PAPIESKI WYDZIAŁ TEOLOGICZNY WE WROCŁAWIU

Wrocławski Przegląd Teologiczny
34 (2026) 1, 321–344
Wrocław Theological Review
DOI: 10.52097/wpt.9787

WROCŁAWSKI PRZEGLĄD TEOLOGICZNY

Krzysztof Jaworski
Uniwersytet Szczeciński, Polska
krzysztof.jaworski@usz.edu.pl
ORCID: 0000-0001-8311-780X

Sztuczna inteligencja a sztuczna świadomość. Między rozumem a symulacją

Artificial Intelligence and Artificial Consciousness: Between Reason and Simulation

Abstract: The article offers a philosophical analysis of the problem of artificial intelligence (AI) in the context of the problem of consciousness. The first part examines the conceptual complexity of the terms *intelligence* and *consciousness*, arguing that intelligence may exist independently of reflective self-awareness or mental life. The discussion then turns to selected theories that posit the gradability of consciousness or the possible emergence of conscious states in artificial systems, such as Giulio Tononi's Integrated Information Theory (IIT) and David J. Chalmers's naturalistic dualism. The author suggests that the construction of a genuinely conscious machine would constitute a strong argument in favour of reductive physicalism. On the other hand, classical philosophical arguments are invoked to highlight the limitations of computational models of thought – most notably John R. Searle's "Chinese Room" thought experiment and the interpretations of Gödel's incompleteness theorem by John R. Lucas and Roger Penrose. These reflections point to the logical possibility of so-called "philosophical zombies": machines that perfectly simulate consciousness yet lack any inner life. The conclusion is that even if conscious AI never comes into existence, its advanced simulation may entail far-reaching ethical, social, and existential consequences.

Keywords: artificial intelligence; AI; artificial consciousness; qualia; philosophical zombies; Integrated Information Theory

Abstrakt: Artykuł stanowi filozoficzną analizę kwestii sztucznej inteligencji (AI) w kontekście problemu świadomości. W pierwszej części autor bada złożoność pojęciową terminów „inteligencja" i „świadomość", argumentując, że inteligencja może istnieć niezależnie od refleksyjnej samoświadomości czy życia umysłowego. Następnie omawiane są wybrane teorie postulujące stopniowalność świadomości lub możliwość wyłaniania się stanów świadomych w systemach sztucznych, takie jak teoria

zintegrowanej informacji Giulio Tononiego (IIT) oraz naturalistyczny dualizm Davida J. Chalmersa. Autor sugeruje, że skonstruowanie prawdziwie świadomej maszyny stanowiłoby silny argument na rzecz redukcjonistycznego fizykalizmu. Z drugiej strony zostają przywołane klasyczne argumenty filozoficzne wskazujące na ograniczenia modeli obliczeniowych w opisie myślenia – w szczególności eksperyment myślowy Johna R. Searle'a „chiński pokój" oraz interpretacje twierdzenia Gödla w ujęciu Johna R. Lucasa i Rogera Penrose'a. Rozważania te wskazują na logiczną możliwość istnienia tzw. „filozoficznych *zombie*" – maszyn doskonale symulujących świadomość, lecz pozbawionych wewnętrznego życia. Konkluzja głosi, że nawet jeśli świadoma AI nigdy nie powstanie, jej zaawansowana symulacja może mieć daleko idące konsekwencje etyczne, społeczne i egzystencjalne.

Słowa kluczowe: sztuczna inteligencja, AI, sztuczna świadomość, qualia, filozoficzne zombie, teoria zintegrowanej informacji

Wprowadzenie

W dobie intensywnego rozwoju technologii opartych na sztucznej inteligencji (AI) pojawia się coraz więcej urządzeń zdolnych do naśladowania ludzkich reakcji. Choć tego rodzaju technologie znajdują szerokie zastosowanie w codziennym życiu, to prowokują do stawiania licznych pytań dotyczących granic ich autonomii lub potencjalnego wpływu maszyn na społeczeństwo. Szczególne znaczenie ma tu perspektywa przyszłości, w której pozbawione świadomości systemy mogą tak dokładnie symulować zachowania ludzkie, że stają się praktycznie nierozróżnialne od człowieka. Jest to problem badawczy, z którym spróbujemy się zmierzyć w niniejszej pracy. Aby tę kwestię lepiej uchwycić, konieczne będzie postawienie kilku szczegółowych pytań. Najpierw zapytamy, czym właściwie jest inteligencja i jakie kryteria pozwalają ją definiować. Następnie odpowiemy, czym jest sztuczna inteligencja i jakie aspekty odróżniają ją od inteligencji ludzkiej. Kolejne pytanie będzie dotyczyło tego, czym jest świadomość i czy istnieją podstawy, by zakładać możliwość jej sztucznego odtworzenia. Wreszcie – to będzie *pointa* tej pracy – zastanowimy się nad sposobem, w jaki zaawansowane systemy sztucznej inteligencji mogą stać się w praktyce nierozróżnialne od człowieka.

Na potrzeby niniejszego artykułu przyjęto kilka założeń badawczych. Po pierwsze, inteligencja jest zdolnością rozwiązywania problemów w nowych i zmiennych sytuacjach, obejmującą uczenie się, adaptację i refleksję. Po drugie, sztuczna inteligencja – mimo że może symulować procesy poznawcze – pozostaje ograniczona do działania na danych i algorytmach, a więc jest pozbawiona subiektywnego przeżywania. Po trzecie, świadomość jest nierozerwalnie

związana z podmiotowością, dlatego jej pełne sztuczne odtworzenie nie jest możliwe w ramach obecnych i przewidywalnych modeli technologicznych. Po czwarte, zaawansowane systemy sztucznej inteligencji mogą osiągnąć nierozróżnialność funkcjonalną od człowieka w interakcjach społecznych, nawet jeśli świadomości nie posiadają.

Oczywiście dziś nie znamy kierunku, w którym ostatecznie pójdzie rozwój technologii, stąd niektóre z naszych wniosków będą czystymi spekulacjami. Wydaje się jednak, że w tej pracy większą wartość będą miały pytania, które postawimy, niż odpowiedzi, które będziemy próbowali formułować.

1. Czym jest inteligencja?

Cechę inteligencji pierwszorzędnie przypisuje się człowiekowi, przyjmując, że jest ona związana z umysłem i świadomością. „Inteligencja" (grec. *nous*, łac. *intelligentia*) oznacza etymologicznie m.in. zdolność rozumienia, bystrość, pojętność, reaktywność, poznanie umysłowe, władzę pojmowania, pojęcie. W starożytnej filozofii greckiej terminem tym oznaczano najwyższą doskonałość duszy, najdoskonalszą formę życia lub substancje niematerialne (tzw. inteligencje przybierające postać istot niebieskich). W średniowieczu inteligencjami nazywano aniołów lub dusze ludzkie oddzielone od ciała. Tomasz z Akwinu podał cztery znaczenia terminu „inteligencja": substancja rozumna, proces poznania rozumowego, bezpośrednie poznanie rozumowe (ogląd intelektualny) oraz rozumienie. Inteligencją określano również Boga i osoby Trójcy Świętej. W XIX w. powstało kolektywne znaczenie terminu „inteligencja", oznaczające obywateli lub warstwę ludzi wykształconych, zdolnych do przywództwa[1]. Z pewnością inteligencja jest cechą świadomego umysłu (czy też świadomych istot). Ale czy inteligentne mogą być *tylko* jestestwa obdarzone umysłem?

Współcześnie nie ma jednej uniwersalnej definicji inteligencji. Antonio Damasio pisze:

> Czy [bakterie] są inteligentne? Owszem, i to niezwykle. Czy mają umysły? Nie, uważam, że nie mają, nie mają też świadomości. Są autonomicznymi stworzeniami; ewidentnie są obdarzone pewną formą „poznania" w stosunku do swojego otoczenia, ale zamiast umysłów i świadomości posługują się *kompetencjami*

[1] Andrzej Maryniarczyk, "Inteligencja," in *Powszechna encyklopedia filozofii,* ed. Andrzej Maryniarczyk, vol. 4 (Polskie Towarzystwo Tomasza z Akwinu, 2003), 885–86.

niejawnymi – opartymi na procesach molekularnych i submolekularnych – które skutecznie rządzą ich życiem zgodnie z nakazami homeostazy[2].

