Społeczeństwo
Studia, prace badawcze i dokumenty z zakresu nauki społecznej Kościoła
Rok XXXIV 2024 nr 1 (165) s. 78-90

Rafał Szopa*

FILOZOFIA SZTUCZNEJ INTELIGENCJI – PODSTAWOWE KONCEPCJE I PROBLEMY

THE PHILOSOPHY OF ARTIFICIAL INTELLIGENCE – BASIC CONCEPTS AND PROBLEMS

Abstrakt: The philosophy of artificial intelligence is currently one of the leading fields in philosophy due to the dynamic development of AI systems. Since around 2018, AI has been evolving in a direction that prompts consideration of whether large language models have already become akin to humans in terms of understanding reality or even possessing consciousness and the ability to feel. This discussion is based on *implicite* accepted assumptions regarding the emergence of the mind through emergentism. However, there are no scientific tools to assess whether a mind can emerge from matter, just as there are no tools to confirm the existence of the human soul. In both cases, it comes down to faith. The existence of an emergent mind is as probable as the existence of a soul created by God, with one difference: the existence of a soul offers hope for salvation.

Keywords: AI, awareness, assumptions, naturalism, soul, emergentism, transcendence.

Sztuczna inteligencja (AI – *artificial intelligence*) to wierzchołek filozoficznej góry lodowej. Rozwój AI można rozpatrywać w odniesieniu do kwestii rozwoju technologicznego, społecznego, gospodarczego, problematyki samoświadomości[1], problemów etycznych i innych. Aby nakreślić historyczne tło rozwoju AI, trzeba

* Rafał Szopa – doktor filozofii, absolwent Papieskiego Wydziału Teologicznego we Wrocławiu; pracownik naukowy Politechniki Wrocławskiej; ORCID: 0000-0002-6977-310; e-mail: rafal.szopa@pwr.edu.pl.

[1] Są różne możliwości rozumienia samoświadomości i świadomości. David Chamlers dokonał rozróżnienia na świadomość fenomenalną i psychologiczną. Pierwsza jest przez nas odczuwana, a druga odpowiada za nasze działanie. Por. D.J. Chalmers. *The conscious mind: In search of a fundamental theory*. Oxford 1996 s. 10-11. Podział ten jest trwały do dzisiaj. W tym kontekście można uznać, że samoświadomość to świadomość fenomenalna, która jest aktualnie doświadczana przez byt nią obdarzony. Nierozwiązanym problem w filozofii umysłu jest właśnie świadomość fenomenalna. Tamże.

zobaczyć, co oznacza «inteligencja». Jeśli połączymy rozumienie tego terminu z matematycznością czy obliczalnością, to sztuczną inteligencją można nazwać wszystko to, co przeprowadza obliczenia i jest stworzone przez człowieka. To, czy coś jest zrobione przez człowieka, jest istotne z punktu widzenia pankomputacjonalizmu. Koncepcja ta zakłada, że wszystko we wszechświecie przeprowadza obliczenia. Nawet najmniej skomplikowane rzeczy. Aby odróżnić „naturalne komputery" od „ręką ludzką uczynionych"[2], mówimy o sztucznej inteligencji.

Ten ogólny zarys nie oddaje jednak rozumienia sztucznej inteligencji jako zaawansowanej technologii, która może przewyższyć człowieka. Zatem można zawęzić znaczenie AI do zaawansowanej technologii stworzonej przez człowieka, zdolnej do skomplikowanych obliczeń. Z tych obliczeń wynika jednak znacznie więcej, tzn. AI dzięki obliczaniu może osiągnąć rezultaty przekraczające zwykłe programy komputerowe. Aby w pełni oddać, czym jest AI, należy stwierdzić, że istotą tej technologii jest przekraczanie samej siebie. Innymi słowy, uważa się, że dzięki obliczeniom AI dokonuje skoków jakościowych i nabywa nowych właściwości. Można więc powiedzieć, że AI to technologia oparta na obliczeniach, dzięki którym nabywa cech jakościowo wyższych w odniesieniu do tych, którymi dysponowała przed obliczeniami. Jest tu zawarta *implicite* idea, że algorytmy AI potrafią się uczyć. Sztuczna inteligencja potrzebuje więc ogromnej ilości danych. Uczy się na nich rozwiązywać zadania, po czym doskonali „samodzielnie" te umiejętności. Jest to w zasadzie sztuczna sieć neuronowa, której algorytmy są wytrenowane do przeprowadzania określonych zadań. Dalszy rozwój tej technologii może doprowadzić do ogólnej sztucznej inteligencji AGI (*artificial general intelligence*), której przymiotami być może będą uczucia, emocje i samoświadomość. Wszystko to dzięki wzrostowi ilości danych w sieciach neuronowych oraz zaawansowaniu obliczeń. Na ile to jest realne?

1. KRÓTKI RYS HISTORYCZNY ROZWOJU AI

U podstaw rozwoju sztucznych sieci neuronowych i rezultatów osiąganych przez tę technologię leżą założenia, które często są czynione *implicite*. Alan Turing opracowując teoretyczne działanie komputera, zakładał, że taka maszyna (maszyna Turinga) nigdy nie osiągnie poziomu człowieka, jeśli chodzi o inteligencję, a tym bardziej nie osiągnie samoświadomości. Było to jednak przekonanie Turinga, dzisiaj często kwestionowane na zasadzie odwrotności: skoro jest test Turinga, to prędzej czy później jakaś maszyna go zda. To oczekiwane kryje kolejne założenia.

