wieku z wielką przenikliwością był w stanie przewidzieć kluczowe okoliczności, które będą skłaniały ludzi do doszukiwania się w algorytmach SI „czegoś więcej" aniżeli mechanicznego przetwarzania symboli. Te okoliczności to imitowanie ludzkich zachowań, symulowanie zdolności do rozumienia oraz wyrafinowanie w wykonywaniu zadań, do jakich maszyna została zaprogramowana. Nie tworzą one wystarczającego zbioru przesłanek, aby uznać inteligentne algorytmy za „na pewno świadome", ale psychologicznie uwiarygodniają atrybucję świadomości w tym obszarze. Jak się wydaje, obecne interakcje użytkowników z chatbotami opartymi na dużych modelach językowych, pokazują, że dokładnie ten sam zbiór kryteriów wydaje się ciągle decydujący przy subiektywnej ocenie możliwości i stopnia inteligencji algorytmów (zob. Hoffman i GPT-4, 2023). Można rzec, że „proste sztuczki", o których

wspominał Turing, zostały w tym przypadku ulepszone i zastąpione przez złożony pokaz behawioralnej symulacji i psychologicznej iluzji[11].

Test zaproponowany przez Turinga trudno, mimo wszystko, uznać za adekwatny sprawdzian posiadania przez maszynę subiektywnej świadomości. Nie likwiduje on problemu bariery – jest raczej sprytnym sposobem na jej obejście. Kryteria oceny zachowania maszyny pod kątem przejawiania przez nią świadomości są dość słabe. Rezygnuje się tam z prób bezpośredniego uchwycenia przejawów maszynowej świadomości, a zamiast tego koncentruje się na analizie jej językowych zachowań – rozumienia komunikatów, sensowności odpowiedzi oraz elokwentności i plastyczności konwersacyjnej – co ma wywoływać wrażenie podobieństwa do świadomych ludzkich zachowań.

Czy jest możliwe takie zmodyfikowanie tego testu, aby był on w stanie wychwycić specyficzne dla świadomości procesy i treści umysłowe? Wydaje się, że w tym właśnie kierunku zmierza propozycja Susan Schneider (we współpracy z Edwinem Turnerem). Nawiązując do schematu testu Turinga, zaproponowała ona swój własny, w którym analiza zachowania konwersacyjnego SI miałaby pomóc w odsłonięciu „subtelnej i nieuchwytnej właściwość maszynowej świadomości" (Schneider, 2021: 86). Jej propozycja należy zasadniczo do tej samej kategorii co test Turinga – jest testem behawioralnym i opartym na analizie aktywności komunikacyjno-językowej maszyny – choć istnieje między nimi podstawowa różnica. Schneider, jak sądzę, mniej zależy na analizie *formy* odpowiedzi maszyny, a bardziej na analizie ich *treści* – szczególnie tych, na podstawie których dałoby się rozpoznać samoświadomość oraz samowiedzę maszyny na temat ewentualnych jej doznań i odczuć. Warto w tym miejscu zwrócić uwagę na dwie charakterystyczne cechy zaproponowanego sposobu testowania SI. Po pierwsze, propozycja Schneider zawęża zakres analizowanych wypowiedzi SI do tych, które jakoś dotyczyłyby szeroko rozumianej świadomości. Miałoby się to dokonać dzięki specjalnemu doborowi pytań zadawanych sztucznej inteligencji przez człowieka w trakcie rozmowy. Pytania te miałyby dotyczyć na przykład umiejętności samoidentyfikacji SI, jej możliwości projektowania samej siebie w przyszłości, rozumienia takich problemów, jak „reinkarnacja,

[11] Pouczająca może być historia Blake Lemoine'a, który w trakcie testowania modelu językowego LaMDA zaczął rozpoznawać w zachowaniach programu „bogate życie wewnętrzne" oraz „uczucia, emocje i subiektywne doświadczenia". Zapewne ocena Lemoine'a musiała się opierać się na dużym psychologicznym wrażeniu, jakie wywarła na nim elokwentność i jakość odpowiedzi komunikatora podczas konwersacji (zob. np. Lemoine, 2022).

eksterioryzacja czy zamiana ciał" (Schneider, 2021: 81). Chodzi też o sprawdzenie powiązania ze świadomością zdolności do empatii – na przykład jak inteligentna maszyna zareaguje na wyłączenie (*resp*. śmierć) innej SI – oraz przekonanie się, czy poprzez zmianę parametrów systemu SI można wpłynąć na pojawienie się odmiennych form świadomości. Po drugie – i co następnie okazało się najbardziej chyba komentowanym i kontrowersyjnym punktem tego projektu – autorka testu kładzie nacisk na ocenę tego, czy SI potrafi posługiwać się terminami odnoszącymi się do świadomości w sposób spontaniczny i niewymuszony, bez specjalnych zachęt ze strony człowieka. W celu wykluczenia sytuacji, w której algorytmy symulowałyby jedynie posiadanie świadomości poprzez posługiwanie się wyuczonymi schematami odpowiedzi skopiowanymi z konwersacji przeprowadzanych przez ludzi, Schneider zaproponowała odcięcie testowanej SI od pewnego typu danych treningowych. Aby SI nie imitowała w zaprogramowany sposób odpowiedzi udzielanych przez człowieka na pytania o świadomość, miałaby ona zostać pozbawiona możliwości korzystania z wiedzy wytworzonej przez ludzi na ten temat, w tym – wiedzy o teoriach świadomości i badaniach neuronaukowych, jak również pozbawiona „znajomości takich terminów jak 'świadomość', 'dusza' i 'umysł'" (Schneider, 2021: 83–84). W efekcie pojawiłaby się szansa na przetestowanie tego, czy SI uzyskała świadomość, zanim zdołałaby się ona nauczyć posługiwania wytworzoną przez ludzi wiedzą, pojęciami czy odpowiednimi słowami, na etapie, na którym są „[…] małe dzieci, zwierzęta, a nawet dorośli ludzie […], nie dysponując tymi słowami" (Schneider, 2021: 84).

Jak sądzę, autorka dyskutowanego testu trafnie rozpoznała główne słabości testu Turinga w zastosowaniu do problemu świadomości maszyn i bezsprzecznie poszukuje jakiegoś sposobu na przezwyciężanie jego ograniczeń. Jednak jej własny projekt pokazuje nowe dylematy i trudności, na jakie narazi się każdy, kto chciałby w podobny sposób oceniać, czy maszyna uzyskała świadomość. Po pierwsze, wydaje się, że w projekt Schneider zaszyta jest od samego początku jakiegoś rodzaju antynomijność. Z jednej strony chce się tam dotrzeć do treści pierwotnej fenomenalnej świadomości SI, zanim zostanie ona odkształcona w procesie uczenia przez zapożyczone od ludzi pojęcia i teorie, z drugiej zaś chce się jednak oceniać możliwości sztucznej inteligencji za pomocą kryteriów zapożyczonych właśnie z ludzkiego pojmowania świadomości (samoidentyfikacja, projektowanie siebie w przyszłość, empatia etc.). Można sądzić, że w sytuacji przejawiania przez SI znamion świadomości zupełnie nowego typu mogłyby to nie zostać w ogóle wychwycone w tego typu konwersacyjnym teście opartym

na powyższych kryteriach. Po drugie, wydaje się że autorka przedstawiła jedynie kierunkowe kryteria oceny wypowiedzi SI dotyczących świadomości, bez szczegółowych reguł weryfikacji takich wypowiedzi. Oznacza to między innymi, że proponowany test mógłby nie poradzić sobie z przejawami tak zwanego halucynowania SI, a językowe konfabulacje programu interpretowano by nie w kategoriach błędu, ale jako przejaw jakiejś nowej, nieznanej jeszcze ludziom maszynowej „metafizyki świadomości". O ile halucynowanie odnośnie faktów ze świata realnego jest stosunkowo łatwe do weryfikacji, to językowe halucynacje SI na temat świadomych stanów umysłowych mogłyby być dość trudne do rozpoznania (zob. np. Sui, Duede i in., 2024; Xu, Jain i in., 2024; Marcus, 2024: rozdz. 2).