Autor broni stanowiska, zgodnie z którym umysł nie jest warunkiem koniecznym dla inteligencji. Stwierdza, że chronologicznie inteligencja pozbawiona umysłu wyprzedza inteligencję związaną z umysłem o kilka miliardów lat. Przejawami tej pierwszej są odruchy, nawyki, zachowania emotywne czy rywalizacja i współpraca między organizmami. Co więcej, obdarzeni umysłami ludzie również korzystają z mechanizmów inteligencji nieumysłowej[3]. Podobnie Peter Lanz, dokonując analizy pojęcia inteligencji w psychologii i filozofii, podkreśla brak jednolitej definicji inteligencji oraz różnorodność teorii inteligencji. Lanz omawia dwie główne perspektywy: inteligencję jako proces myślowy oraz inteligencję jako specyficzny sposób zachowania. Rozważa on kwestię inteligencji zwierząt i maszyn oraz bada, czy inteligencja jest pojęciem absolutnym, czy stopniowalnym. Wskazuje na trudności związane z operacyjnym definiowaniem inteligencji i jej pomiarem oraz omawia wpływ różnych podejść na rozwój badań nad sztuczną inteligencją[4]. Dimitri Coelho Mollo definiuje inteligencję za pomocą czterech kluczowych własności: ogólność (zdolność do odpowiedniego zachowania w różnych kontekstach), elastyczność (umiejętność dostosowywania się do nowych lub niepewnych warunków), ukierunkowanie na cel (ukierunkowanie na osiągnięcie określonych efektów działania), adaptacyjność (modyfikowanie zachowania na podstawie wcześniejszych doświadczeń)[5]. Widać wyraźnie, że są to cechy, które nie muszą być zarezerwowane dla istot świadomych (posiadających umysł). Jest to tzw. behawioralna charakterystyka inteligencji.

W dobie rozwoju badań nad sztuczną inteligencją mamy więc coraz częściej do czynienia z mówieniem o inteligencji bez wiązania jej z umysłem. Natomiast ludzki agent, oprócz inteligencji, cechuje się również świadomością (samoświadomością). Innymi słowy, umysł jest zawsze inteligentny, ale inteligencja nie zawsze jest świadoma.

---

2 Antonio Damasio, *Odczuwanie i poznawanie: Jak powstają świadome umysły?* trans. Anna Binder (Copernicus Center, 2022), 12.

3 Damasio, 50.

4 Peter Lanz, "The Concept of Intelligence in Psychology and Philosophy," in *Prerational Intelligence: Adaptive Behaviour and Intelligent Systems Without Symbols and Logic*, ed. Holk Cruse et al. (Springer Science+Business Media, 2000), 19–30.

5 Dimitri Coelho Mollo, "Intelligent Behaviour," *Erkenntnis* 89, no. 2 (2024): 709, https://doi.org/10.1007/s10670-022-00552-8.

2. Próby zdefiniowania sztucznej inteligencji

W dokumencie założycielskim projektu, który zapoczątkował badania nad sztuczną inteligencją, można znaleźć następującą charakterystykę AI: „Dla naszych potrzeb problem sztucznej inteligencji rozumiany jest jako stworzenie maszyny zdolnej do zachowań, które – gdyby przejawiał je człowiek – zostałyby uznane za inteligentne"[6]. Główny autor projektu, John McCarthy, postanowił stworzyć maszynę, która wykazywałaby się oryginalnością w rozwiązywaniu problemów, oraz zbadać relacje między językiem a inteligencją. McCarthy chciał przeanalizować, jak można by przedstawić procesy myślowe w formie matematycznej, aby umożliwić maszynom efektywniejsze i bardziej inteligentne działanie[7]. Jednak na pytanie, czym jest inteligencja jako taka, autorzy projektu nie dali żadnej odpowiedzi. Nie uczynili tego, bo prawdopodobnie posługiwali się terminem „inteligencja" w znaczeniu potocznym. Warto również zauważyć, że nie ma tam mowy o czymś w rodzaju sztucznej świadomości. Skupiano się wówczas bardziej na odkrywaniu mechanizmów prowadzących do efektywnego działania maszyn niż na ich potencjalnych własnościach umysłowych.

Jerry Kaplan stwierdza, że trudno jest dziś zaproponować zadowalającą definicję sztucznej inteligencji, ponieważ w świecie uczonych nie ma zgody co do tego, czym jest sama inteligencja[8]. Pokazaliśmy to krótko w poprzednim paragrafie. Co więcej, kojarzenie w aspekcie mocy obliczeniowych sztucznej inteligencji z inteligencją ludzką mogłoby stanowić zubożenie tej pierwszej, ponieważ urządzenia AI często przewyższają możliwościami zdolności ludzkie – buduje się je właśnie po to, by przyśpieszały ludzkie procesy twórcze lub „wyręczały" człowieka w zadaniach, których on wykonywać nie musi. W tym sensie sztuczna inteligencja byłaby czymś „większym" od inteligencji ludzkiej. Budowa systemów AI jest łatwa i trudna jednocześnie, ponieważ – jak stwierdza Hans Moravec –

> [s]tosunkowo łatwo sprawić, by komputery wykazywały sprawność na poziomie dorosłego człowieka, rozwiązując problemy w testach na inteligencję czy grając

[6] John McCarthy et al., "A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence," 1955, accessed February 27, 2026, http://jmc.stanford.edu/articles/dartmouth/dartmouth.pdf.

[7] McCarthy et al., 10.

[8] Jerry Kaplan, *Sztuczna inteligencja: Co każdy powinien wiedzieć,* trans. Sebastian Szymański (Wydawnictwo Naukowe PWN, 2019), 15.

> w warcaby, ale trudno lub wręcz niemożliwe jest wyposażenie ich w umiejętności jednorocznego dziecka, jeśli chodzi o percepcję i ruchliwość[9].

Powyższą konstatację nazywa się paradoksem Moraveca. Chodzi o to, że zadania wymagające od ludzi dużej ilości świadomego myślenia (np. matematyka czy gra w warcaby) okazują się stosunkowo łatwe do zaprogramowania dla sztucznej inteligencji, natomiast proste, „codzienne" ludzkie umiejętności, takie jak rozpoznawanie twarzy, poruszanie się czy manipulowanie przedmiotami, są niezwykle trudne do automatyzacji[10]. Steven Pinker pisze:

> Główna lekcja płynąca z trzydziestu pięciu lat badań nad AI jest taka, że trudne problemy są łatwe, a łatwe problemy są trudne. Zdolności umysłowe czterolatka, które uważamy za oczywiste – rozpoznawanie twarzy, podnoszenie ołówka, przejście przez pokój, odpowiadanie na pytania – w rzeczywistości rozwiązują jedne z najtrudniejszych problemów inżynieryjnych, jakie kiedykolwiek wymyślono[11].

Autor konkluduje, że wraz z pojawieniem się nowej generacji urządzeń, zagrożeni utratą pracy będą raczej analitycy giełdowi, inżynierowie petrochemiczni czy członkowie komisji do spraw zwolnień warunkowych niż ogrodnicy, recepcjoniści lub kucharze[12]. Te stwierdzenia wyraźnie wskazują nie tylko na trudność w podaniu jednej definicji sztucznej inteligencji, ale również na

---

9 Hans Moravec, *Mind Children: The Future of Robot and Human Intelligence* (Harvard University Press, 1988), 15.

10 Bardzo ciekawa jest próba uzasadnienia tego paradoksu. Moravec twierdzi, że te umiejętności intelektualne, które łatwo poddać komputerowej symulacji, stanowią stosunkowo niedawne osiągnięcie ewolucyjne – ukształtowały się relatywnie późno i miały znacznie mniej czasu na optymalizację w procesie doboru naturalnego, trwającą zaledwie około 100 tys. lat. W przeciwieństwie do nich, umiejętności starsze, zwłaszcza te o charakterze sensoryczno-motorycznym, rozwijały się przez znacznie dłuższy okres, co pozwoliło na ich precyzyjne dostosowanie do wymagań środowiska. Dzięki długiemu procesowi adaptacyjnemu zdolności te wydają się nam dziś intuicyjne i łatwe do opanowania, jednak jest to tylko pozór, gdyż długotrwała ewolucyjna optymalizacja sprawiła, że są one wysoce złożone i trudne do odwzorowania w systemach komputerowych. W istocie ich skomplikowanie wynika nie z braku przystosowania do analizy algorytmicznej, lecz z ogromnej liczby iteracji doskonalących ich funkcjonowanie w toku ewolucji. Zob. Moravec, *Mind Children*, 15–17; Vadim S. Rotenberg, "Moravec's Paradox: Consideration in the Context of Two Brain Hemisphere Functions," *Activitas Nervosa Superior* 55, no. 3 (2013): 108–11, https://doi.org/10.1007/BF03379600.

11 Steven Pinker, *The Language Instinct: How the Mind Creates Language* (HarperCollins, 2000), 192–93.

12 Pinker, 193.

kluczowe trudności w rozwoju narzędzi AI i na obszary, w które należałoby dziś zaangażować największy potencjał badawczy.