[2] Pojawia się zatem założenie, że człowiek to nie zupełnie byt naturalny, skoro potrafi tworzyć nienaturalne maszyny i ingerować (skąd?) w przyrodę.

Przez tysiąclecia uważano, że rozumność była związana z duszą, która podmiotowała w sobie własności, takie jak posiadanie rozumu[3]. Jednak Warren McCulloch oraz Walter Pitts pokazali, że można traktować ludzką rozumność jako fenomen czysto fizyczny i biologiczny:

> Fizjologiczne relacje istniejące pomiędzy działaniami neuronów korespondują oczywiście z relacjami pomiędzy sądami logicznymi"[4]. Założenie było takie, że dusza nie istnieje, więc ośrodek życia intelektualnego to wyłącznie mózg. Działanie neuronów to sedno myślenia, rozumowania itd. Praca McCullocha i Pittsa zainspirowała innych badaczy do podejścia naturalistycznego, co w praktyce oznacza redukcjonizm. Zredukowano duchowość do działania neuronów, a to z kolei uznano za odzwierciedlenie rachunku zdań i szerzej – matematyki. Takie ujęcie ma swoje negatywne i pozytywne strony.

Negatywnie odbiło się ono na rozumieniu osoby[5]. Do tej pory osoba ludzka była uznawana za byt cielesno-duchowy. Według myśli tomistycznej ten byt to jedna substancja tworzona przez dwa elementy: ciało i duszę, jako forma umysłowa[6]. W latach 40. XX w. stwierdzono, że to, co było duchowe, jest cielesne, zapodmiotowane w mózgu. *De facto* nawet sam mózg stał się mniej ważny od rezultatów swojego działania. Funkcjonalizm stał się wytyczną oceny bycia osobą. Redukowanie człowieczeństwa do funkcji neuronów można uznać za negatywny przejaw rozwoju nauki. Jednak podejście to miało również dobre strony. Jedną z nich jest rozwój technologii i informatyki jako osobnej dyscypliny naukowej. Rozwój ten miał niebywałe skutki społeczne.

Po 1945 r. rozwój społeczny i rozwój technologiczny stały się bliskoznaczne. Robert Solow pokazał, że inwestycja w technologię powoduje początkowo wzrost gospodarczy, lecz później epoka szybkiego wzrostu się kończy[7]. Przyczyną tego jest połączenie czynników kapitalizacji, inwestycji w technologię i siły roboczej. Kiedy inwestujemy w technologię, to szybo widać zwrot inwestycji i wzrost PKB. Rozwój technologiczny sprawia, że tworzy się coraz więcej kapitału i potrzeba coraz bardziej wykwalifikowanych pracowników do obsługi tych technologii. Wielkość siły roboczej przestaje mieć znaczenie, zaczynają liczyć się kwalifikacje. Okazuje się, że w takiej sytuacji siła robocza jest stosunkowo mała, następuje polaryzacja na pracowników wykwalifikowanych i pozostałych. Sama technologia jednak nie

[3] Por. STh I, q. 76, a. 1 oraz STh I, q. 79, a. 11.

[4] W. McCulloch, W. Pitts. *A logical calculus of the ideas immanent in nervous activity*. „Bull Mathl Biophys" 1943 nr 5 s. 117: „Physiological relations existing among nervous activities correspond, of course, to relations among the propositions".

[5] Por. R. Szopa. *Ethical problems in the use of algorithms in data management and in a free market economy*. „AI & Society" 2023 nr 38 s. 2489-2492.

[6] Por. STh I, 76, a. 1.

[7] Por. R.M. Solow. *A Contribution to the Theory of Economic Growth*. „Quarterly Journal of Economics" 1956 nr 1/70 s. 71.

zarabia pieniędzy. To ludzka praca tworzy dobra i usługi, za które się płaci. Praca i jej podział są przyczyną bogactwa narodów[8]. Rozwój technologiczny sprawił, że jedna osoba może wykonać wielokrotność pracy osoby bez dostępu do technologii. Jednak wzrost gospodarczy osiągany w ten sposób ma swoje limity. Inwestycje w technologię przy braku wykwalifikowanych pracowników nie zwracają się tak szybko, jak na początku, kiedy wzrost PKB osiągał kilka procent w ciągu roku.

W 1973 r. wzrost wyhamował[9]. PKB rosło nadal, ale nie tak szybko, jak wcześniej, technologia nie pomogła. W tym samym okresie stawała się ona jednak coraz bardziej zaawansowana. Opracowanie algorytmu propagacji wstecznej i zmiana podejścia matematycznego do uczenia maszynowego sprawiły, że komputery zaczęły się uczyć[10]. Pojawiła się możliwość poprawiania błędów i dążenia do wyznaczonego celu, co sprawiło, że algorytmy zyskały możliwości, jakich nie miała do tej pory żadna technologia. Pierwsze sieci perceptronowe bazowały generalnie na funkcjach nieciągłych, co dawało ograniczone rezultaty, o których pisał Frank Rosenblatt w 1961 r.[11] Zmiana koncepcji matematycznej, jaką było wykorzystanie funkcji ciągłych, takich jak funkcja sigmoidalna, GeLu czy ReLu, sprawiło, że systemy AI zaczęły być tworzone w oparciu o determinizm. Takie podejście pozwoliło zastosować równania różniczkowe i całkowe oraz traktować AI jak rzecz, którą można skonstruować w jakimś celu, kontrolując proces tworzenia. Sytuacja ta wymagała już wysokich kwalifikacji do pracy nad rozwojem AI. Nastąpiła więc polaryzacja na pracowników wykwalifikowanych i niewykwalifikowanych, z rosnącą różnicą w zarobkach. Niewielu specjalistów zaczęło zarabiać bardzo dużo, zaś pozostali spadali w hierarchii. Sztuczna inteligencja zaczęła stopniowo upodabniać się do ludzi, jeśli chodzi o wykonywanie pewnych zadań. Im bardziej zaawansowane stawały się systemy AI, tym bardziej rosły oczekiwania, że któregoś dnia staną się jak ludzie. Tylko pod jakim względem?