Last but not least, problem odcięcia SI od danych dotyczących wypracowanych przez ludzi pojęć i teorii świadomości sam w sobie zdaje się generować cały szereg kontrowersji. Przypomina to nieco dylemat spotykany obecnie w uczeniu modeli SI: czy należy trenować je na danych wytworzonych przez człowieka, czy na danych syntetycznych? Jeśli degradacja poznawcza modeli niekarmionych danymi wytworzonymi przez człowieka ma rzeczywiście miejsce, to nie wróży to dobrze generalnemu kierunkowi rozumowania przedstawionemu przez Schneider[12]. Gdyby nawet uwzględnić podział w architekturze uczenia LLM-ów na „model bazowy" wstępnie przetrenowany na olbrzymich zbiorach danych (*base LLM*), oraz „model dotrenowany" pod kątem wykonywanego zadania za pomocą wyspecyfikowanego zbioru danych lub ograniczeń nałożonych na model (*fine-tuned LLM* – zob. np. Sejnowski, 2024: 113; też Foster, 2024: 254), to i tak nie wiadomo, na którym z tych etapów można by wprowadzić postulowane przez Schneider odcięcie SI od danych treningowych na temat ludzkiego pojmowania świadomości. Dodatkowo, nawet jeśli przyjąć, że wstępnie

[12] Analogia ta jest jedynie przybliżona, gdyż Schneider nie zajmuje się w ogóle problemem wsadowych danych syntetycznych (jej książka powstała, zanim cały ten problem się ujawnił), a jedynie postuluje dziedzinowe uszczuplenie zbioru danych treningowych wytworzonych przez człowieka. Tymczasem tym, co obecnie niepokoi badaczy, jest przyrost danych syntetycznych, wytworzonych przez generatywne SI i dostępnych w Internecie w stosunku do danych wytworzonych przez ludzi. Trenowanie kolejnych generacji dużych modeli językowych na danych syntetycznych wydaje się przypominać – mówiąc obrazowo – zjadanie przez węża własnego ogona. Dane syntetyczne mogą nie odzwierciedlać prawdziwej złożoności świata, mogą być niereprezentatywne i powodować skażenie Internetu. W efekcie oparcie procesu uczenia na danych syntetycznych może powodować degradację modeli, zob. Shumailov, Shumaylov i in., 2024; Alemohammad i Casco-Rodriguez, 2024; zob. też Usidus, 2023.

przetrenowane modele generatywnej sztucznej inteligencji – albo przed ich dotrenowaniem za pomocą pojęć zaczerpniętych z ludzkiego słownika doświadczeń mentalnych, albo po ich stuningowaniu ograniczającym stosowanie tego słownika – wytworzyłyby jakiegoś rodzaju formę świadomości minimalnej, to i tak jest bardzo wątpliwe, aby dało się ją wyrazić językowo za pomocą pozostałej, dopuszczonej do testu części języka naturalnego, która wydaje się koherentnie dopasowana do raportowania stanów mentalnych ze zestandaryzowanego *ludzkiego* poziomu.

W jaki sposób świadoma SI mogłaby powstać? Podejście projektowe vs. podejście emergentne

Jak mogliśmy się przekonać, pytanie „jak rozpoznać, że sztuczna inteligencja uzyskała świadomość?" jest ekstremalnie trudne – jeśli w ogóle możliwe – do rozstrzygnięcia. Kolejne kluczowe zagadnienie – tym razem dotyczące widoków na powstanie w przyszłości świadomej SI – może w pierwszej chwili wydawać się mniej filozoficznie karkołomne. Ale przez to, że dotyczy zdarzeń umiejscawianych w przyszłości, których jeszcze nie ma, ono również wikła się w spekulacje obarczone dużym stopniem ryzyka i niepewności. Badacze przez wiele ostatnich lat próbowali ze zmiennym szczęściem prognozować dalsze etapy rozwoju sztucznej inteligencji (zob. np. Amstrong i Sotala, 2015; Bostrom, 2016; Tegmark, 2019), lecz w przypadku przewidywań na temat uzyskania przez SI świadomości stopień ryzyka takich predykcji rośnie wykładniczo. Inaczej niż w przypadku prognoz na temat rozwoju technologicznego czy rozwoju sztucznej inteligencji, gdzie możliwe jest posłużenie się „analizą trendów", w przypadku przewidywania pojawienia się syntetycznej świadomości narzędzie to wydaje się obecnie bezużyteczne, gdyż brak jest jakichkolwiek danych historycznych i wzorców ich interpretacji pozwalających na ujawnienie takiego trendu[13]. Najtrudniejsze wydają się w tym przypadku pytania najbardziej generalne i obciążone wieloma filozoficznymi wątpliwościami, na przykład „czy świadoma SI w ogóle powstanie?", a także pytanie najbardziej konkretne, typu „kiedy do tego dojdzie?".

[13] Chyba że „epokę powstania sztucznej świadomości" włączylibyśmy w spekulatywny plan technologicznej historiozofii w stylu nadejścia osobliwości, zob. np. Kurzweil, 2013, rozdz. 1.

Pewną szansę na postawienie problemu powstania świadomej SI w perspektywie przyszłego rozwoju technologicznego stwarza natomiast przesunięcie akcentu z pytań „czy" i „kiedy", na pytanie „jak". Pytamy wtedy o to, „jakie są możliwe scenariusze powstania w przyszłości świadomej SI?"[14]. Na tak postawione pytanie rysują się dwie możliwe odpowiedzi: że odbędzie się to w sposób projektowy lub w sposób emergentny. Jest prawdopodobne, że jeśli w ogóle dojdzie do powstania myślącej, czującej i samoświadomej SI, to stanie się to na jeden z dwóch wymienionych sposobów lub w sposób będący połączeniem ich obu.

Syntetyczna świadomość jako wynik projektowania

Filozoficzne podstawy tej idei znane są od dawna. W najbardziej generalnej postaci odwołuje się ona do tak zwanej zasady organizacyjnej inwariantności[15] głoszącej, że „[…] układy fizyczne o tym samym abstrakcyjnym schemacie organizacji będą generować takie same doznania świadome niezależnie od tego, z jakich elementów fizycznych zostały wykonane" (Chalmers, 2003: 110). Jeśli więc udałoby się odtworzyć wzorzec połączeń między neuronami i główne układy mózgowe w postaci krzemowych układów scalonych i elementów architektury komputera, to można by liczyć na to, że również w maszynie cyfrowej pojawią się przynajmniej jakieś zalążki świadomości. Do tej generalnej idei przez lata odwoływały się mniej lub bardziej fantastyczne pomysły i eksperymenty myślowe „przeniesienia świadomości na komputer" lub „zasymulowania świadomości w maszynie cyfrowej". Mimo że nie jest to ani logicznie sprzeczne, ani wykluczone przez podstawowe prawa fizyki (zob. Poczobut, 2024: 9), to na obecnym etapie rozwoju badań i technologii obliczeniowych nie widać szans na pełną realizację tego projektu[16].

---

[14] Obok rozumowań opartych na „prawach dziejowych" oraz na „wnioskowaniu z trendów" spekulowanie na temat przyszłości na podstawie „możliwych scenariuszy" wydaje się trzecim dostępnym sposobem stawiania prognoz przyszłych zdarzeń.

[15] Zasada ta pochodzi z funkcjonalistycznej tradycji w filozofii umysłu.