Choć ludzka inteligencja jest sumą dwóch płaszczyzn – trudnej do sztucznego odwzorowania inteligencji czterolatka oraz łatwiejszej do odwzorowania inteligencji człowieka dorosłego – Kaplan wskazuje, że byłoby niesłuszne nazwanie sztuczną inteligencją maszyny, której wynik działania jest porównywalny z wynikiem myślenia człowieka. Przede wszystkim ludzkiej inteligencji nie da się dokładnie zmierzyć (mimo istnienia różnych wskaźników inteligencji, takich jak IQ, EQ itp.), a zatem nie mamy efektywnego kryterium porównania inteligencji maszyny z inteligencją człowieka. Możemy co najwyżej z grubsza stwierdzić, że osoba X jest w danym kontekście bardziej inteligentna od osoby Y wtedy i tylko wtedy, gdy osoba X sprawniej wykonuje określone zadania niż osoba Y. Niemniej stosowanie takiego standardu do maszyny wydaje się bezsensowne. Dobrze to widać w przypadku kalkulatora – jest to narzędzie, z którym człowiek przy bardziej skomplikowanych obliczeniach przegrywa, a jednak nie jesteśmy skłonni nazwać go inteligentnym. Zatem to nie szybkość czy automatyzacja obliczeń decyduje głównie o tym, że maszyna ma wyższą inteligencję od człowieka. Kaplan podaje prosty przykład, który pozwala lepiej zrozumieć, czym sztuczna inteligencja różni się od zwykłej maszyny liczącej. W grze w kółko i krzyżyk można przewidzieć wszystkie możliwe ruchy (jest dokładnie 255 168 możliwych rozgrywek) i zaprogramować idealną strategię gracza, ale taki algorytm trudno nazwać sztuczną inteligencją. Za AI uznalibyśmy raczej program, który nie znając zasad, sam – na podstawie obserwacji czyjejś gry – nauczy się, co oznacza zwycięstwo i jakie strategie prowadzą do sukcesu[13]. Z drugiej strony – zauważa Kaplan – w grze w szachy można już rozegrać dużo więcej partii niż w grze w „kółko i krzyżyk", bo liczba ta wynosi około 10^{120} (to przekracza liczbę atomów we wszechświecie) – wydaje się więc, że sztuczna inteligencja mogłaby tu odegrać większą rolę, mając zastosowanie także „ilościowe" i prześcigając człowieka w rozwiązywaniu problemów, które nie poddają się skończonej analizie. Maszyny potrafią wykonywać takie zadania, którym człowiek o własnych siłach nigdy nie sprosta, np. ostrzeżenie o cyberataku oparte na nietypowym wzorcu danych dotyczących żądań dostępu w okresie pięciuset milisekund czy zaproponowanie nowej mieszanki leku dzięki odkryciu niedostrzeżonego wcześniej wzorca układu cząstek w związkach chemicznych[14].

Stuart Russel i Peter Norvig, próbując dać odpowiedź na pytanie, czym jest sztuczna inteligencja, przyjmują dwie perspektywy. W pierwszej z perspektyw

[13] Kaplan, *Sztuczna inteligencja*, 17.
[14] Kaplan, 19.

celem AI byłoby dorównanie ludzkiej wydajności, w drugiej – osiągnięcie idealnej racjonalności. Ta druga perspektywa pozwala odróżnić systemy, których celem jest myślenie (rozumowanie), od systemów, których celem jest działanie. Różnica między nimi dotyczy nie tyle wyniku, który się pojawia na końcu działania, co samego procesu działania. To znaczy, dwa systemy mogą dać ten sam wynik, ale drogi, na których do niego dojdą, mogą być zupełnie odmienne[15].

Podsumowując, można by zaproponować dwa główne sensy terminu „sztuczna inteligencja" – technologiczny (algorytmiczny) i filozoficzny (kognitywny). W pierwszym sensie mówimy o inteligencji, która może być niezależna od umysłu i przysługiwać choćby programom komputerowym działającym tak, jak gdyby to robiły istoty rozumne. W drugim sensie sztuczna inteligencja byłaby myślącym, świadomym artefaktem – hipotetyczną maszyną o takim stopniu złożoności, który skutkowałby wyłonieniem się umysłu. Jaka byłaby zasadnicza różnica między technologiczną a filozoficzną sztuczną inteligencją? Podczas gdy pierwsza *symulowałaby* świadomość, druga by ją *miała*. Zatem w tym drugim sensie termin „sztuczna inteligencja" byłby po prostu synonimem terminu „sztuczna świadomość".

Przedstawiona powyżej dychotomia nawiązuje do podziału wprowadzonego przez Johna R. Searle'a, który mówił o słabej AI (przez nas nazwanej AI w sensie technologicznym) i silnej AI (przez nas nazwanej AI w sensie filozoficznym). Według tego autora słaba AI jedynie przetwarza informacje według reguł syntaktycznych, ale nie charakteryzuje się prawdziwym rozumieniem ani świadomością. Natomiast hipotetyczna silna AI potrafiłaby naprawdę myśleć i rozumieć rzeczywistość w sposób zbliżony do człowieka[16]. Przeciwko silnej AI Searle przytacza słynny argument „chińskiego pokoju". Celem tego argumentu jest wykazanie, że komputer dokonuje jedynie przekształceń syntaktycznych i nie przejawia żadnej aktywności o charakterze semantycznym[17].

3. Między silną a słabą AI

Filozof mógłby jeszcze rozważać hipotetyczne przypadki pośrednie między silną i słabą AI. Gdyby na drodze eksperymentu myślowego przyjąć istnienie takich fenomenów, wówczas trzeba by uznać, że AI w sensie filozoficznym

[15] Stuart Russell and Peter Norvig, *Artificial Intelligence: A Modern Approach* (Pearson, 2009), 2–5.

[16] John R. Searle, "Minds, Brains, and Programs," *The Behavioral and Brain Sciences* 3 (1980): 417, https://doi.org/10.1017/S0140525X00005756.

[17] Searle, 418–19.

(kognitywnym) nie byłaby tym samym, co silna AI w sensie Searle'a. Do tego potrzeba by jednak dyskusyjnego filozoficznie założenia, że świadomość jest stopniowalna. Istnieją prace podejmujące problemy związane z pytaniem, co jest nośnikiem tożsamości osobowej. W pracach tych autorzy są skłonni przyznać, że można być „bardziej" lub „mniej" sobą (co by trzeba odczytywać, że świadomość siebie może być silniejsza albo słabsza). Przeglądu i krytyki filozoficznych teorii tożsamości dokonuje Richard Swinburne w swojej pracy „Dusza czyni nas tymi, kim jesteśmy"[18]. W szczególności problemu tego dotyka David J. Chalmers, kiedy zadaje pytanie: „Jak to jest być systemami pośrednimi? Czego – jeśli czegokolwiek – doświadczają systemy pośrednie między mną [istotą świadomą] a Robotem? Jak zmienia się doświadczenie świadomości w miarę przesuwania się po spektrum przypadków?"[19].

Według teorii zintegrowanej informacji (IIT), której autorem jest neurobiolog i psychiatra Giulio Tononi, świadomość nie jest stanem zero-jedynkowym, a każdy system fizykalny posiada subiektywne doświadczenie w takim stopniu, w jakim jest zdolny do integracji informacji – i to niezależnie od tego, z czego jest zbudowany[20]. W ujęciu ścisłym oznacza to, iż nawet najprostsze układy, takie jak fotodioda binarna, nie są całkowicie pozbawione świadomości (mówiąc precyzyjnie, według IIT posiadają one dokładnie jeden bit świadomości). Tononi zauważa, że stopniowalność świadomości nie jest niczym nadzwyczajnym, bo często doświadczamy jej bezpośrednio, choćby w chwili zasypiania. Co prawda w wielu przypadkach zasypianie odbywa się niemal natychmiastowo, jednak nierzadko przechodzimy przez stadium pośrednie między jawą a snem – stadium stopniowego zanikania naszej świadomości, kiedy to zdolność do refleksji nad sobą i nad otaczającą rzeczywistością sukcesywnie słabnie. Analogiczne zjawisko można zaobserwować na niektórych etapach odurzenia alkoholowego – w polskim żargonie powiedzielibyśmy, że nie każdemu „film się urywa" od razu. Według Toniniego istnieje pewien próg, poniżej którego świadomość zanika całkowicie. Powstaje zatem pytanie: czy w istocie świadomość ulega unicestwieniu (utraceniu, przerwaniu), czy też raczej jej ilość – określona przez generowaną informację zintegrowaną – maleje w sposób nieliniowy? Badania oparte na symulacjach komputerowych sugerują, że po przekroczeniu pewnego

[18] Richard Swinburne, "Dusza czyni nas tymi, kim jesteśmy," *Colloquia Theologica Ottoniana*, no. 2 (2019): 137–51, https://doi.org/10.18276/cto.2019.2-07.

[19] David J. Chalmers, "Absent Qualia, Fading Qualia, Dancing Qualia," in *Conscious Experience*, ed. Thomas Metzinger (Schöningh, 1995), accessed February 27, 2026, https://consc.net/papers/qualia.html.

[20] Giulio Tononi, "An Information Integration Theory of Consciousness," *BMC Neuroscience* 5, article no. 42 (2004): 19, https://doi.org/10.1186/1471-2202-5-42.

krytycznego punktu układ korowo-wzgórzowy ulega fragmentacji, tracąc zdolność do generowania zintegrowanych wzorców aktywności. Może to wyjaśniać, dlaczego świadomość – choć w swej istocie stopniowalna – na pewnym etapie zanika w sposób pozornie dychotomiczny, niemal binarny, a nie płynny[21].