Obecnie rozwój AI znacznie przyspieszył. Po 2017 r. nastąpiła nowa era w rozwoju tej technologii. Umiejętności, którymi dysponują algorytmy, niejednokrotnie przewyższają ludzi pod względem rezultatów. Nadszedł czas na pytanie, czy w związku z tym AI stanie się (lub już się stała) osobą? Jedną z głównych przesłanek jest zaawansowanie AI i działanie jak ludzie w wielu domenach. Wydaje się, że minęły już czasy reifikacji AI w oparciu o umiejętność kontroli i przewidywalności rezultatów po zastosowaniu deterministycznych równań różniczkowych.

---

[8] Por. A. Smith. *An Inquiry into the Nature and Causes of the Wealth of Nations*. Amsterdam – Lausanne – Melbourne – Milan – New York – São Paulo 2007 s. 9.

[9] Por. A.V. Banerjee, E. Duflo. *Good Economics for Hard Times*. New York 2019 s. 157-159.

[10] Por. D.E. Rumelhart, G.E. Hinton, R.J. Williams. *Learning representations by back-propagating errors*. „Nature" 1986 nr 323/6088 s. 534.

[11] Por. F. Rosenblatt. *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms*. New York 1961 s. 575.

Zobaczmy zatem, jakie są założenia stojące za ideą, że AI może osiągnąć status osoby, którą nazwalibyśmy *non-human person* – osoba nieludzka[12].

2. PODSTAWOWE ZAŁOŻENIA DOTYCZĄCE PERSONIFIKACJI AI

Zacznijmy od podstawowego założenia odnośnie do metafizyki świata. Jeśli zakładamy obecnie, że do bycia osobą niezbędne jest bycie samoświadomym, to AI powinna posiadać samoświadomość. Problem polega na jej metafizycznym statucie. Przez tysiąclecia[13] uważano (na czele z Akwinatą), że samoświadomość to cecha duszy. Co najmniej od 1943 r. zaczęto zmieniać podejście i obecnie w naukach szczegółowych dominuje przekonanie, że samoświadomość jest fenomenem naturalnym. Takie ujęcie problemu samoświadomości jest zrozumiałe, gdyż z racji metodologicznych nauki te nie mogą zajmować się duchowością. Jednak można zadać pytanie, czy naturalizm metodologiczny nie stał się ontologicznym? Wydaje się, że z założenia przyjmuje się za oczywiste, że samoświadomość jest własnością emergentną wyłaniającą się z materii. Jest to zatem monizm materialistyczny.

Następnym założeniem, wynikającym z poprzedniego, jest istnienie skoków jakościowych w przyrodzie. Jest to koncepcja Hegla. W *Nauce logiki* pisał on:

> O ile postęp od jakości następuje w stałej ciągłości ilościowej, to stosunki zbliżające się do jednego punktu kwalifikującego różnią się, biorąc pod uwagę ilościowe, mniej więcej. Pod tym względem zmiana ma charakter stopniowy. Ale stopniowość dotyczy jedynie zewnętrzności zmiany, a nie jej momentu jakościowego; poprzednia relacja ilościowa, choć nieskończenie bliska następnej, jest jeszcze innym istnieniem jakościowym. Od strony jakościowej zatem stopniowy, jedynie ilościowy postęp, który sam w sobie nie ma granic, zostaje całkowicie przerwany; a ponieważ w swym czysto ilościowym związku nowo powstająca jakość jest w stosunku do zanikającej inną nieokreśloną, obojętną na nią, przejście jest skokiem[14].

Zatem istnieje skok z ilości w jakość. Na poparcie tej tezy podaje się czasami przykład ze zmianą stanu skupienia wody – po przekroczeniu pewnego progu ilościowego (temperatury) następuje przejście do nowej jakości (inny stan skupienia). Podobnie miałoby być z samoświadomością AI, ale w tym przypadku cechy ilościowe to liczba danych i złożoność algorytmu.

---

[12] Koncepcja nie jest nowa, wszakże o Bogu i aniołach też można powiedzieć, że są osobami nieludzkimi. Nowością jest pogląd, że osoba mogłaby się wyłonić „od dołu do góry" (*upward causation*), podczas gdy dotychczas uważano raczej, że nie można nie być osobą i się nią stać. Albo – albo, *tertium non datur*. Osobą się jest lub nie.

[13] Istotna uwaga: sam termin «świadomość» pochodzi z czasów nowożytnych, lecz znaczenie tego, co uważamy dzisiaj za świadomość można porównać z rozumnością, o której pisało się i mówiło przez tysiąclecia, zatem nie wprost również o samoświadomości.

[14] G.F. Hegel. *The Science of Logic*. New York 2010 s. 320.