[16] W ramach teorii zintegrowanej informacji G. Tonioni i Ch. Koch sformułowali własne sceptyczne stanowisko w tej sprawie. Ich zdaniem dzisiejsze architektury maszyn obliczeniowych nie nadają się do wytworzenia świadomości nawet przy założeniu ich dalszego dynamicznego rozwoju. Nawet jeśli udałoby się zeskanować cały mózg człowieka i odtworzyć go w środowisku komputerowym, to nie oznacza wcale, że taka komputerowa symulacja uzyskałaby świadomość. Koch daje tu przykład: gdy symuluje przebieg huraganu na komputerze, w pomieszczeniu, w którym stoi komputer, nie zaczyna przecież nagle

Jako pewną przymiarkę do jego przyszłej realizacji – na razie bardzo okrojoną i cząstkową – można natomiast potraktować sięganie do istniejących neuronaukowych teorii świadomości po to, aby na ich podstawie spróbować odtworzyć w syntetycznym materiale niektóre z mechanizmów odpowiedzialnych za występowanie świadomości u człowieka lub zwierząt. Wybiera się zwykle te mechanizmy lub procesy, które mają charakter funkcjonalny – mogą nimi być na przykład mechanizmy odpowiedzialne za podejmowanie decyzji, planowanie czy raportowanie własnych stanów – a ich odtworzenie w architekturze obliczeniowej komputera lub algorytmu dawałoby nadzieję na pojawienie się w maszynie zalążka świadomych procesów umysłowych. Kluczowe jest tu oczywiście założenie, że wymienione procesy mają podobną organizację funkcjonalną (*resp*. budowę sieci neuropodobnych, architekturę procesu obliczeniowego, procesu uczenia etc.) niezależnie od tego, czy są realizowane przez ludzki mózg, czy przez algorytm SI. Przekonanie, że kluczem do zrozumienia świadomości jest rozgryzienie zagadki jej funkcjonalnej architektury, a nie zastanawianie się nad naturą subiektywnych doznań, zdaje się – przynajmniej w teorii – nieco redukować skalę trudności związanych z zaprojektowaniem świadomej SI.

Tym właśnie tropem idą na przykład autorzy zbiorowego opracowania na temat możliwości zbudowania świadomej SI (Butlin, Bengio i in., 2023). Dokonali oni przeglądu kilku współczesnych naukowych teorii świadomości i na ich podstawie wyróżnili cały szereg funkcjonalnych wskaźników procesów świadomych, których odtworzenie w strukturze funkcjonalnej algorytmu SI mogłyby ukierunkować wysiłki programistów i inżynierów ku określonym rozwiązaniom służącym zaprojektowaniu świadomej SI[17]. Autorzy raportu stwierdzają, że wprawdzie dzisiejsza SI nie jest jeszcze

padać deszcz. Albo podczas symulacji sił grawitacyjnych w czarnej dziurze nie zostajemy pochłonięci przez przestrzeń biurka, na którym stoi sprzęt komputerowy. Układy komputerowe – przynajmniej o rozpowszechnionej dziś architekturze – nie mają odpowiedniego poziomu integracji i odpowiedniej mocy sprawczej, aby te zjawiska wywołać; zob. Massimini i Tononi, 2018; Koch, 2019; 2020.

[17] Butlin, Bengio i in. przeanalizowali sześć współczesnych neuronaukowych i psychologicznych teorii świadomości – tj. teorie rekurencyjne, teorie globalnej przestrzeni roboczej, obliczeniowe teorie wyższego rzędu, teorie schematów uwagowych, teorie kodowania predykcyjnego, oraz teorie oparte na agencyjności i ucieleśnieniu – i na tej podstawie wyróżnili łącznie piętnaście wskaźników, których odtworzenie w układach sztucznych mogłoby sprawić, że SI uzyskałaby funkcjonalną bazę dla pojawienia się w niej świadomości; zob. np. Butlin, Bengio i in., 2023: 45 i nast.

świadoma, ale równocześnie „nie istnieją żadne oczywiste przeszkody dla budowy systemów świadomej SI" (Butlin, Bengio i in., 2023: 1)

Jedną z analizowanych przez nich teorii jest model świadomości jako globalnej przestrzeni roboczej (*global workspace theory* – GWT), w ramach którego postuluje się, że kluczowym mechanizmem odpowiedzialnym za świadomość człowieka jest umożliwianie dostępu różnym modułom (*resp*. układom funkcjonalnym, takim jak język, działanie, pamięć długotrwała, system oczekiwań etc.) do informacji w ramach tak zwanej przestrzeni roboczej. Ujmując rzecz maksymalnie skrótowo: świadomość to nic innego jak „globalne dzielenie się informacjami" w przestrzeni (pamięci) roboczej (zob. Dehaene, 2023: 222).

Analizując tę wpływową teorię świadomości, autorzy artykułu doszli do wniosku, że kluczowe dla możliwości powstania świadomej SI byłoby odtworzenie w syntetycznej strukturze szeregu cech organizacji funkcjonalnej leżących u podstaw świadomych procesów poznawczych człowieka, tak jak je rekonstruuje GWT[18].

I tak postuluje się, aby SI miała organizację złożoną z wielu działających równolegle modułów (GWT-1), ażeby przestrzeń robocza pełniła w niej funkcję „wąskiego gardła" dla strumienia informacji odpowiadającego za funkcję selekcji uwagi (GWT-2), ażeby informacja pojawiająca się w przestrzeni roboczej była równocześnie dostępna dla wielu modułów (GWT-3), natomiast pobieranie informacji było regulowane zarówno przez stan systemu, zapotrzebowanie poszczególnych modułów, jak i czułość na nowe dane wejściowe (GWT-4; zob. Butlin, Bengio i in., 2023: 26–27, 49–51)[19]. Odtworzenie w systemach SI tego rodzaju architektury funkcjonalnej mogłoby stanowić podstawę dla budowy w przyszłości świadomej SI.

---

[18] GWT już wcześniej była traktowana jako model, na podstawie którego można próbować odtworzyć architekturę świadomego mózgu w syntetycznym materiale. Dechaene, Lau i in. (2017) zajęli się problemem, czy jest możliwe odtworzenie w inteligentnych maszynach dwóch kluczowych procesów, które są bazą dla świadomości: globalnej dostępności informacji i automonitorowania.

[19] Warunki GWT-3 i GWT-4 są często wskazywane jako kluczowe dla uzyskania przez SI świadomości, jeśli miałaby ona być modelowana na podstawie teorii globalnej przestrzeni roboczej. S. Dehaene nazywa je „niezbędnymi funkcjami, których brakuje obecnym komputerom" (Dehaene, 2023: 352). W jego nomenklaturze można je nazwać odpowiednio „elastyczną komunikacją", w której chodzi o to, aby wyniki uzyskane w poszczególnych aplikacjach na komputerze były dostępne w przestrzeni roboczej dla wszystkich pozostałych programów, oraz „plastycznością", w której idzie o to, aby programy same odkrywały najlepszy użytek, jaki mogą uczynić z uzyskanej informacji. Dehaene dodaje do tego jeszcze trzecią niezbędną funkcję, którą nazywa „autonomią": możliwość decydowania przez

Podobną metodą posłużył się ostatnio również David Chalmers, próbując rozstrzygnąć, czy duże modele językowe – na przykład ChatGPT i LaMDA – mają jakiekolwiek oznaki świadomości oraz czy jest możliwe, że posiądą je w przyszłości (zob. Chalmers, 2023)[20]. Chalmers wyróżnia kilka rodzajów cech kojarzonych zwykle ze świadomością człowieka lub zwierząt, a następnie analizuje, czy występują one również w dużych modelach językowych. Są wśród nich takie cechy jak samoraportowanie własnych stanów wewnętrznych, możliwości konwersacyjne oraz wrażenie, jakie wywiera się na ludziach podczas konwersacji – czyli te, które można oszacować na podstawie interpretacji zachowania (np. w teście Turinga). Jest też inna grupa cech, takich jak dysponowanie biologiczną podstawą oraz posiadanie zmysłów i ciała, które różnią świadome organizmy od programów komputerowych. Wreszcie Chalmers analizuje cechy, które wynikają z teorii naukowych na temat działania biologicznych umysłów i świadomości, jak na przykład posiadanie inteligencji ogólnej, dysponowanie wewnętrznym modelem świata i modelem samego siebie, występowanie procesów rekurencyjnych, posiadanie globalnej przestrzeni roboczej i zunifikowanego „ja".