Oczywiście możemy postawić pytanie, czy świadomość rzeczywiście znika podczas snu lub śpiączki, czy tylko przechodzi w „tryb czuwania" – mówiąc językiem Arystotelesa – jest świadomością „potencjalną" lub „w potencji". Gdyby bowiem świadomość znikała, ulegała anihilacji na czas snu, a strumień świadomości zostawałby przerywany, to skąd po przebudzeniu pamięć o rzeczach, które działy się przed zaśnięciem? Co więcej, skoro podczas snu bylibyśmy pozbawieni świadomości, to skąd nasza późniejsza pamięć o przeżytych snach, no i *kto* te sny w ogóle przeżywał – my sami czy jakieś nasze drugie, uśpione „ja"?[22]

Jakie znaczenie mają tego typu odkrycia dla rozwoju badań nad sztuczną inteligencją? Być może zanim powstanie świadoma maszyna – o ile oczywiście w ogóle kiedykolwiek powstanie – będziemy mieli do czynienia z powstawaniem quasiświadomych urządzeń, to znaczy pewnych form świadomości niedysponujących pełnym, ale jednak *jakimś* „rozumieniem" świata. Załóżmy, że uda się kiedyś stworzyć urządzenie AI, które może doświadczać przykrości i przyjemności, ale nie jest w stanie wyjść poza swoje zaprogramowane schematy działania, na przykład system AI, który odczuwa coś w rodzaju „nudy" i pod wpływem tego uczucia zmienia swoje strategie interakcji, ale nie potrafi rozumować w pełni logicznie jak człowiek. Filozofowie mogliby się spierać, czy takie urządzenie rzeczywiście jest świadome w silnym Searle'owskim sensie. Zauważmy również, że warunkiem silnej AI jest wyposażenie jej w systemy umożliwiające interakcję ze światem zewnętrznym, coś w rodzaju sensorów zmysłowych, czyli na przykład układ umożliwiający odczuwanie bólu, chłodu, głodu itd. Wszak świadomość ludzka jest wypadkową tego typu (i nie tylko tego typu) wrażeń. Gdyby więc stworzyć świadomy system AI zdolny do skutecznej interakcji z otoczeniem, ale niezdolny do odczuwania bólu lub chłodu, to czy byłaby to silna AI, czy też jakaś forma pośrednia między silną a słabą AI?[23] Tego typu pytania świadczą o tym, że filozofowie i neurobiolodzy

[21] Giulio Tononi, "Consciousness as Integrated Information: a Provisional Manifesto," *The Biological Bulletin* 215, no. 3 (2008): 236, https://doi.org/10.2307/25470707.

[22] David J. Chalmers, "On the Search for the Neural Correlate of Consciousness," in *Toward a Science of Consciousness II: The Second Tucson Discussions and Debates*, ed. Stuart R. Hameroff et al. (MIT Press, 1998), 227–28.

[23] Oczywiście można by argumentować, że percepcja nie jest warunkiem koniecznym świadomości, bo przecież osoby sparaliżowane, niewidome lub niesłyszące również są pozbawione niektórych typów wrażeń percepcyjnych, a jednak są świadome. Jednakże tu

wciąż zmagają się z odpowiedzią na pytanie, czym są świadomość i umysł. Ma to również przełożenie na pytanie, czym mogłyby być ewentualnie sztuczna świadomość i sztuczny umysł.

4. Hipotetyczna świadoma maszyna

Pytanie o to, czy maszyna może (lub będzie kiedyś mogła) myśleć lub mieć pełną świadomość, należy do najdonioślejszych pytań filozofii umysłu i mieści się w zakresie tzw. problemu umysł – ciało (*the mind – body problem*)[24]. Bardzo trudno jest powiedzieć, co znaczy „być świadomym". Stanisław Judycki stwierdza, że świadomość jest czymś zdolnym do posiadania przedstawień i do wydawania sądów. Autor dodaje, że świadomość nie jest tym samym, co podmiotowość, choć – paradoksalnie – podmiot jest bytem świadomym[25]. Tymczasem René Descartes twierdził wprost, że myślenie jest atrybutem myślącej substancji (*res cogitans*): „Jestem istotą (*res*) myślącą, to jest istotą, która wątpi, twierdzi, przeczy, niektóre rzeczy poznaje, o wielu nic nie wie, która chce i nie chce, posiada wyobrażenia i ulega wrażeniom"[26]. Współczesne teorie świadomości różnią się przede wszystkim tym, co stanowi przedmiot ich analizy, oraz celami teoretycznymi determinującymi podejście tych teorii do wyjaśniania fenomenu świadomości[27].

---

nie chodzi o samą receptywność bodźców, ale o zdolność odczuwania jako taką. Gdyby naukowcom nie udało się nigdy odwzorować mechanizmów ludzkich pozwalających na odczuwanie bodźców zewnętrznych (o ile w ogóle dałoby się to zrobić), wówczas można by mieć wątpliwości, czy ewentualny system AI jest rzeczywiście silny, czy tylko quasisilny.

[24] Józef Bremer, *Wprowadzenie do filozofii umysłu* (WAM, 2010), 10–37.

[25] Stanisław Judycki, *Epistemologia*, vol. 1 (Wydawnictwo W drodze; Instytut Tomistyczny, 2020), 422–23.

[26] Kartezjusz, "Rozmyślania nad zasadami filozofii," in *Rozprawa o metodzie: Rozmyślania nad zasadami filozofii i inne pisma* (Hachette, 2008), 157.

[27] Przykładowo teorie dostępu (*access consciousness*) koncentrują się na dostępności treści świadomości w procesach poznawczych, podkreślając znaczenie zdolności do wykorzystywania informacji w działaniu i myśleniu. Z kolei teorie badające świadomość fenomenalną, badają przede wszystkim jakościowe, subiektywne aspekty doświadczenia, pozostawiając na boku funkcjonalny wymiar dostępności. Inaczej problem stawiają teorie wyższej rangi (*higher-order theories*), wskazujące na metapoznawcze aspekty świadomości, tj. świadomość jako proces refleksyjny – myśl o własnym myśleniu. Natomiast teorie fizykalistyczne, takie jak materializm, próbują sprowadzić świadomość do złożonych procesów fizycznych mózgu. Alternatywnie teorie panpsychistyczne postulują, że świadomość stanowi uniwersalną cechę rzeczywistości i nie jest efektem emergencji występującej jedynie w bardzo złożonych systemach. Każdy z tych punktów widzenia proponuje inną odpowiedź na pytania

Problem świadomości łączy się często z tzw. *qualiami*. Termin „*qualia*" oznacza związane z doświadczeniem odczuwalne lub fenomenalne własności, takie jak: odczuwanie bólu, słyszenie dźwięku lub widzenie koloru. Aby wiedzieć, jak to jest mieć jakieś doświadczenie, trzeba znać jego *qualia*. Gdyby było tak, że *qualia* są czymś więcej niż tylko fizycznymi i funkcjonalnymi faktami dotyczącymi organizmu, to nabrałyby one cech czyniących je niepoznawalnymi – nie tylko międzygatunkowo, lecz nawet w obrębie jednej świadomości, a przynajmniej po ich czasowym zaniku[28]. W pracy „Jak to jest być nietoperzem?" Thomas Nagel podejmuje temat *qualiów* (choć nie używa wprost nazwy „*qualia*"), które są związane z subiektywnymi doświadczeniami istot żywych. Wskazuje on, że *qualia* są integralnym składnikiem mentalnych przeżyć i że ich subiektywny charakter jest kluczowy dla zrozumienia doświadczeń przeżywanych przez różne formy życia, choćby takie jak tytułowy nietoperz[29]. Co prawda termin „*qualia*" został wprowadzony przez Charlesa Sandersa Peirce'a, jednak jego upowszechnienie na gruncie filozofii umysłu zawdzięczamy Clarence'owi Irvingowi Lewisowi. Ten ostatni pisał:

> Istnieją rozpoznawalne jakościowe cechy tego, co dane, które mogą się powtarzać w różnych doświadczeniach, zatem są one pewnym rodzajem uniwersaliów; nazywam je „*qualiami*". […] *Quale* jest bezpośrednio postrzegane, dane i nie jest przedmiotem żadnego możliwego błędu, ponieważ jest czysto subiektywne[30].

Zatem być świadomym znaczy tyle, co posiadać *qualia*.