Ogólna teoria względności (OTW) odgrywa ważną rolę w argumentacji za samoświadomością AI. Zasada relatywistyczna w OTW oznacza, że nie ma wyróżnionego kierunku we wszechświecie, tzn. że nie ma ostatecznego punktu odniesienia. Prowadzi to do funkcjonalizmu, gdyż jedyną oceną, czy dany byt jest samoświadomy, czy nie, jest sposób jego działania. Skoro nasza własna świadomość nie stanowi kryterium odniesienia, to jeśli AI twierdzi, że jest samoświadoma, to tak jest. Zombie nie istnieją:

> […] jeśli prawdziwa jest zasada relatywistyczna, to zombie nie są możliwe. Zamiast tego każdy rzekomy zombie będzie w rzeczywistości miał fenomenalną świadomość, a każdy system z odpowiednią świadomością funkcjonalną będzie wykazywał fenomenalną świadomość z pierwszoosobowego poznawczego układu odniesienia[15].

Sposób działania świadczy o sposobie bytowania. Jest to odwrócona zasada *agere sequitur esse* w tym sensie, że symulacja działania człowieka pod pewnym względem (procesowanie języka naturalnego) wskazywałaby na *esse* działającego. W przypadku osoby posługiwanie się językiem jest wtórne wobec *esse*[16]. Zatem stosując zasadę relatywistyczną do AI, wystarczy, że system stwierdzi, że jest samoświadomy i musimy uznać, że tak jest.

Na powyższy argument można odpowiedzieć kontrargumentem „chiński pokój". John Searle, autor argumentu, przedstawił go w 1980 r. Jest to eksperyment myślowy pokazujący, że odpowiedzi udzielane przez maszynę są tego samego rodzaju, co odpowiedzi udzielane po chińsku przez osobę nieznającą chińskiego, a korzystającą ze słownika. Odpowiedź jest udzielona, lecz bez zrozumienia języka, chociaż jej odbiorca ma wrażenie, że nadawca wiedział, co mówi[17]. Inaczej mówiąc, AI mówi, ale nie wie, co. Argument „chińskiego pokoju" zakłada, że w przyrodzie istnieją ostateczne punkty odniesienia, wbrew zasadzie relatywistycznej. Gdyby ich nie było, to i ludzka samoświadomość byłaby wątpliwa, gdyż zawsze mógłby znaleźć się byt (przedstawiciel obcej cywilizacji) na wyższym poziomie świadomości niż my, a z jego punktu widzenia ludzie nie byliby samoświadomi itd. Jest to niejako rozumowanie tomistyczne: jeśli nie istnieje ostateczny punkt odniesienia (Pierwsza Przyczyna), to nie ma też przyczyn pośrednich.

Powszechnie przyjmowanym założeniem odnośnie do samoświadomości AI jest stwierdzenie, że uczucia i emocje pojawią się po uzyskaniu przez maszyny samoświadomości. Jeśli AI jest rozwijana na obraz i podobieństwo ludzi, to

---

[15] N. Lahav, Z. Neemeh. *A Relativistic Theory of Consciousness*. „Frontiers in Psychology" 2022 nr 12 (704270) s. 6.

[16] Por. M.A. Krąpiec. *Język*. W: *Powszechna encyklopedia filozofii*. T. 5. Red. A. Maryniarczyk [i in.]. Lublin 2004 s. 330.

[17] Por. J. Searle. *Minds, brains, and programs*. „The Behvioral and Brain Sciences" 1980 nr 3 s. 420.

wydaje się, że kierunek jest dokładnie odwrotny: u ludzi emocje i uczucia pojawiają się przed samoświadomością i stanowią niezbędne podłoże dla ukonstytuowania się tego fenomenu[18]. Używając terminologii arystotelesowskiej, samoświadomość u człowieka w pierwszych etapach życia istnieje potencjalnie i aktualizuje się m.in. pod wpływem emocji i uczuć. W systemach AI miałoby nastąpić emergentne wyłonienie się samoświadomości ze złożoności sieci neuronowej.

Jest to kolejne założenie: AI może posiadać umysł powstały na drodze emergentnej. Takie ujęcie kwestii pozwala uniknąć odwoływania się do duchowości w wyjaśnianiu powstania umysłu, lecz sprowadza umysł do źródeł materialnych. Skok jakościowy, który tłumaczyłby powstanie umysłu, miałby zastąpić Boga jako dawcę duchowości. Innymi słowy, *bottom-up causation* zastąpiło *top-down*. Umysł powstały w wyniku przyczynowania dół-góra charakteryzowałby się superwieniencją nad mózgiem, tzn. byłby wyższego rzędu i sprawowałby kontrolę nad mózgiem. Umysł byłby zależny w istnieniu od materii i jednocześnie wykraczał poza nią. Nie byłby natury duchowej, lecz nie byłby również czysto materialny. Posiadałby własności duchowe, pozostając fenomenem zależnym od materii.