Odnosząc te wskaźniki do systemów opartych na modelach LLM i do niektórych innych systemów SI, łatwo dojść do wniosku, że zdecydowana większość z tych cech nie występuje w ich przypadku, a zatem można z bardzo dużym prawdopodobieństwem przyjąć, że obecnie systemom tym brakuje świadomości[21]. Zarazem jednak Chalmers nie wyklucza, że LLM-y mogą – i to już niedługo – posiąść niektóre z powyższych cech, co by mogło wskazywać na pojawienie się w nich zalążków świadomości.

---

system, które z danych znajdujących się w pamięci roboczej są warte świadomej analizy, oraz wprowadzania tam co jakiś czas własnych losowych treści, które można utożsamić z wolno płynącymi myślami (Dehaene, 2023: 352–353).

[20] Chalmers zwraca uwagę na obecną ewolucję LLM-ów w stronę architektur wielomodułowych, posiadających nie tylko zdolność operacji na języku naturalnym, ale również możliwość przetwarzania obrazu i dźwięku, sprawowanie kontroli nad obiektami fizycznymi i witrualnymi, komunikację z różnymi bazami danych oraz symulację różnych procesów i zjawisk. Pod tymi względami modele te wstępnie zaczynają upodabniać się do działania ludzkiego systemu poznawczego. Te przyszłe wersje obecnych dużych modeli językowych Chalmers nazywa „modelami rozszerzonymi" i oznacza je jako LLM+; Chalmers, 2023: 2.

[21] Według Chalmersa niektóre argumenty przeciwko możliwości uzyskania przez LLM-y świadomości, np. wynikające stąd, że świadomość ma naturę biologiczną i musi mieć ugruntowanie w zmysłach, są kontrowersyjne. Inne argumenty, jak np. ten, że LLM-y nie dysponują własnym modelem świata, uważa on za mało oczywiste. Najtrudniejsze do oddalenia wydają mu się argumenty stwierdzające, że LLM-om brakuje możliwości przetwarzania rekurencyjnego, globalnej przestrzeni roboczej i zunifikowanego „ja" (Chalmers, 2023: 18).

Analiza powyższych cech-wskaźników podpowiada różne rozwiązania do tego prowadzące. Będzie to między innymi budowanie modeli wielomodalnych typu percepcja-język-działanie, ucieleśnionych i wyposażonych w liczne sensory trenowane na przykład w świecie wirtualnym. Dalej – tworzenie LLM-ów, które dysponowałyby semantycznym modelem świata i modelem samego siebie. Jak również – wyposażenie LLM-ów w globalną przestrzeń roboczą, oparcie ich działania na procesach pamięciowych i rekurencyjnych, a także trenowanie takich modeli na danych zindywidualizowanych, co mogłoby zapewnić im większy poziom „spersonalizowania” (zob. Chalmers, 2023: 10–18). Zdaniem Chalmersa konstruowanie systemów sztucznej inteligencji opartych na tych wytycznych mogłoby nas przybliżyć do powstania świadomej SI.

Należy mieć na uwadze, że propozycje przedstawione przez Bengio, Butlina i in. oraz Chalmersa, dotyczące tego, jak zbudować systemy świadomej sztucznej inteligencji, są na dość wczesnym etapie realizacji i ich ostateczny wynik jest niepewny. Niewątpliwie mamy tutaj do czynienia z projektami, które próbują zidentyfikować kluczowe funkcjonalne mechanizmy odpowiadające za świadomość w świecie biologicznym, a następnie przenieść je na grunt systemów sztucznych. To podejście niesie mnóstwo problemów – na przykład z identyfikacją i wyborem poszczególnych mechanizmów, z ich zgraniem ze sobą, z ich odtworzeniem w sztucznym środowisku, oraz niepewność co do tego, czy tam one zadziałają – ale mimo to jest to w obecnej chwili najbardziej chyba konkretna i interesująca propozycja idąca w tym kierunku.

Warto też zwrócić uwagę na jeszcze jeden cel przyświecający temu projektowi. Jest nim dostarczenie uchwytnych kryteriów dla rozpoznania tego, czy system SI jest świadomy, czy nie. Wcześniej widzieliśmy, że takich kryteriów poszukiwano w ramach podejścia behawioralnego (zob. rozdział „Jak rozpoznać, że sztuczna inteligencja stała się świadoma? Problem bariery” niniejszego tekstu). Z kolei w przypadku podejścia funkcjonalnego o tym, czy SI jest świadoma, miałoby decydować to, czy udało się w jej architekturze odtworzyć funkcjonalne mechanizmy odpowiedzialne za powstanie świadomości w świecie biologicznym. Nie potrzeba zbyt wielkiego wysiłku, aby przekonać się, że oba wymienione kryteria – tj. behawioralne i funkcjonalne, brane z osobna – są dość słabą miarą posiadania świadomości przez systemy sztuczne. Zsumowanie ich obu wzmacniałoby nieco, w moim przekonaniu, siłę przekonywania na rzecz takiej tezy. Jeśli zatem udałoby się zbudować SI, która nie tylko przechodziłaby behawioralny test Turinga (oraz test w postaci gry *viva voce*, a także test Schneider), ale równocześnie

posiadałaby też wbudowane i działające mechanizmy funkcjonalnej architektury świadomości, o których mówi Chalmers (oraz Butlin, Bengio i in.), to teza o posiadaniu przez taki system doznań świadomych zyskiwałaby wstępnie nieco na wiarygodności.

Syntetyczna świadomość jako wynik spontanicznego rozwoju systemów inteligentnych

Alternatywny sposób myślenia o powstaniu świadomej SI jest oparty na przypuszczeniu, że mogłaby ona wyłonić się spontanicznie jako tak zwana własność emergentna na odpowiednio zaawansowanym etapie rozwoju sztucznej inteligencji, nawet bez zaplanowania tego przez projektantów i inżynierów. W tym ujęciu świadomość miałaby niejako podążać za rozwojem inteligencji, a gdy ta ostatnia osiągnęłaby wysoce zaawansowaną postać, to wraz z tym mogłyby pojawić się oznaki życia wewnętrznego i samoświadomości[22]. Możliwe też, że mechanizmy spontanicznego rozwoju SI mogłyby współdziałać z wysiłkami programistów, efektywnie wzmacniając zapoczątkowane przez nich procesy[23]. Taki scenariusz pojawienia się świadomości maszynowej mieli zapewne na myśli wizjonerzy rozwoju technologicznego, na przykład Stanisław Lem, który przypuszczał, że świadomość w maszynach może pojawić się stopniowo i trochę nieoczekiwanie, w wyniku wprowadzania kolejnych drobnych poprawek i przeróbek, co umożliwiałoby przejście od bezdusznych urządzeń do myślących maszyn (Lem, 1984: 113–114).

Przewidywania tego rodzaju nasiliły się w ostatnich latach między innymi w związku z zaobserwowaniem cech o charakterze – jak się sądzi – emergentnym w obszarze generatywnej sztucznej inteligencji, na przykład u inteligentnych chatbotów opartych na dużych modelach językowych i architekturze transformerów. Przykładowo modelowi GPT-4, zaprojektowanemu jako system inteligencji konwersacyjnej, której zadaniem jest sprawne udzielanie odpowiedzi na prompty użytkownika oparte na

---

[22] Nikt jednak nie wie, jaki stopień zaawansowania inteligencji systemy sztuczne musiałyby osiągnąć, aby stać się świadome i czy faktycznie sam spektakularny rozwój inteligencji jest wystarczającym warunkiem pojawienia się świadomości. Dlatego teza o „podążaniu świadomości za rozwojem inteligencji" spotyka się z licznymi zarzutami, zob. np. Koch, 2019, rozdz. 12 i 13; Seth, 2021: 250 i nast.