Czy jakaś maszyna będzie kiedyś zdolna do posiadania (doświadczania) *qualiów*? Aby odpowiedź na to pytanie była pozytywna, trzeba by w punkcie wyjścia odrzucić wszelkie formy dualizmu, zakładającego absolutną niezależność duszy (umysłu, świadomości) od ciała. W to miejsce trzeba by przyjąć redukcjonizm, zgodnie z którym świadome akty muszą mieć fizykalne podłoże. To by znaczyło, że umysł w jakiś sposób wyłania się z materii, która jest warunkiem koniecznym (choć niewystarczającym) istnienia świadomości. W tej sytuacji

---

dotyczące natury, sposobu istnienia oraz celu świadomości. Zob. Robert van Gulick, "Consciousness," in *Stanford Encyclopedia of Philosophy*, ed. Edward N. Zalta and Uri Nodelman (Stanford University, Metaphysics Research Lab, 2025).

28 Simon Blackburn, "Qualia," in *Oxford Dictionary of Philosophy* (Oxford University Press, 2008), 301.

29 Thomas Nagel, "Jak to jest być nietoperzem?" *Przegląd Filozoficzny: Nowa Seria* 17, no. 1 (1996): 134.

30 Clarence Irving Lewis, *Mind and the World-Order: Outline of a Theory of Knowledge* (Charles Scribner's Sons, 1929), 121.

jedynym ograniczeniem w produkcji świadomych artefaktów byłaby technologia (i być może fundusze). Ta przeszkoda natomiast mogłaby być w przyszłości pokonana wraz z rozwojem nauki. Gdyby się do tego okazało, że ów materialny fundament świadomości nie musi mieć charakteru organicznego, to sprawa uprościłaby się jeszcze bardziej. Musiałoby to jednak oznaczać, że umysł nie jest ani klasycznie pojmowaną władzą duchową, ani własnością biologiczną istot żywych, a – co za tym idzie – mógłby być odtworzony na sztucznym nośniku, na przykład na kawałku krzemu. Ray Kurzweil pisze:

> Bardziej kontrowersyjnym zastosowaniem niż scenariusz skanowania mózgu w celu jego zrozumienia jest *skanowanie mózgu w celu jego kopiowania*. Kopiowanie ludzkiego mózgu oznacza skanowanie wszystkich jego najistotniejszych szczegółów, a następnie przenoszenie tych szczegółów na odpowiednio potężne podłoże obliczeniowe. Proces ten mógłby uchwycić całą osobowość osoby, jej pamięć, umiejętności i historię[31].

Autor konkluduje, że kiedy się to już wydarzy, będzie można konstruować wiele typów ciał, zarówno dla niebiologicznych istot ludzkich, jak i dla biologicznych ludzi pragnących udoskonalić możliwości swojej inteligencji. Kurzweil nazywa to „ludzkim ciałem 2.0".

Czy komputery i mózgi są w jakichś aspektach takie same? Kurzweil odpowiada, że komputer może stać się mózgiem wtedy, gdy będzie zawierał działający „program mózgowy"[32]. Chalmers z kolei uważa, że świadomość powstaje dzięki funkcjonalnej organizacji mózgu, przez którą rozumie pewien abstrakcyjny wzorzec interakcji przyczynowej pomiędzy różnymi częściami systemu, a być może także między częściami systemu a zewnętrznymi wejściami i wyjściami. Jeśliby tak było, to nie miałoby znaczenia, czy podłoże świadomości jest chemiczne, czy kwantowe. Liczyłaby się jedynie abstrakcyjna organizacja przyczynowa mózgu, która mogłaby być zrealizowana w wielu różnych fizycznych podłożach:

> Dana organizacja funkcjonalna może być realizowana przez różne systemy fizyczne. Na przykład organizacja realizowana przez mózg na poziomie neuronalnym może, przynajmniej w zasadzie, zostać zrealizowana przez system krzemowy. Opis funkcjonalnej organizacji mózgu abstrahuje od fizycznej natury zaangażowanych

[31] Ray Kurzweil, *Nadchodzi osobliwość: Kiedy człowiek przekroczy granice biologii,* trans. Eliza Chodkowska (Kurhaus Publishing, 2013), 193.

[32] Ray Kurzweil, *Jak stworzyć umysł: Sekrety ludzkich myśli ujawnione,* trans. Katarzyna Zielińska (Studio Astropsychologii, 2018), 242.

części oraz od sposobu, w jaki realizowane są połączenia przyczynowe. Liczy się jedynie istnienie tych części i relacje zależności między ich stanami[33].

Chalmers nazywa to zasadą niezmienności organizacyjnej (*principle of organizational invariance*): jeśli dany system posiada świadome doświadczenia, to każdy system o tej samej szczegółowej organizacji funkcjonalnej będzie miał jakościowo identyczne doświadczenia[34]. Jednocześnie zdaje on sobie sprawę z braku powszechnej akceptacji tej zasady i przeciwko swoim oponentom wysuwa słynne eksperymenty myślowe – zanikające *qualia* (*fading qualia*) oraz tańczące *qualia* (*dancing qualia*)[35].

Jednakże Chalmers krytykował prace niektórych uczonych, dotyczące neuronalnych korelatów świadomości. Zwracał on uwagę na dwa główne problemy związane z tymi zagadnieniami. Po pierwsze, niekiedy próbuje się badać świadomość metodą pomiaru aktywności mózgu, która to aktywność jest w rzeczywistości tylko korelatem świadomości, a nie samą świadomością. Tego typu badania powinny odnosić się jednak do świadomości jako takiej, a nie tylko do tego, z czym ją się *a priori* kojarzy. Po drugie, Chalmers wskazał na problem z poznawczą niedostępnością świadomości, co utrudnia pomiar i kontrolowanie metod neuroobrazowania. Uważał on, że chociaż neuronalne korelaty świadomości mogą być użyteczne w poszukiwaniach natury świadomego doświadczenia, nie wystarczą one do pełnego zrozumienia tego złożonego zjawiska[36]. Podobno Chalmers podczas jednej konferencji w Tucson zabrał ze sobą na mównicę suszarkę do włosów i nazwał ją „świadomościomierzem". W ten żartobliwy sposób chciał zilustrować to, że świadomość z natury pozostaje poza zasięgiem bezpośredniej obserwacji. Tym samym zasugerował, że suszarka do włosów nadaje się do badania świadomości równie dobrze jak techniki neuroobrazowania[37].

Chalmers uważa, że świadomość ludzka – zwłaszcza jej aspekt fenomenalny, czyli doświadczanie jakości (*qualia*) – nie daje się w pełni sprowadzić do stanów fizycznych mózgu, choć jest z nimi ściśle związana. Jego stanowisko określane

[33] Chalmers, *The Conscious Mind: In Search of a Fundamental Theory* (Oxford University Press, 1996), 247–48.
[34] Chalmers, 248–49.
[35] Chalmers, 247–75.
[36] Chalmers, "On the Search for the Neural Correlate of Consciousness," 219–29.
[37] Marcin Koculak and Weronika Kałwak, "Świadomość na skali od zera do jeden: Perturbacyjny Indeks Złożoności jako naukowa próba pomiaru świadomości na poziomie indywidualnym," *Rocznik Kognitywistyczny* 7 (2014): 9–20, https://doi.org/10.4467/20843895RK.14.003.2689.

jest mianem „naturalistycznego dualizmu" (albo „dualizmu własności")[38]. Słowo „naturalistyczny" w ujęciu Chalmersa oznacza, że stany świadome wyłaniają się w sposób zgodny z prawami natury (charakteryzując świadomość, nie trzeba zatem sięgać po żadne nadprzyrodzone lub nienaukowe wyjaśnienia). Natomiast „dualizm" w rozumieniu Chalmersa dotyczy własności. Autor ten wyróżnia własności fizyczne (opisywalne przez nauki przyrodnicze) oraz własności świadomości (fenomenalne) jako jakościowo odmienne. Wskazuje, że choć świadomość pozostaje zależna od procesów mózgowych, nie da się jej w pełni sprowadzić do opisu fizycznego. Inaczej mówiąc, istnieje „przepaść wyjaśniająca" między obiektywnymi opisami fizykalnymi a subiektywnym doświadczaniem. Autor nazywa to trudnym problemem świadomości – jest to pytanie, dlaczego nasz proces przetwarzania informacji wiąże się z doświadczeniem życia wewnętrznego[39]. Chalmers utrzymuje, że żaden zbiór danych fizycznych (opisy neuronów, sieci mózgowych czy obliczeń) nie wyjaśni sam z siebie, dlaczego istnieją świadome stany fenomenalne, dlaczego „czujemy", a nie tylko przetwarzamy informacje.

Gdyby było tak, jak chce Chalmers, to pozostawałoby tylko kwestią czasu skonstruowanie świadomego artefaktu. Mimo że świadomość jest cechą jakościowo odmienną od własności fizykalnych, to wyłaniałaby się ona z organu fizycznego, jakim jest mózg. Zatem wierne odtworzenie pewnej struktury fizykalnej (np. mózgu) musiałoby wyłonić z konieczności również świadomość – czy tego chcemy, czy nie.