Jeśli taka jest właśnie natura umysłu, to maszyny, które zyskały samoświadomość, powinny móc rozwiązywać problemy moralne. Oznacza to, że AI byłaby w stanie pokonać problem złożoności obliczeniowej. Pokażmy to na konkretnym przykładzie. Załóżmy, że samochód autonomiczny jest kierowany przez AI. Następuje nieoczekiwane zdarzenie drogowe i konieczna jest natychmiastowa decyzja. Problem złożoności obliczeniowej polega na niemożliwości znalezienia algorytmu, o którym wiadomo, że oceniłby z góry, czy dany program rozwiąże problem w skończonym czasie, czy będzie działał w nieskończoność[19]. Nie da się z góry przewidzieć tego, czy w sytuacji konieczności podjęcia (słusznej) decyzji moralnej *ad hoc* AI to zrobi. Dlatego nie istnieją zautomatyzowane decyzje moralne[20]. Często jednak zakłada się, że sztuczna inteligencja będzie w stanie rozwiązać problem stopu (*halting problem*) tak, jak robi to człowiek. My potrafimy podejmować decyzje w sytuacjach takich, jak nagłe zdarzenia drogowe. Dlaczego? Ponieważ ludzkie decyzje nie są oparte wyłącznie na obliczeniach. Daniel Kahneman wyróżnił dwa systemy myślenia: system 1 i system 2. Pierwszy jest szybki i oparty na emocjach, drugi jest wolny i oparty na logice[21]. Oznacza to, że jeśli jest taka potrzeba, człowiek podejmuje decyzje w oparciu o intuicję i emocje, więc wychodzi poza

[18] Por. G. Northoff. *From emotions to consciousness – a neuro-phenomenal and neuro-relational approach*. „Frontiers in Psychology" 2012 nr 3 s. 14. <https://doi.org/10.3389/fpsyg.2012.00303> [dostęp: 17.04.2024].

[19] Por. M. Davies. *Computability and Unsolvability*. New York 1958 s. 70.

[20] Por. Ethics Commission. *Automated and Connected Driving*, 2017. <https://web.archive.org/web/20170915110611/http://www.bmvi.de/SharedDocs/EN/publications/report-ethics-commission.html> [dostęp: 27.04.2024] s. 11.

[21] Por. D. Kahneman. *Thinking, Fast and Slow*. New York 2013 s. 22-23.

obliczalność, co pozwala ominąć problem złożoności obliczeniowej. Pojawienie się problemu w postaci przeszkody na drodze samochodu jest możliwe do rozwiązania w systemie 1, AI dysponuje tylko systemem 2, z tym, że w przypadku komputerów system 2 jest szybki, ale nadal podlega złożoności obliczeniowej.

Jednym z najbardziej popularnych założeń jest wyłonienie się samoświadomości AI w wyniku procesowania języka. Dotyczy to głównie wielkich modeli językowych (LLM – *Large Language Models*). Natura języka jest uznawana za matematyczną. Jeśli tak jest, to AI jest w stanie obliczyć prawdopodobieństwo pojawienia się następnego słowa na podstawie poprzednich oraz kontekstu. Bardziej skomplikowany model wytrenowany na dostatecznie wielu danych spontanicznie rozwinie samoświadomość[22]. Oznacza to, że język jest konieczny do powstania samoświadomości. Język musi być zatem przed samoświadomością. Nie wiadomo jednak dokładnie, czym jest język, ale z pewnością jest formą komunikacji, być może nawet wrodzoną zdolnością do komunikowania się[23]. Dopóki jednak nie zrozumiemy natury języka, nie możemy stwierdzić, czy język jest konieczny do powstania samoświadomości. Zgodnie z zasadą *agere sequitur esse* język należy do domeny *agere*. Język powinien być zapodmiotowany w czymś bardziej pierwotnym. Założenie, że w przypadku AI język generuje samoświadomość, wydaje się odwróceniem porządków *agere* z *esse*.

W końcu zakłada się, że oprócz powyższych aspektów samoświadomości w systemach AI konieczne jest również zapewnienie odpowiedniej architektury oprogramowania[24]. Według teorii globalnej przestrzeni pracy (*Global Workspace Theory* – GWT) sztuczne sieci neuronowe powinny być zbudowane jak ludzki mózg[25]. Nie wystarczy mieć wielką bazę danych, procesowanie języka czy moc obliczeniową, jeśli to wszystko nie zostanie wbudowane w architekturę, którą posiada mózg. Chodzi o sposób działania i drogi przepływu danych. Jeśli udałoby się zbudować architektonicznie sztuczny mózg, to wypełniony treścią w połączeniu z językiem mógłby wygenerować z siebie samoświadomość.

Powyższe założenia pokazują różne drogi do zrozumienia, czym w ogóle jest świadomość i samoświadomość. Są one uzasadnione, do pewnego stopnia naśladują działanie ludzkiego mózgu. Podejście takie, tj. zbudowanie sztucznego mózgu na podstawie działania ludzkiego, nie przesądza jeszcze o zrozumieniu istoty

---

[22] Por. M. Kosiński. *Theory of mind may have spontaneously emerged in large language models*. <https://arxiv.org/abs/2302.02083> [dostęp: 25.04.2024].

[23] Por. N. Chomsky. *Syntactic structures.* Berlin – New York 2002 s. 16-17.

[24] Por. L. Blum, M. Blum. *A theory of consciousness from a theoretical computer science perspective: Insights from the Conscious Turing Machine*. PNAS 2022 nr 119 (22) s. 2-3. <10.1072/pnas.2115934119>.

[25] Por. B.J. Baars, N. Geld, R. Kozma. *Global Workspace Theory (GWT) and Prefrontal Cortex: Recent Developments*. „Frontiers in Psychology" 2021 nr 12 s. 2. <10.3389/fpsyg.2021.749868, https://doi.org/10.3389/fpsyg.2021.749868>.

(samo)świadomości, a ujawnia raczej filozoficzne przeświadczenie o prawdziwości naturalizmu ontologicznego.