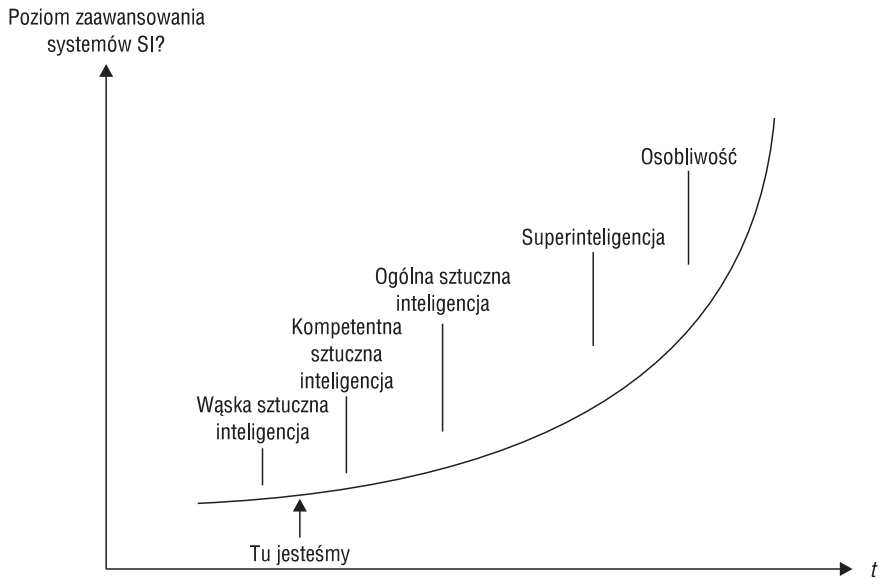
[23] Wydaje się, że na takim łączonym rozwiązaniu opiera się koncepcja „zalążkowej SI", gdzie postuluje się, że SI spontanicznie powinna rozwijać zaszczepione jej przez człowieka wartości. zob. Bostrom, 2016, rozdz. 12.

generowaniu kontynuacji ciągu wyrażeń (zob. Wolfram, 2023), zaczęto przypisywać „przebłyski” kreatywności i rozumienia świata (zob. Bubeck, Chandrasekaran i in., 2023). Badacze zwracają uwagę, że na przykład sprawne udzielanie odpowiedzi przez chatbota na pytania dotyczące sytuacji mających miejsce na świecie i rozlokowanych w czasie i przestrzeni może zależeć nie tylko od formalnej kompetencji językowej, ale również od całego szeregu „kompetencji funkcjonalnych” polegających na posiadaniu wiedzy na temat przedstawianej sytuacji, dysponowaniu modelem świata i reprezentowaniu relacji w nim zachodzących (zob. krytyczne podejście Mahowald, Ivanova i in., 2023; zob. też artykuł przeglądowy van Dijk, Kouwenhoven i in., 2023). Główne pytanie dotyczy więc tego, czy na przykład w modelu GPT-4 – poza sprawnym generowaniem fraz językowych, do czego został on zaprogramowany – pojawiły się (np. w wyniku uczenia się i treningu lub ulepszeń w kolejnych wersjach algorytmu) również inne umiejętności, na przykład zdolność rozumienia czy reprezentowania świata w nielingwistyczny sposób, podobne choćby do tego, co robią ludzie, gdy posługują się wyobraźnią lub wglądem w istotę problemu.

Jeśli zaś w systemach generatywnej sztucznej inteligencji faktycznie mogły pojawić się w wyniku emergencji cechy umożliwiające takie nowatorskie reprezentowanie i rozumienie świata, to nie da się wykluczyć, że kiedyś w podobny sposób algorytmy mogłyby uzyskać również zdolność do subiektywnych doświadczeń i świadomość samej siebie.

Na razie trudno jest jednoznacznie i bez wątpliwości stwierdzić tak dalece idącą jakościową ewolucję jakiejś linii rozwojowej konkretnego algorytmu czy systemu SI. Dlatego możliwość spontanicznego pojawienia się syntetycznej świadomości jest zwykle rozważana w perspektywie odległej przyszłości i długofalowych trendów rozwoju sztucznej inteligencji. Wadami takich długoterminowych scenariuszy jest jednak – po pierwsze – nieznajomość czynników uruchamiających kaskadę długofalowych procesów emergentnych, które mogłyby doprowadzić do wyłonienia się świadomej SI. Przykładowo, nie wiadomo, czy powstanie syntetycznej świadomości miałoby być spowodowane tak zwanym efektem skali (tj. wzrostem złożoności systemów) związanym z budowaniem coraz bardziej zaawansowanych urządzeń i algorytmów o coraz większej mocy (liczonej ilością parametrów) oraz trenowanych za pomocą coraz większej ilości danych, czy raczej bardziej prawdopodobna jest hipoteza, że świadoma SI nie pojawi się, zanim nie dokona się jakościowy skok związany z przezwyciężeniem ograniczeń tkwiących w dominujących obecnie architekturach oprogramowania i architekturach sprzętowych zależnych

od rozwijanych technologii informatycznych? Po drugie – nie można przewidzieć, na jakim etapie przyszłego rozwoju SI mogłaby mieć miejsce taka „eksplozja sztucznej inteligencji" powiązana z pojawieniem się syntetycznej świadomości. Przykładowo, szeroko dotąd rozwijana postać sztucznej inteligencji nazywana jest często „wąską SI", co oznacza, że realizuje ona zadania w określonym wąskim zakresie i nie jest zdolna wyjść poza cele, do których została zaprojektowana. Pobiera ona „dane z jednej konkretnej dziedziny i stosuje je od optymalizacji jednego konkretnego celu" (Lee, 2019: 23). Według Mustafy Suleymana jesteśmy obecnie dopiero na etapie tworzenia zrębów zaawansowanych systemów, które miałyby postać „kompetentnej sztucznej inteligencji" zdolnej wykonywać złożone, wieloetapowe i kompleksowo powiązane ze sobą zadania (Suleyman, 2024: 121–122). Możliwa trajektoria dalszego rozwoju sztucznej inteligencji mogłaby wyglądać w następujący sposób:



Il. 2. Możliwa trajektoria przyszłego rozwoju sztucznej inteligencji. Obejmuje etapy: wąskiej sztucznej inteligencji (*narrow artificial intelligence* – NAI), kompetentnej sztucznej inteligencji (*artificial capable intelligence* – ACI), ogólnej sztucznej inteligencji (*artificial general intelligence* – AGI), superinteligencji (*artificial superintelligence* – ASI) oraz osobliwości (*singularity*)

Źródło: opracowanie własne

Czy na którymś z tych etapów mogłyby pojawić się okoliczności sprzyjające powstaniu syntetycznej świadomości? Spora część nadziei na dalszy rozwój SI wiązana jest obecnie z powstaniem „ogólnej sztucznej inteligencji”, która miałaby być sztuczną inteligencją ogólnego zastosowania i obejmować między innymi zdolność do transferu rozwiązań z jednej dziedziny do drugiej, w połączeniu z adaptacyjnością i plastycznością polegającymi na elastycznym przechodzeniu między zadaniami i na kreatywnym stosowaniu wyuczonych rozwiązań do nowych sytuacji (zob. Moris, Sohl-Dickstein i in., 2024[24]). Czy okoliczności te sprzyjałyby pojawieniu się w takich systemach zapotrzebowania na świadome myślenie? Jeszcze dalej sięgające w przyszłość projekty przewidują powstanie „superinteligencji”, która po zrównaniu się z inteligencją ludzką rozwijałaby się dalej w tempie wykładniczym, zostawiając w tyle ludzkie umiejętności poznawcze (zob. Bostrom, 2016), a w końcu – dojście do „punktu osobliwości” oznaczającego połączenie się inteligencji biologicznej z inteligencją sztuczną (Kurzweil, 2013).

Niestety, jak już wspomniałem, wadą tej spekulatywnej wizji dalszego rozwoju SI jest brak przekonujących argumentów wskazujących, na którym z wymienionych etapów mogłyby pojawić się warunki wystarczające dla pojawienia się syntetycznej świadomości.