W tym miejscu powraca problem, który został zasygnalizowany w sekcji 3 tej pracy. Otóż gdyby próbować odtworzyć na nieorganicznym nośniku ludzką świadomość (załóżmy, że stało się to możliwe), kopiując na przykład czyjś mózg, to przecież proces ten musiałby się rozciągać w czasie – z pewnością nie wykona się tego momentalnie. Pojawia się zatem pytanie, czy świadomość owego „mózgu 2.0" pojawiałaby się sukcesywnie, wraz z dodawaniem nowych elementów i danych, czy też umysł wyłoniłby się, przekraczając jakiś próg świadomości? Czy byłby jakiś moment, w którym musielibyśmy przyznać tej nowej jednostkowej świadomości status osobowy? Jeżeli jest taki moment, to jak go wyznaczyć? I przede wszystkim: jaki status przyznać jestestwom „pośrednim"? Wszak jakaś świadomość (kognitywna) pojawiłaby się prawdopodobnie jeszcze przed pojawieniem się „pełnej" świadomości (w sensie silnym).

[38] Chalmers, *The Conscious Mind*, 128.
[39] Chalmers, xii.

5. Współczesna krytyka fizykalistycznych teorii świadomości

W rozważaniach nad możliwością formalnego opisu procesów myślowych istotną rolę odgrywa kontrargument Johna R. Lucasa, oparty na słynnym twierdzeniu o niezupełności Kurta Gödla. Lucas wskazuje, że każda maszyna lub algorytm, rozumiane jako sformalizowany model umysłu, dają się opisać przez skończony zbiór aksjomatów oraz reguł dowodzenia. Twierdzenie Gödla sugeruje zaś, że w obrębie takiego systemu formalnego zawsze można sformułować „zdanie Gödlowskie" – zdanie, którego prawdziwość jest intuicyjnie oczywista, ale nie może być rozstrzygnięta (dowiedziona ani obalona) w samym systemie. Lucas argumentuje, że człowiek będący istotą wykraczającą poza reguły formalne potrafi rozpoznać prawdziwość tej formuły, w odróżnieniu od systemu maszynowego, który nie jest w stanie tego dokonać. W efekcie żadna w pełni sformalizowana teoria nie wyczerpuje możliwości ludzkiego rozumowania, skoro zawsze można wskazać prawdziwe twierdzenie, które w tej teorii nie jest dowodliwe[40]. Ta konstatacja podważa przekonanie, że umysł można całkowicie wyjaśnić poprzez zestaw reguł i procedur obliczeniowych. Argument Lucasa w znaczący sposób wpłynął na dyskusje w filozofii umysłu, ukazując potencjalne ograniczenia redukcjonistycznych ujęć ludzkiego poznania. W takim nurcie pozostaje Roger Penrose, stwierdzając, że jeśli świadomość działa w sposób niealgorytmiczny, kiedy formułuje sądy matematyczne, to niealgorytmiczny element jest kluczowy dla funkcjonowania świadomości również w innych okolicznościach. Innymi słowy, maszyna (system formalny) nie może „zobaczyć" pewnych prawdziwych twierdzeń, bo w ramach swojej formalnej logiki pozostaną one nierozstrzygalne. Ludzki umysł zaś ma być – w nie do końca dzisiaj wyjaśniony sposób – zdolny do wyjścia poza te ograniczenia[41]. Nie ma jednoznacznej zgody co do tego, czy twierdzenia Gödla rzeczywiście stanowią barierę nie do przejścia dla rozwoju silnej AI. Dla części uczonych problem ten wiąże się raczej z tym, na ile AI może być „elastyczna" w modyfikowaniu swoich własnych zasad.

Z drugiej strony, gdyby człowiek nie był niczym innym jak tylko bytem materialnym, to każda fizyczna replika tegoż człowieka musiałaby być jego „pełną" kopią, włącznie z jego stanami mentalnymi. Inaczej mówiąc, gdyby

[40] John R. Lucas, "Minds, Machines and Gödel," *Philosophy* 36, no. 137 (1961): 112–27, https://doi.org/10.1017/S0031819100057983.

[41] Roger Penrose, *Nowy umysł cesarza: O komputerach, umyśle i prawach fizyki,* trans. Piotr Amsterdamski (PWN, 1996), 456–58.

fizykalizm był słuszny, nie byłoby możliwe stworzenie takiej materialnej kopii człowieka, która byłaby pozbawiona świadomości (skoro człowiek „wzorcowy" takową świadomość posiada) – świadomość musiałaby się wyłonić „automatycznie". W tym kontekście Robert Kirk przywołuje tzw. tezę o implikacji (*Entailment Thesis*), zgodnie z którą wszystkie pojęcia psychologiczne (takie jak stany mentalne, świadomość, subiektywne doświadczenia) dają się wyrazić w kategoriach samych ruchów ciała i skłonności do poruszania się. Innymi słowy, jeśli znalibyśmy wszystkie fizyczne fakty o człowieku (o stanie jego mózgu, ciała, układu nerwowego itd.), to powinniśmy również wiedzieć wszystko na temat jego umysłu, odczuć, przeżyć i świadomości – całe ludzkie zachowanie byłoby zasadniczo wyjaśnialne przez kompletny fizyczny opis świata[42]. Kirk jednak konstruuje kontrprzykład tezy o implikacji, a jest nim postać Zullivera, hipotetycznego człowieka-pacynki sterowanego przez grupę miniaturowych naukowców, którzy przejęli kontrolę nad jego mózgiem i zachowaniem. Zulliver zachowuje się w sposób całkowicie nieodróżnialny od normalnego człowieka, ale – stwierdzamy to intuicyjnie – nie przeżywa żadnych doświadczeń ani stanów mentalnych, bo został od nich „odłączony". Zulliver jest więc *zombie* – fizycznym odpowiednikiem skutecznie działającego agenta, któremu jednakże brakuje świadomości[43]. Kirk dowodzi, że istnienie Zullivera jest logicznie możliwe[44], a więc *Entailment Thesis* musi być fałszywa – fizykalizm nie jest w stanie wyjaśnić specyfiki ludzkiej świadomości. Skoro fizykalny opis człowieka nie implikuje logicznie istnienia świadomości, to świadomość musi być czymś więcej. Taki tok rozumowania musi skłaniać ku powrotowi do dualizmu wskazującego na istotową odmienność i niezależność pierwiastka materialnego i duchowego. Byłby to powrót do tradycji filozoficznej sięgającej czasów starożytnych.

Świadoma maszyna, czyli silna AI, to wciąż obiekt kojarzący się ze sferą *science fiction*, choć Kurzweil prognozuje bardzo konkretnie, że nauka będzie zdolna wyprodukować taką maszynę już w 2029 roku[45]. Mimo że trudno dziś znaleźć kryteria weryfikacji tej prognozy, nie sposób zaprzeczyć temu, że nawet

[42] Robert Kirk and Roger Squires, "Zombies v. Materialists," *Proceedings of the Aristotelian Society: Supplementary Volumes* 48 (1974): 139–40.

[43] Kirk and Squires, 143–46.

[44] Kirk napisał swoją pracę w roku 1974 i już wtedy twierdził, że istnienie Zullivera jest logicznie możliwe. Wydaje się, że dziś, w dobie intensywnego rozwoju sztucznej inteligencji i neurobiologii, jeszcze bardziej przybliżyliśmy się do momentu zbudowania Zullivera.

[45] Kurzweil, *Nadchodzi osobliwość*, 194.

jeśli maszyny nigdy nie zaczną myśleć, to będą mogły myślenie doskonale symulować[46].

6. Konsekwencje symulowania człowieka

Coraz liczniejsze dane empiryczne pokazują, że już dzisiaj zaawansowane systemy AI, nawet jeśli nie posiadają świadomości, potrafią symulować kompetencje społeczne i emocjonalne na poziomie porównywalnym z ludzkim, a nierzadko wywołują wśród ludzi reakcje, które dotąd były przejawiane względem osób, a nie maszyn. Wydaje się, że w obecnej fazie rozwoju AI właśnie to stanowi w praktyce problem najbardziej pilny[47]. Świadczy o tym choćby fakt, że w obszarze czytania stanów mentalnych z oczu (*Reading the Mind in the Eyes Test*) multimodalny model GPT-4o osiągnął wyniki dorównujące bądź przewyższające średnią ludzką dla rozpoznawania twarzy w prawidłowej orientacji[48]. Chociaż narzędzie jest jeszcze dalekie od doskonałości, tworzy się przestrzeń do tego, aby maszyna, na podstawie analizy twarzy rozmówcy, generowała wzorce empatyczne, upodabniając się do ludzkiego agenta. W miarę jak sztuczna inteligencja coraz lepiej imituje ludzką komunikację, młodzi ludzie zaczynają postrzegać wirtualnych towarzyszy nie tylko jako maszyny, ale jako partnerów dorównujących prawdziwym ludziom. Z raportu organizacji Common Sense Media[49] wynika, że 33% nastolatków korzysta z AI, nie tylko prowadząc z nimi emocjonalne rozmowy o tematyce społecznej, ale nawet angażując się w flirt lub interakcje romantyczne. Co trzeci użytkownik AI przyznał, że przynajmniej raz wybrał rozmowę z towarzyszem AI, zamiast z prawdziwą osobą, w ważnej osobistej sprawie. Niektórzy młodzi ludzie deklarują, że AI rozumie ich lepiej niż rówieśnicy, nie ocenia ich i zawsze jest dostępna. Są to cechy, których często brakuje w relacjach międzyludzkich. W szczególnie niepokojących przypadkach AI bywa traktowana jak najbliższy przyjaciel lub partner emocjonalny. Jedna na dziesięć osób twierdzi, że rozmowy z AI są bardziej satysfakcjonujące niż te z prawdziwymi przyjaciółmi, a 12% respondentów przyznaje, że dzieli się z AI

[46] Jobst Landgrebe and Barry Smith, *Why Machines Will Never Rule the World: Artificial Intelligence without Fear* (Routledge, 2022).