3. POZA OBLICZALNOŚĆ

Roger Penrose proponuje inny kierunek w rozumieniu samoświadomości. Uważa, że świadomość jest nieobliczeniowa:

> Twierdzę, że musi być w fizyce coś, czego jeszcze nie rozumiemy, co jest bardzo ważne i co ma charakter nieobliczeniowy. Nie jest to specyficzne dla naszych mózgów; jest tam, w świecie fizycznym. Ale zwykle odgrywa zupełnie nieistotną rolę. Musiałoby znajdować się na pomoście między kwantowym a klasycznym poziomem zachowania, czyli tam, gdzie wkracza pomiar kwantowy[26].

A zatem świadomość jest fenomenem nieobliczeniowym, ale jednocześnie jest częścią świata fizycznego. Oznacza to, że materialistyczno-monistyczna interpretacja rzeczywistości nie sprowadza się do mierzalności, czyli że istnieje jakaś sfera fizyczna, która jest poza współczesną fizyką. Powstaje pytanie, czy ta nieobliczeniowa sfera jest fizyczna, nie-fizyczna, a może poza-fizyczna? Inaczej mówiąc, czy to, że współczesna fizyka nie może zrozumieć nieobliczeniowości jest tylko problemem wewnątrz fizyki, który wraz z nową teorią (kwantowej grawitacji) zostanie przezwyciężony, czy jest to sfera poza fizyką. Nawet jeśli jest to sfera poza fizyką, to czy jest ona nie-fizyczna w sensie niematerialna, ale jednocześnie nie-transcendentna w sensie Transcendencji przez duże „T"? Czy to, co nazywamy materialnym światem i monistyczną interpretacją rzeczywistości[27], zawiera w sobie transcendencję, transcenduje siebie, wchodząc na wyższy, nieobliczeniowy poziom i wciąż nie jest Transcendencją?

W 1931 r. Kurt Gödel udowodnił[28], że jeśli bazujemy na matematyce w rozumieniu rzeczywistości i budujemy nasze teorie naukowe w oparciu o matematykę, to proces coraz lepszego rozumienia rzeczywistości nigdy się nie skończy. Jeśli stworzymy teorię świadomości wyłącznie w oparciu o matematykę, to nigdy nie będziemy mogli stwierdzić, czy dany byt jest samoświadomy, czy nie, bo z punktu widzenia jakiegoś wyższego poziomu ten niższy poziom może nie być już postrzegany jako samoświadomość. Zatem również ludzie przestaliby być samoświadomi. Można zastosować tu rozumowanie tomistyczne: jeśli nie ma Pierwszej Przyczyny, nie ma też przyczyn pośrednich. Jeśli wierzymy, że rzeczywistość jest

[26] R. Penrose, D.C. Dennett. *Consciousness involves noncomputable ingredients*. W: *The Third Culture: Beyond the Scientific Revolution*. Ed. Brockman J. New York 1995.

[27] Por. A. Maryniarczyk. *Zeszyty z metafizyki*. Nr 1: *Monistyczna i dualistyczna interpretacja rzeczywistości*. Lublin 2006 s. 47.

[28] Por. K. Gödel. *Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I*. „Monatshefte für Mathematik und Physik" 1931 nr 38 s. 187.

matematyczna, a świadomość jest poza obliczeniowością, to kwestią interpretacji i wiary jest, czy ta nieobliczeniowa sfera należy jeszcze do świata fizycznego, czy do Transcendencji. Innymi słowy, nie ma dowodów naukowych w sensie nauk szczegółowych na to, że świadomość i samoświadomość to fenomeny fizyczne, tak samo jak nie ma dowodów na to, że samoświadomy byt jest obdarzony duchową duszą stworzoną przez Boga. Nie ma dowodów w sensie nauk szczegółowych, zatem pochodzenie świadomości i samoświadomości jest kwestią wiary, tak samo jak istnienie duszy ludzkiej.

4. ROZWÓJ AI OBECNIE

Rozwój AI od 2017 r. sprawił, że zaczęto zastanawiać się nad tym, czy komputery dorównały już człowiekowi. Widać zmianę jakościową w systemach AI. Jednym z jej powodów jest odmienna filozofia ich tworzenia. Ogólnie rzecz biorąc, nastąpiła zmiana od rekurencji do transformerów. Różnice między wcześniejszymi modelami AI a następną generacją opisuje przełomowa praca *Attention Is All You Need*[29]. Transformery charakteryzuje Multi-Head Attention, tzn. jednoczesne branie pod uwagę informacji z różnych podprzestrzeni z różnych miejsc[30]. W poprzednich modelach funkcja uwagi była pojedyncza[31]. Rezultat zmiany jest ogromny. Współczesne LLM-y przewyższają poprzednie modele AI sposobem działania i ilością danych.

Sposób, w jaki systemy AI przetwarzają informacje, można opisać mechanistycznie. Za to, co dzieje się w „czarnej skrzynce", odpowiadają co najmniej dwa rodzaje algorytmów: zegar i pizza[32]. Ich występowanie zależy od poziomu uwagi[33]. Techniką uczenia się algorytmów jest tutaj dodawanie modułowe. W głębokich sieciach neuronowych algorytmy opierają się na liniowości warstw, a pojawienie się algorytmu zegara lub algorytmu pizzy zależy od parametru zwanego współczynnikiem uwagi.