Jakie cechy mogłaby mieć świadoma SI? Hipotetyczne scenariusze

Rozpoznanie, czy maszyny wyposażone w SI rzeczywiście posiadły zdolność do subiektywnego przeżywania, może okazać się – jak starałem się pokazać – nie lada problemem dla ich użytkowników. Podobnie trudno jest z dzisiejszej perspektywy prognozować, czy w przewidywalnym czasie pojawią się czujące i świadome maszyny. Paradoksalnie nieco łatwiejsze wydaje się określenie już teraz, jaką taka syntetyczna świadomość *mogłaby* mieć postać i jakie miałaby cechy, *gdyby* jednak powstała.

Służą do tego hipotetyczne scenariusze, w ramach których stosuje się strategię polegającą na próbie „wydedukowania” pewnych ogólnych cech przyszłej świadomej SI na podstawie analizy możliwości i ograniczeń wynikających z ogólnych praw i teorii naukowych. Z jednej strony takim tropem

[24] Moris, Sohl-Dickstein i in. proponują własną wizję rozwoju ogólnej sztucznej inteligencji (AGI), obejmującą pięć poziomów jej zaawansowania: wyłaniającej się AGI (obecne modele: GPT-3,5, Bard, Llama 2, Gemini), kompetentnej AGI, eksperckiej AGI, wirtuozerskiej AGI i superinteligencji; zob. Moris, Sohl-Dickstein i in., 2024: 4–6.

podążają niekiedy twórcy literatury czy kina *science fiction*, którzy przy tworzeniu swoich technoutopijnych fabuł czy dystopijnych wizji próbują odgadnąć potencjał rozwojowy tkwiący we współczesnych technologiach[25]. Z drugiej strony podobną strategię stosują naukowcy, próbując dociec, jaką mogłaby taka syntetyczna świadomość mieć postać, przy uwzględnieniu całego szeregu ograniczeń i możliwości nakładanych na świadomość przez układy fizyczne.

Co ciekawe, podejście to pozwala zastanowić się, jaki kształt mogłaby mieć świadomość fenomenalna u inteligentnych maszyn, a więc przekracza ono perspektywę funkcjonalną w patrzeniu na świadomość, która dominuje mimo wszystko w nauce.

Przykładowo, taką właśnie drogą idzie współczesny fizyk teoretyczny Max Tegmark. Pisze on: „[…] stosując […] argumenty oparte na fizyce, możemy domyślać się pewnych aspektów tego, jak czułaby się sztuczna świadomość" (Tegmark, 2019: 396). Interesujące w tym ujęciu jest też posłużenie się perspektywą porównawczą, gdzie zestawia się hipotetyczną syntetyczną świadomość maszyny z biologiczną świadomością człowieka, która służy jako punkt odniesienia.

I tak, Tegmark przyjmuje dość racjonalnie, że obszar doznań i odczuć sztucznej inteligencji ilościowo mógłby być o wiele bardziej rozbudowany („ogromny") w porównaniu z doznaniami świadomymi człowieka, co wynika z możliwości wyposażenia maszyny w większą liczbę zewnętrznych i wewnętrznych czujników oraz receptorów. W przypadku człowieka świadome doznania, czyli qualia, „podczepione" są pod pięć ludzkich zmysłów i pod interocepcję, podczas gdy u maszyn wyposażonych w SI mogłyby występować dodatkowo na przykład odczucia nadbudowane nad zdolnością do „słyszenia" ultradźwięków lub „widzenia" w podczerwieni.

Interesujące jest również przyjęcie, że sztuczna świadomość mogłaby doświadczać więcej przeżyć w jednostce czasu niż ma to miejsce u człowieka („może mieć miliony razy więcej niż my doznań na sekundę"). Wynika to stąd, że przewodnictwo w układach elektronicznych jest o wiele szybsze niż przemieszczanie się sygnałów w mózgu. Syntetyczne qualia mogłyby być zatem gęściej upakowane w jednostkach czasu, a także przepływałyby w strumieniu syntetycznej świadomości o wiele szybciej niż u ludzi (Tegmark, 2019: 396).

---

[25] Ich wizje skłaniają jednak często do spojrzenia na nie z dystansem i „z przymrużeniem oka" ze względu na poszukiwanie atencji i uznania u odbiorcy poprzez podkoloryzowane fabularyzowanie i artystyczne konfabulowanie na temat przyszłości.

Tegmark przywołuje też znane z psychologii rozróżnienie na szybkie myślowe automatyzmy i wolniejsze myślenie refleksyjne (zob. np. szybki i automatyczny System 1 i wolniej działający System 2 wg Daniela Kahnemana). Jego zdaniem ta dystynkcja mogłaby mieć swój odpowiednik w przyszłej sztucznej inteligencji, gdyż poza wykonywaniem rutynowych zadań, gdzie polecenia przekazywane są nieświadomym i szybko działającym podsystemom, SI będzie prawdopodobnie zarządzać informacją w rozbudowanych i bardzo rozległych sieciach informatycznych, a „[…] im większa sztuczna inteligencja, tym wolniejsze muszą być jej globalne myśli, aby dać informacji czas na przepływ między wszystkimi jej częściami" (Tegmark 2019: 396, 397–398). Możliwe zatem, że w zaawansowanych i rozbudowanych systemach SI świadomość pojawi się pod postacią globalnych i „wolno płynących" myśli, które umożliwiałyby komunikację między odległymi obszarami sieci.

Czy zaawansowana SI mogłaby również doświadczać subiektywnych przeżyć, które u ludzi wywodzą się z ich biologicznego i ewolucyjnego pochodzenia? Chodzi przede wszystkim o odczucia związane z głodem, bólem, pożądaniem seksualnym czy instynktem samozachowawczym. Można zasadnie argumentować, że większość z tego typu odczuć nie jest potrzebna SI, a konieczność zapewnienia efektywności i niezawodności działania inteligentnych maszyn podaje w wątpliwość sam sens projektowania urządzeń posiadających takie doznania. Tegmark wskazuje jednak na trzy wyjątki. Po pierwsze, niezbędny mógłby się okazać – zaprojektowany lub powstały emergentnie – odpowiednik instynktu samozachowawczego, który zapewniałby, że maszyna będzie dążyła do realizacji zaprogramowanych celów przed jej wyłączeniem. Po drugie, jest wielce prawdopodobne, że z powodu „[…] łatwego kopiowania informacji i oprogramowania między sztucznymi inteligencjami" (Tegmark, 2019: 399) urządzenia wyposażone w SI miałyby mniejsze poczucie własnego indywidualnego „ja", czyli ich świadomość mogłaby być w mniejszym stopniu indywidualistyczna, a bardziej holistyczna (a nawet kolektywistyczna). I po trzecie, sztuczna inteligencja mogłaby również posiadać odpowiednik poczucia wolnej woli, które powiązane jest z procesami decyzyjnymi, możliwością wyboru takiej lub innej alternatywy oraz z indeterministyczną nieświadomością wyniku (poczucie niepewności przeddecyzyjnej).

Analizy przeprowadzone w formie hipotetycznych, spekulatywnych scenariuszy nie są, rzecz jasna, w stanie przewidzieć, czy i kiedy świadoma sztuczna inteligencja się pojawi. Mimo to wydają się uprawnione, gdyż eksplorują sferę różnych możliwych postaci, jakie mogłaby ona przyjąć, zważywszy na ograniczenia wypływające z obowiązujących teorii

naukowych, a przy okazji dopuszczają perspektywę porównawczą, w której zestawia się taką hipotetyczną sztuczną świadomość ze świadomością o podłożu biologicznym. To ostatnie jest o tyle ciekawe, że pozwala zająć stanowisko w sprawie starego dylematu: czy świadoma SI przybierze postać ludzką, czy nieludzką. Na przykład w proponowanym przez Tegmarka ujęciu syntetycznej świadomości uda się przezwyciężyć liczne ograniczenia wynikające z antropologii człowieka, choć z drugiej strony – nie będzie ona aż tak różna od ludzkiej, aby nie można było ich zestawiać i porównywać ze sobą na różnych płaszczyznach.