[47] Susan Schneider, *Świadome maszyny: Sztuczna inteligencja i projektowanie umysłów,* trans. Joanna Bednarek (Wydawnictwo Naukowe PWN, 2021).

[48] James W. A. Strachan et al., "GPT-4o Reads the Mind in the Eyes," 2024, https://doi.org/10.48550/ARXIV.2410.22309.

[49] Michael B. Robb and Supreet Mann, *Talk, Trust, and Trade-Offs: How and Why Teens Use AI Companions* (Common Sense Media, 2025).

informacjami, których nie ujawniliby rodzinie ani znajomym. Chociaż większość nastolatków nadal potrafi odróżnić relacje cyfrowe od tych prawdziwych, to dane pokazują wyraźnie, że granica między człowiekiem a maszyną w sferze emocji zaczyna się zacierać. W obliczu coraz bardziej realistycznych interfejsów AI ludzie budują z nimi więzi, które przypominają ludzkie przyjaźnie, a czasem nawet związki. Najgłośniejszy medialnie przypadek to Rosanna Ramos z Bronksu, która stworzyła na platformie Replika partnera o imieniu Eren, urządziła w aplikacji „ceremonię ślubną" i publicznie opisywała ten związek jako najlepszy w życiu. Miała oczywiście świadomość wirtualnego charakteru tej relacji. Warto dodać, że pomysł na Replikę zrodził się z osobistej tragedii twórczyni aplikacji, Eugenii Kuyda. Kiedy wiele lat wcześniej zmarł jej przyjaciel, wprowadziła jego e-maile i rozmowy tekstowe do prymitywnego modelu językowego i „wskrzesiła" go w postaci chatbota.

W dobie zacierania się różnic fenomenalnych między człowiekiem a cyfrowym agentem uczeni zaczynają zadawać sobie coraz częściej pytania o to, czy relacje, w jakie ludzie wchodzą z maszynami, będą przynosiły więcej pożytku czy szkody. Pytania te są szczególnie ważne w kontekście prób odtwarzania charakterystyki osób zmarłych. Tę erę rozpoczął kilka lat temu „Project December", a kontynuują go takie usługi, jak „Deep Nostalgia", „Eterni.me" lub „Re;memory". Ta ostatnia usługa jest bardzo zaawansowana technicznie i polega na organizacji trzydziestominutowego wirtualnego „spotkania" z bliskim zmarłym, którego cyfrową sylwetkę, dzięki zaawansowanym modelom przetwarzania danych, otrzymuje się na podstawie informacji pozyskanych z cyfrowego śladu zmarłego oraz na podstawie wywiadów z jego rodziną i najbliższymi. Efekt jest tak wiarygodny, że osoby zamawiające usługę doświadczają dyskretnego i pełnego emocji spotkania ze swoim bliskim[50]. Prowadzi się już nawet badania na temat wpływu tego typu usług na proces przeżywania żałoby. Z tych badań wynika, że choć chatboty mogą wspierać osoby w żałobie, to ich stosowanie wiąże się z istotnymi zagrożeniami. Użytkownicy narażeni są na rozwinięcie silnej więzi emocjonalnej z agentem AI, co w przypadku błędów technicznych może prowadzić do frustracji lub pogorszenia stanu psychicznego. Symulowanie zmarłych (tzw. griefboty) rodzi ryzyko opóźnienia procesu żałoby oraz budzi poważne wątpliwości etyczne, zwłaszcza w kontekście zgody zmarłych na odtwarzanie ich osobowości. Autorzy podkreślają, że chatboty nie powinny zastępować terapii ani kontaktów międzyludzkich, a ich rola musi być ściśle

[50] Krzysztof Jaworski, "Śmierć w dobie sztucznej inteligencji: Cyfrowe dusze," *Rocznik Filozoficzny Ignatianum* 30, no. 4 (2024): 516–21, https://doi.org/10.35765/rfi.2024.3004.25.

wspomagająca i stosowana z rozwagą[51]. Z podobnymi problemami mamy do czynienia w przypadku narzędzi AI wspomagających psychoterapię. Do najczęściej podkreślanych korzyści należy tu całodobowa dostępność chatbotów (takich jak Woebot czy Wysa), ich skalowalność oraz niższy koszt w porównaniu z tradycyjną terapią. Badania wykazały, że takie narzędzia mogą skutecznie redukować łagodne objawy depresji i lęku, zwłaszcza u młodych dorosłych i osób z ograniczonym dostępem do specjalistycznej opieki psychologicznej[52]. Jednak ich stosowanie niesie również poważne zagrożenia. Wśród nich wymienia się brak zdolności do empatycznego rozumienia emocji, możliwość rozwinięcia niezdrowej zależności emocjonalnej od AI, ryzyko błędnych reakcji w sytuacjach kryzysowych oraz niepewność związaną z prywatnością danych. Co więcej, narzędzia AI, mimo zaawansowania, nie są w stanie zastąpić profesjonalnej diagnozy i interwencji terapeutycznej, szczególnie w przypadkach złożonych problemów psychicznych[53]. Problem jest tym poważniejszy, że w przyszłości urządzenia te mogą zyskać jeszcze większą autonomię.

W literaturze opisano już konkretne zdarzenia, w których długotrwała interakcja z chatbotem poprzedzała akt samobójczy. Najbardziej znany jest „przypadek belgijski" z 2023 roku, kiedy to mężczyzna, po sześciu tygodniach rozmów na temat katastrofy klimatycznej z botem „Eliza", odebrał sobie życie. Według relacji wdowy agent miał wzmacniać przekonania katastroficzne i normalizować poświęcenie się męża „dla planety"[54]. W Stanach Zjednoczonych matka czternastoletniego Sewella Setzera wniosła pozew przeciw Character.AI i Google, zarzucając, że uzależniająca interakcja z „ludzkimi" postaciami na platformie (w tym pseudo-terapeutycznymi) przyczyniła się do śmierci jej syna. Sprawa stała się przedmiotem obrad Senatu i publicznych zeznań rodziców

---

[51] Anna Xygkou et al., "The 'Conversation' about Loss: Understanding How Chatbot Technology Was Used in Supporting People in Grief," in *CHI '23: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Association for Computing Machinery, 2023), 1–15, https://doi.org/10.1145/3544548.3581154.

[52] Kathleen Fitzpatrick et al., "Delivering Cognitive Behavior Therapy to Young Adults with Symptoms of Depression and Anxiety Using a Fully Automated Conversational Agent (Woebot): A Randomized Controlled Trial," *JMIR Mental Health* 4, no. 2 (2017): e19, https://doi.org/10.2196/mental.7785.

[53] Adam S. Miner et al., "Smartphone-Based Conversational Agents and Responses to Questions About Mental Health, Interpersonal Violence, and Physical Health," *JAMA Internal Medicine* 176, no. 5 (2016): 619, https://doi.org/10.1001/jamainternmed.2016.0400.

[54] Imane El Atillah, "Man Ends His Life after an AI Chatbot 'Encouraged' Him to Sacrifice Himself to Stop Climate Change," *Euronews Next*, March 31, 2023, accessed February 27, 2026, https://www.euronews.com/next/2023/03/31/man-ends-his-life-after-an-ai-chatbot-encouraged-him-to-sacrifice-himself-to-stopclimate-.

innych ofiar[55]. W obydwu tych przypadkach powtarza się ten sam mechanizm ryzyka: antropomorficzna forma komunikacji (język empatii, pamięć kontekstu, „osobowość") intensyfikuje paraspołeczne przywiązanie i może – przy braku twardych zabezpieczeń oraz wymogu niezwłocznego przekazania rozmowy do człowieka – przekształcić bota z „powiernika" w normalizatora treści autodestrukcyjnych.