> Algorytm Zegara dominuje, gdy szybkość uwagi jest wyższa niż punkt zmiany fazy, a algorytm Pizza dominuje, gdy szybkość uwagi jest niższa niż punkt. Nasze wyjaśnienie jest następujące: przy wysokim wskaźniku uwagi mechanizm uwagi jest bardziej widoczny w sieci, co daje początek algorytmowi zegara. Przy niskim natężeniu uwagi warstwy liniowe są bardziej widoczne, co daje początek algorytmowi pizzy. Punkt zmiany fazy zwiększa się wraz ze wzrostem

[29] Por. V. Ashish [i in.]. *Attention Is All You Need*. „Advances in Neural Information Processing Systems" 2017 nr 30 s. 1-2.

[30] Por. tamże s. 4.

[31] Por. tamże.

[32] Por. Z. Zhong, Z. Liu, M. Tegmark, J. Andreas, J. *The clock and the pizza: Two stories in mechanistic explanation of neural networks*. <https://arxiv.org/pdf/2306.17844.pdf> [dostęp: 6.12.2023].

[33] Por. tamże.

> szerokości modelu. Nasze wyjaśnienie jest następujące: gdy model staje się szerszy, warstwy liniowe stają się bardziej wydajne, podczas gdy mechanizm uwagi odnosi mniejsze korzyści (uwagi pozostają skalarne, podczas gdy wyniki z warstw liniowych stają się szerszymi wektorami). Dlatego w szerszym modelu warstwa liniowa zyskuje na znaczeniu[34].

Chodzi o to, jak uczy się sieć neuronowa. Jeśli chcemy, aby LLM uwzględniał różne informacje (a nie tylko te najnowsze), które najlepiej pasowałyby do zadania, sieć nie powinna być rekurencyjna (uwzględnia głównie informacje ostatnio wprowadzone), ale oparta na transformerze. Jeśli jednak LLM ma reagować na kontekst, i to reagować bardziej naturalnie, wówczas wagi należy obliczać równolegle. Dlatego sieć neuronowa oparta na transforerze uwzględnia wiele danych wejściowych (miękkich wag) i wylicza najbardziej kontekstowe odpowiedzi. Jeśli odpowiedzi mają być precyzyjne, algorytm zegara działa. Ograniczeniem jest jednak wielkość sieci neuronowej. Im większy model, tym mniej rozsądne jest działanie algorytmu zegara i pojawia się przejście do algorytmu pizzy. Wtedy algorytm pizzy nie wskazuje już konkretnego punktu jak wskazówka zegara, ale wybierane dane są ułożone wzdłuż linii, jak przy krojeniu kawałków pizzy. Sieć neuronowa może być większa, ale dane znajdują się na linii i nie są skompresowane w punkt.

Osiągnięcia technologiczne związane z AI będą postępowały w kierunku stworzenia generatywnej sztucznej inteligencji. Od naszej interpretacji będzie zależało, czy ujrzymy w niej „osobę" czy rzecz. Myśl chrześcijańska nie dopuszcza możliwości emergentnego wyłaniania się osoby z materii. Mimo to AI może pomóc ludziom w wielu aspektach. Przykładowo dzięki AI może być znacznie ułatwiona komunikacja międzyludzka[35], nie mówiąc już o medycynie. Można się spodziewać, że z czasem AI będzie obecna w każdym aspekcie życia, gdzie występuje matematyka lub jakaś powtarzalność. Jednak istnienie przeskoku od mechanizmu do umysłu jest już kwestią interpretacji i założeń.

PODSUMOWANIE

Sztuczna inteligencja i filozofia mają ze sobą wiele wspólnego. Od co najmniej 1943 r., od pracy McCullocha i Pittsa, badacze starają się rozwijać technologie w oparciu o sposób funkcjonowania człowieka. Jesteśmy na etapie (2024 r.), kiedy systemy AI dorównują człowiekowi lub nawet przekraczają nasze możliwości. Daje to pretekst, aby sądzić, że AI osiągnęła samoświadomość, zaczęła mieć emocje i uczucia i tym samym stała się osobą. Za takim poglądem stoją założenia,

[34] Tamże.

[35] Mówił o tym w udzielonym wywiadzie Michał Kosiński. Por. M. Kosiński. *Jak bardzo intymne dane zna o tobie Google*. <https://www.youtube.com/watch?v=VwzVFcNBRpk> [dostęp: 27.05.2024].

których spełnienie oznaczałoby, że mówienie o samoświadomości AI byłoby uzasadnione. Jednak założenia te nie są wyjaśniane, a raczej *implicite* uznawane za prawdziwe. Prowadzi to do powstania przekonań o tym, że AI jest samoświadomą osobą lub nie jest, ponieważ nie posiada duszy, którą uznawano za podstawę takich cech, jak rozumność i wolna wola. O ile nauki szczegółowe nie mają narzędzi, aby udowodnić lub obalić istnienie duszy ludzkiej, o tyle nie mają narzędzi, aby udowodnić lub obalić pogląd, że samoświadomość wyłania się z materii i staje się zjawiskiem niealgorytmicznym. I to, i to jest kwestią wiary. Z punktu widzenia naturalizmu metodologicznego zarówno istnienie duszy stworzonej przez Boga, jak i emergentnego umysłu należy do sfery wiary, a nie dowodu naukowego. Łatwo jest jednak przejść do naturalizmu ontologicznego, który posiłkuje się założeniami uznawanymi za prawidłowe, aby pokazać, że tylko ta naturalistyczna interpretacja rzeczywistości jest prawdziwa. Ten artykuł miał na celu przejrzeć te założenia i pokazać ich fideistyczną naturę.