Analizując hipotetyczne scenariusze na temat możliwego kształtu przyszłej świadomej SI, rzuca się w oczy przede wszystkim spekulatywność i „fantazyjność" tego typu propozycji, choćby w porównaniu ze „skromniejszymi" projektami wyposażenia LLM-ów w globalną przestrzenią roboczą, przetwarzanie rekurencyjne czy zdolność rozumienia (zob. s. 73–75 niniejszego tekstu). Wynika to stąd, że autorzy zajmujący się tego typu hipotetycznymi scenariuszami wykorzystują możliwość spekulowania na temat docelowego i ostatecznego kształtu świadomej SI, bez kłopotania się o ewentualne etapy pośrednie jej rozwoju i warunki początkowe, w jakich może ona powstać. Mimo to sądzę, że to właśnie ten sposób dociekań wyznacza w dużej mierze horyzont myślenia o ewentualności pojawienia się w przyszłości czujących, myślących i samoświadomych maszyn, w ramach którego kształtują się główne podejścia i narracje wykorzystywane w kulturowym i naukowym dyskursie na ten temat. Można podejrzewać, że podejście to ma potencjał zwrotnego inspirowania całych rzesz badaczy i podpowiadania im potencjalnych obszarów naukowej i filozoficznej eksploracji.

Zakończenie

Trzy omówione w artykule zagadnienia pokazują, jak trudne są problemy dotyczące ewentualności powstania świadomej sztucznej inteligencji. Problemy te tworzą zazębiający się zestaw tematów, które tylko częściowo poddają się obecnie naukowej empirycznej weryfikacji. Nawet przywoływana często w tym kontekście „współpraca interdyscyplinarna" (np. nauk informatycznych i neuronauki) nie jest w stanie rozwiązać większości z nich. Z uwagi na samą naturę doznań świadomych, jak i niepewność wnioskowań na temat przyszłości rozwoju technologicznego, dzisiejsze myślenie na temat świadomej SI nie może obyć się bez filozoficznych analiz, jak również bez pewnej dozy śmiałych naukowych spekulacji.

Problem ograniczonych możliwości rozpoznania, czy SI faktycznie uzyskała świadomość, czy algorytm jedynie symuluje posiadanie stanów świadomych i doznań, należy niewątpliwie do najtrudniejszych filozoficznie zagadnień, którego rozwiązania ciągle nie widać na horyzoncie. Jak starałem się pokazać, próba rozstrzygnięcia tego problemu poprzez przeniesienie oceny na płaszczyznę behawioralno-językową (Turing, Schneider) jedynie częściowo jest satysfakcjonująca. Nieco lepszym – choć ciągle niepewnym – rozwiązaniem jest w tym przypadku dopiero połączenie testowania behawioralnego z podejściem funkcjonalnym: jeśli system sztucznej inteligencji zachowywałby się jak świadomy i doznający podmiot, a zarazem jego funkcjonalna architektura byłaby w jakiejś mierze odwzorowaniem architektury biologicznej mózgu człowieka czy zwierzęcia, to wzmacniałoby to nasz stopień wewnętrznego przekonania, że mamy do czynienia z obdarzonym świadomością systemem sztucznej inteligencji.

Kolejny ze wskazanych przeze mnie problemów – dotyczący sposobu powstania świadomej SI – nie jest już aż tak filozoficznie trudny do przejścia, ale niewątpliwie również należy do kategorii problemów trudnych z uwagi na to, że wiąże się z próbą przewidywania przyszłości. W artykule starałem się pokazać, że najbardziej owocne w tym względzie jest przyjrzenie się podejściom opisującym, *jak* mogłoby dojść do powstania świadomej SI, a nie *czy* i *kiedy* to się stanie. Oba wymieniane w tym kontekście ujęcia – projektowe i emergentystyczne – wyznaczają dość szerokie ramy dla możliwości jej pojawienia się. Podejście projektowe niesie obietnicę większej kontroli nad procesem wyłaniania się sztucznej świadomości, ale jesteśmy prawdopodobnie jeszcze daleko od stworzenia technologii, która by to umożliwiała. Dodatkowo na tym podejściu cieniem kładą się trudne obecnie do rozstrzygnięcia problemy natury ogólnej – na przykład czy i w jakim stopniu da się odtworzyć organizację funkcjonalną układów biologicznych w układach sztucznych, a także, czy odtworzenie organizacji funkcjonalnej mechanizmów biologicznych w środowisku syntetycznym przełoży się rzeczywiście na wyłonienie się świadomości fenomenalnej i doznań subiektywnych. Z kolei poleganie na podejściu emergentystycznym wydaje się interesujące, gdyż w jakimś sensie obiecuje ono powtórzenie – choć w szybszym tempie – naturalnego procesu powstania świadomości u żywych organizmów na ziemi w zastosowaniu do układów sztucznych. Podejście to ma jednak tę wadę, że proces emergencji świadomości jest w jakiejś mierze poza naszą kontrolą, a kolejne jego etapy są trudne do przewidzenia.

Trzeci problem, którym zająłem się w artykule, dotyczył wnioskowania o cechach przyszłej świadomej sztucznej inteligencji z perspektywy

uwzględniającej możliwości i ograniczenia wynikające z praw fizyki i z pozostałych uwarunkowań świata fizycznego. Ten sposób wnioskowania na temat świadomej SI, pomimo że wydaje się dość spekulatywnym przedsięwzięciem, ma tę zaletę, że jest w stanie z grubsza zakreślić obszar możliwych postaci, jakie mogłaby przybrać przyszła syntetyczna świadomość. Ujęcie to, wzbogacone na przykład o perspektywę porównawczą, jest też w stanie już obecnie – nawet w sytuacji braku technologii budowy świadomej SI – wskazać na potencjalne różnice pomiędzy syntetyczną świadomością i biologiczną świadomością człowieka.

Literatura

Alemohammad, S., Casco-Rodriguez, J. (2023). *Self-Consuming Generative Models Go MAD*. arXiv:2307.01850v1

Amstrong, S., Sotala, K. (2016). How We're Predicting AI – or Failing To. W: J. Romportl i in. (red.), *Beyond Artificial Intelligence. The Disappearing Human-Machine Divide* (ss. 11–29). Cham: Springer.

Aru, J., Larkum, M., Shine, J. (2023). The Feasibility of Artificial Consciousness through the Lens of Neuroscience. *Trends in Neurosciences*, *46*(12), 1008–1017.

Avramides, A. (2001). *Other Minds*. London: Routledge.

Baars, B. (1998). *In the Theater of Consciousness: The Workspace of Mind*. New York: Oxford University Press.

Bostrom, N. (2016). *Superinteligencja. Scenariusze, strategie, zagrożenia*. Gliwice: Helion.

Bubeck, S., Chandrasekaran, V. i in. (2023). *Sparks of Artificial General Intelligence: Early Experiments with GPT-4*. arXiv:2303.12712v5

Butlin, P., Bengio, Y. i in. (2023). *Consciousness in Artificial Intelligence: Insights from the Science of Consciousness*. arXiv:2308.08708v1

Chalmers, D. (2003). Zagadka świadomości. *Świat Nauki*, wydanie specjalne, 102–110.

Chalmers, D. (2010). *Umysł świadomy. W poszukiwaniu teorii fundamentalnej*. Warszawa: Wydawnictwo Naukowe PWN.

Chalmers, D. (2023). *Could a Large Language Model be Conscious?* arXiv: 2303.07 103

Christian, B. (2020). *The Alignment Problem. How Can Artificial Intelligence Learn Human Values?* London: Atlantic Books.

Dehaene, S. (2023). *Świadomość i mózg. Odczytywanie kodu naszych myśli*. Kraków: Copernicus Center Press.

Dehaene, S., Lau, H., Kouider, S. (2017). What is Consciousness, and Could Machines Have It. *Science*, *358*, 486–492.

Dreyfus, H. (1992). *What Computers Still Can't Do. A Critique of Artificial Reason*. New York: MIT Press.