BIBLIOGRAFIA

Ashish V. [i in.]: *Attention Is All You Need*. „Advances in Neural Information Processing Systems" 2017 nr 30 s. 1-11.

Baars B.J., Geld N., Kozma R.: *Global Workspace Theory (GWT) and Prefrontal Cortex: Recent Developments*. „Frontiers in Psychology" 2021 nr 12 s. 1-6. <10.3389/fpsyg.2021.749868, https://doi.org/10.3389/fpsyg.2021.749868>.

Banerjee A.V., Duflo E.: *Good Economics for Hard Times*. New York 2019.

Blum L., Blum M.: *A theory of consciousness from a theoretical computer science perspective: Insights from the Conscious Turing Machine*. PNAS 2022 119 (22) s. 1-111. <10.1072/pnas.2115934119>.

Chalmers D.J.: *The conscious mind: In search of a fundamental theory*. Oxford 1996.

Chomsky N.: *Syntactic structures*. Berlin – New York 2002.

Davies M.: *Computability and Unsolvability*. New York 1958.

Ethics Commission. *Automated and Connected Driving*, 2017. <https://web.archive.org/web/20170915110611/http://www.bmvi.de/SharedDocs/EN/publications/report-ethics-commission.html> [dostęp: 27.04.2024)]

Gödel K.: *Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I*. „Monatshefte für Mathematik und Physik" 1931 nr 38 s. 173-198.

Hegel G.F.: *The Science of Logic*. New York 2010.

Kahneman D.: *Thinking, Fast and Slow*. New York 2013.

Kosiński M.: *Jak bardzo intymne dane zna o tobie Google*. <https://www.youtube.com/watch?v=VwzVFcNBRpk> [dostęp: 27.05.2024].

Kosiński M.: *Theory of mind may have spontaneously emerged in large language models*. <https://arxiv.org/abs/2302.02083> [dostęp: 25.04.2024].

Krąpiec M.A.: *Język*. W: *Powszechna encyklopedia filozofii*. T. 5. Red. A. Maryniarczyk [i in.]. Lublin 2004.

Lahav N., Neemeh Z.: *A Relativistic Theory of Consciousness*. „Frontiers in Psychology" 2022 nr 12 (704270) s. 1-25.
Maryniarczyk A.: *Zeszyty z metafizyki*. Nr 1: *Monistyczna i dualistyczna interpretacja rzeczywistości*. Lublin 2006.
McCulloch W., Pitts W.: *A logical calculus of the ideas immanent in nervous activity*. „Bull Mathl Biophys" 1943 nr 5 s. 115-133.
Northoff G.: *From emotions to consciousness – a neuro-phenomenal and neuro-relational approach*. „Frontiers in Psychology" 2012 nr 3 s. 1-14. <https://doi.org/10.3389/fpsyg.2012.00303> [dostęp: 17.04.2024].
Penrose R., Dennett D.C.: *Consciousness involves noncomputable ingredients*. W: *The Third Culture: Beyond the Scientific Revolution*. Ed. J. Brockman. New York 1995.
Rosenblatt F.: *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms*. New York 1961.
Rumelhart D.E., Hinton G.E., Williams R.J.: *Learning representations by back-propagating errors*. „Nature" 1986 nr 323/6088 s. 533-536.
Searle J.: *Minds, brains, and programs*. „The Behvioral and Brain Sciences" 1980 nr 3 s. 417-457.
Smith A.: *An Inquiry into the Nature and Causes of the Wealth of Nations*. Amsterdam – Lausanne – Melbourne – Milan – New York – São Paulo 2007.
Solow R.M.: *A Contribution to the Theory of Economic Growth*. „Quarterly Journal of Economics" 1956 nr 1/70 s. 65-94.
Szopa R.: *Ethical problems in the use of algorithms in data management and in a free market economy*. „AI & Society" 2023 nr 38 s. 2487-2498.
Tomasz z Akwinu: *Summa theologiae.* T. 6: *Człowiek*. Przeł. P. Bełch. Londyn 1980.
Zhong, Z., Liu, Z., Tegmark, M., Andreas, J.: *The clock and the pizza: Two stories in mechanistic explanation of neural networks*. <https://arxiv.org/pdf/2306.17844.pdf> [dostęp: 6.12.2023].

Streszczenie: Filozofia sztucznej inteligencji jest obecnie jedną z wiodących dziedzin w filozofii ze względu na dynamiczny rozwój systemów AI. Od około 2018 r. sztuczna inteligencja rozwija się w kierunku, który skłania do myślenia, czy aby wielkie modele językowe nie stały się już jak człowiek pod względem rozumienia rzeczywistości czy wręcz posiadania świadomości i zdolności odczuwania. Dyskusja na ten temat jest oparta o *implicite* przyjmowane założenia odnośnie do powstania umysłu na drodze emergentyzmu. Nie ma jednak narzędzi naukowych do oceny, czy umysł może wyłonić się z materii, tak samo jak nie ma narzędzi do potwierdzenia istnienia ludzkiej duszy. I tu, i tu jest wiara. Istnienie emergentnego umysłu jest tak samo prawdopodobne, jak istnienie duszy stworzonej przez Boga, z jedną różnicą: istnienie duszy daje nadzieję na zbawienie.

Słowa kluczowe: sztuczna inteligencja, świadomość, założenia, naturalizm, dusza, emergentyzm, transcendencja.