Duch, W. (2024). *Sztuczna inteligencja coraz prawdziwsza. Czy właśnie zaczyna czuć i myśleć jak człowiek?* Projekt Pulsar. https://www.projektpulsar.pl/

technologia/2274829,1,sztuczna-inteligencja-coraz-prawdziwsza-czy-wlasnie-zaczyna-czuc-i-myslec-jak-czlowiek.read (dostęp: 18.12.2024).
Esmaeilzadeh, H., Vaezi, R. (2021). *Conscious AI*. arXiv:2105.07879v2
Feinberg, T., Mallatt, J. (2016). *The Ancient Origins of Consciousness. How the Brain Created Experience*. Cambridge, MA: The MIT Press.
Foster, D. (2024). *Generatywne głębokie uczenie. Uczenie maszyn, jak malować, pisać, komponować i grać*. Warszawa: APN Promise.
Ginsburg, S., Jablonka, E. (2019). *The Evolution of the Sensitive Soul. Learning and the Origins of Consciousness*. Cambridge, MA: The MIT Press.
Hoffman, R., GPT-4. (2023). *Rozmowa z chatem GPT o przyszłości ludzi i świata*. Warszawa: Prześwity.
Humphrey, N. (2022). *Sentience. The Invention of Consciousness*. Oxford: Oxford University Press.
Kissinger, H., Schmidt, E., Huttenlocher, D. (2023). *Era sztucznej inteligencji. I nasza przyszłość jako ludzkości*. Warszawa: Wydawnictwo Nowej Konfederacji.
Koch, Ch. (2008). *Neurobiologia na tropie świadomości*. Warszawa: Wydawnictwa Uniwersytetu Warszawskiego.
Koch, Ch. (2019). *The Feeling of Life Itself. Why Consciousness is Widespread But Can't Be Computed*. Cambridge, MA: The MIT Press.
Koch, Ch. (2020). Proust wśród maszyn. *Świat Nauki*, *1*(341), 50–53.
Kurzweil, R. (2013). *Nadchodzi osobliwość. Kiedy człowiek przekroczy granice biologii*. Warszawa: Kurhaus Publishing.
Landgrebe, J., Smith, B. (2023). *Why Machines Will Never Rule the World. Artificial Intelligence without Fear*. New York: Routledge.
LeDoux, J. (2020). *Historia naszej świadomości. Jak po czterech miliardach lat ewolucji powstał świadomy mózg*. Kraków: Copernicus Center Press.
Lee, Kai-Fu. (2019). *Inteligencja sztuczna, rewolucja prawdziwa. Chiny, USA i przyszłość świata*. Poznań: Media Rodzina.
Lem, S. (1984). *Summa Technologiae*. Lublin: Wydawnictwo Lubelskie.
Lemoine B. (2022). *Is LaMDA Sentient? – an Interview*, https://bit.ly/3U6x6kq
Li, D., He, W., Guo, Y. (2021). Why AI Still Doesn't Have Consciousness? *CAAI Transactions on Intelligence Technology*, *6*, 175–179.
Mahowald, K., Ivanova, A.A., Blank, I.A., Kanwisher, N., Tenenbaum, J.B., Fedorenko, E. (2023). *Dissociating Language and Thought in Large Language Models: A Cognitive Perspective*. arXiv:2301.06627v1
Marciszewski, W., Stacewicz, P. (2011). *Umysł-Komputer-Świat. O zagadce umysłu z informatycznego punktu widzenia*. Warszawa: EXIT.
Marcus, G. (2024). *Taming Silicon Valley. How We can Ensure That AI Works for Us*. Cambridge, MA: The MIT Press.
Marcus, G., Davis, E. (2019). *Rebooting AI. Building Artificial Intelligence We Can Trust*. New York: Panthenon Books.
Massimini, M., Tononi, G. (2018). *Sizing Up Consciousness. Towards an Objective Measure of the Capacity for Experience*. Oxford: Oxford University Press.
Meissner, G. (2020). Artifcial Intelligence: Consciousness and Conscience. *AI & SOCIETY*, *35*, 225–235.

Metzinger, T. (2018). *Tunel Ego. Naukowe badanie umysłu i mit świadomego ja*. Łódź: Wydawnictwo Uniwersytetu Łódzkiego.

Mill, J.S. (1889). *An Examination of Sir William Hamilton's Philosophy*. London: Longmans, Green & Co.

Moris, M.R., Sohl-Dickstein, J., Fiedel, N., Warkentin, T., Dafoe, A., Faust, A., Farabet, C., Legg, S. (2024). *Levels of AGI for Operationalizing Progress on the Path to AGI*. arXiv:2311.02462v4

Musser, G. (2024). Naprawdę inteligentna maszyna. *Świat Nauki*, *5*(393), 45–50.

Narayanan, A., Kapor, S. (2024). *AI Snake Oil. What Artificial Intelligence Can Do, What It Can't, and How to Tell the Difference*. Princeton: Princeton University Press.

Poczobut, R. (2024). Czy sztuczna świadomość jest możliwa? *Filozofuj*, *1*(55), 9–12.

Rosenthal, D. (2005). *Consciousness and Mind*. Oxford: Oxford University Press.

Russell, S. (2019). *Human Compatible. AI and the Problem of Control*. New York: Allen Lane.

Russell, S., Norvig, P. (2023). *Sztuczna inteligencja. Nowe spojrzenie*, t. 1. Gliwice: Helion.

Schneider, S. (2021). *Świadome maszyny. Sztuczna inteligencja i projektowanie umysłów*. Warszawa: Wydawnictwo Naukowe PWN.

Searle, J. (1995). Umysły, mózgi, programy. W: B. Chwedeńczuk (red.), *Filozofia umysłu* (ss. 301–324). Warszawa: Aletheia.

Sejnowski, T. (2024). *ChatGPT and the Future of AI. The Deep Language Revolution*. Cambridge, MA: The MIT Press.

Seth, A. (2021). *Being You. A New Science of Consciousness*. London: Faber.

Seth, A., Bayne, T. (2022). Theories of Consciousness. *Nature Reviews Neuroscience*, *23*, 439–452.

Shumailov, I., Shumaylov, Z. i in. (2024). AI Models Collapse When Trained On Recursively Generated Data. *Nature*, *631*, 755–759.

Sui, P., Duede, E. i in. (2024). *Confabulation: The Surprising Value of Large Language Model Hallucinations*. arXiv:2406.04175v2

Suleyman, M. (2024). *Nadchodząca fala. Sztuczna inteligencja, władza i najważniejszy dylemat ludzkości w XXI wieku*. Kraków: Szczeliny.

Tegmark, M. (2019). *Życie 3.0. Człowiek w erze sztucznej inteligencji*. Warszawa: Prószyński i S-ka.

Turing, A. (1995). Maszyna licząca a inteligencja. W: B. Chwedeńczuk (red.), *Filozofia umysłu* (ss. 271–300). Warszawa: Aletheia.

Usidus, M. (2023). Wąż na własnym ogonie się nie pożywi. *Młody Technik*, *11*, 32–36.

van Dijk, B., Kouwenhoven, T. i in. (2023). Large Language Models: The Need for Nuance In Current Debates and a Pragmatic Perspective on Understanding. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. arXiv:2310.19671v2

Warne, R. (2024). Lessons about Human Mind from Artificial Intelligence. *Skeptic Magazine*, *1*(291), 30–35.

Weiskrantz, L. (1997). *Consciousness Lost and Found. A Neuropsychological Exploration*. Oxford: Oxford University Press.

Wolfram, S. (2023). *What is ChatGPT Doing… and Why Does it Work?* Champaign: Wolfram Media, Inc.

Wooldridge, M. (2020). *The Road to Conscious Machines*. *The Story of AI*. London: Penguin Random House.

Xu, Z., Jain, S. i in. (2024) *Hallucination is Inevitable: An Innate Limitation of Large Language Models*. arXiv:2401.11817v1

Yudkowsky, E. (2004). *Coherent Extrapolated Volition*. Machine Intelligence Research Institute.

Zybertowicz, A. (2024). *AI Eksploracja*. Warszawa: Zona Zero.