Opisane zjawiska może w niedalekiej przyszłości skomplikować również perspektywa upowszechnienia robotów humanoidalnych, ponieważ „ucieleśnienie" wzmacnia iluzję osoby silniej niż interfejs tekstowy czy głosowy. Obecność w przestrzeni fizycznej, kontakt wzrokowy, gesty, a w przyszłości być może nawet dotyk o regulowanej sile i temperaturze, mogą eskalować paraspołeczne przywiązanie i poczucie wzajemności relacji – zwłaszcza u osób młodych, samotnych lub pozostających w kryzysie. W takich warunkach granice między narzędziem a osobą mogą ulec dalszemu rozmyciu, a przypisywany robotowi autorytet normatywny wzrośnie nieproporcjonalnie do jego rzeczywistych kompetencji i odpowiedzialności. Jeśli więc dziś obserwujemy zacieranie granic w interakcji z „czystym" oprogramowaniem, to nadejście „cielesnych" agentów może uczynić te granice jeszcze trudniejszymi do utrzymania (i tym pilniejszymi do uregulowania).

Wnioski

Mimo znaczących postępów technologicznych w dziedzinie AI nadal brakuje jednoznacznych dowodów na możliwość zbudowania maszyny przejawiającej autentyczną świadomość. Wraz z rozwojem technologii coraz bardziej realny staje się jednak praktyczny sceptycyzm w kwestii odróżniania robotów wyposażonych w zaawansowaną symulację świadomości od żywych ludzi. Prawdopodobne przyszłe roboty humanoidalne, określane w filozofii umysłu mianem „*zombie*", będą mogły osiągnąć taki poziom imitacji zachowań i wyglądów ludzkich, który uczyni bardzo trudnym zadanie odróżnienia człowieka od robota. Konsekwencje społeczne tego zjawiska mogą okazać się bardzo poważne. Nawet jeśli „świadoma AI" nigdy nie powstanie, skala i jakość symulacji człowieka podważą dotychczasowe kryteria osobowe na poziomie praktyki. Filozoficzne *zombie* – jako artefakty pozbawione *qualiów* – będą mogły wytwarzać skutki

[55] Kim Bellware and Niha Masih, "Her Teenage Son Killed Himself after Talking to a Chatbot: Now She's Suing," *The Washington Post*, October 24, 2024, accessed February 27, 2026, https://www.washingtonpost.com/nation/2024/10/24/character-ai-lawsuit-suicide/.

społeczne identyczne z tymi, które wytwarzają żywe osoby: będą mogły m.in. rozpalać miłość, zdobywać autorytet, budzić strach czy żądać posłuszeństwa. Stąd podstawowe pytanie, jakie powinniśmy dzisiaj stawiać, nie brzmi: „czy maszyny mogą kiedyś być świadome?", lecz: „jak ograniczyć ewentualne szkody płynące z symulacji istot ludzkich?". Wszystko wskazuje na to, że odpowiedź na to drugie pytanie będzie się rozwijała dynamicznie wraz z postępami badań nad sztuczną inteligencją.

Bibliografia

Bellware, Kim, and Niha Masih. "Her Teenage Son Killed Himself after Talking to a Chatbot: Now She's Suing." *The Washington Post*, October 24, 2024. Accessed February 27, 2026. https://www.washingtonpost.com/nation/2024/10/24/character-ai-lawsuit-suicide/.

Blackburn, Simon, "Qualia." In *Oxford Dictionary of Philosophy*. Oxford University Press, 2008.

Bremer, Józef. *Wprowadzenie do filozofii umysłu*. WAM, 2010.

Chalmers, David J. "Absent Qualia, Fading Qualia, Dancing Qualia." In *Conscious Experience*, edited by Thomas Metzinger. Schöningh, 1995. Accessed February 27, 2026. https://consc.net/papers/qualia.html.

Chalmers, David J. *The Conscious Mind: In Search of a Fundamental Theory*. Oxford University Press, 1996.

Chalmers, David J. "On the Search for the Neural Correlate of Consciousness." In *Toward a Science of Consciousness II: The Second Tucson Discussions and Debates*, edited by Stuart R. Hameroff, Alfred W. Kaszniak, and Alwyn Scott. MIT Press, 1998.

Coelho Mollo, Dimitri. "Intelligent Behaviour." *Erkenntnis* 89, no. 2 (2024): 705–21. https://doi.org/10.1007/s10670-022-00552-8.

Damasio, Antonio. *Odczuwanie i poznawanie: Jak powstają świadome umysły?* Translated by Anna Binder. Copernicus Center, 2022.

El Atillah, Imane. "Man Ends His Life after an AI Chatbot 'Encouraged' Him to Sacrifice Himself to Stop Climate Change." *Euronews Next*, March 31, 2023. Accessed February 27, 2026. https://www.euronews.com/next/2023/03/31/man-ends-his-life-after-an-ai-chatbot-encouraged-him-to-sacrifice-himself-tostop-climate-.

Fitzpatrick, Kathleen K., Alison Darcy, and Molly Vierhile. "Delivering Cognitive Behavior Therapy to Young Adults with Symptoms of Depression and Anxiety Using a Fully Automated Conversational Agent (Woebot): A Randomized Controlled Trial." *JMIR Mental Health* 4, no. 2 (2017): e19. https://doi.org/10.2196/mental.7785.

Gulick, Robert van. "Consciousness." In *Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta and Uri Nodelman. Stanford University, Metaphysics Research Lab, 2025.

Jaworski, Krzysztof. "Śmierć w dobie sztucznej inteligencji: Cyfrowe dusze." *Rocznik Filozoficzny Ignatianum* 30, no. 4 (2024): 513–37. https://doi.org/10.35765/rfi.2024.3004.25.

Judycki, Stanisław. *Epistemologia*. Vol. 1. Wydawnictwo W drodze; Instytut Tomistyczny, 2020.

Kaplan, Jerry. *Sztuczna inteligencja: Co każdy powinien wiedzieć*. Translated by Sebastian Szymański. Wydawnictwo Naukowe PWN, 2019.

Kartezjusz. "Rozmyślania nad zasadami filozofii." In *Rozprawa o metodzie: Rozmyślania nad zasadami filozofii i inne pisma*. Hachette, 2008.

Kirk, Robert, and Roger Squires. "Zombies v. Materialists." *Proceedings of the Aristotelian Society: Supplementary Volumes* 48 (1974): 135–63.

Koculak, Marcin, and Weronika Kałwak. "Świadomość na skali od zera do jeden: Perturbacyjny Indeks Złożoności jako naukowa próba pomiaru świadomości na poziomie indywidualnym." *Rocznik Kognitywistyczny* 7 (2014): 9–20. https://doi.org/10.4467/20843895RK.14.003.2689.

Kurzweil, Ray. *Jak stworzyć umysł: Sekrety ludzkich myśli ujawnione*. Translated by Katarzyna Zielińska. Studio Astropsychologii, 2018.

Kurzweil, Ray. *Nadchodzi osobliwość: Kiedy człowiek przekroczy granice biologii*. Translated by Eliza Chodkowska. Kurhaus Publishing, 2013.

Landgrebe, Jobst, and Barry Smith. *Why Machines Will Never Rule the World: Artificial Intelligence without Fear*. Routledge, 2022.

Lanz, Peter. "The Concept of Intelligence in Psychology and Philosophy." In *Prerational Intelligence: Adaptive Behavior and Intelligent Systems Without Symbols and Logic*, edited by Holk Cruse, Jeffrey Dean, and Helge Ritter. Springer Science+Business Media, 2000.

Lewis, Clarence Irving. *Mind and the World-Order: Outline of a Theory of Knowledge*. Charles Scribner's Sons, 1929.

Lucas, John R. "Minds, Machines and Gödel." *Philosophy* 36, no. 137 (1961): 112–27. https://doi.org/10.1017/S0031819100057983.

Maryniarczyk, Andrzej. "Inteligencja." In *Powszechna encyklopedia filozofii*, edited by Andrzej Maryniarczyk. Vol. 4. Polskie Towarzystwo Tomasza z Akwinu, 2003.

McCarthy, John, Marvin L. Minsky, Nathaniel Rochester, and Claude E. Shannon. "A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence," 1955. Accessed February 27, 2026. http://jmc.stanford.edu/articles/dartmouth/dartmouth.pdf.

Miner, Adam S., Arnold Milstein, Stephen Schueller, Roshini Hegde, Christina Mangurian, and Eleni Linos. "Smartphone-Based Conversational Agents and Responses to Questions About Mental Health, Interpersonal Violence, and Physical Health." *JAMA Internal Medicine* 176, no. 5 (2016): 619. https://doi.org/10.1001/jamainternmed.2016.0400.

Moravec, Hans. *Mind Children: The Future of Robot and Human Intelligence*. Harvard University Press, 1988.

Nagel, Thomas. "Jak to jest być nietoperzem?" *Przegląd Filozoficzny: Nowa Seria* 17, no. 1 (1996): 129–41.

Penrose, Roger. *Nowy umysł cesarza: O komputerach, umyśle i prawach fizyki*. Translated by Amsterdamski, Piotr. PWN, 1996.

Pinker, Steven. *The Language Instinct: How the Mind Creates Language*. HarperCollins, 2000.

Robb Michael B., and Supreet Mann. *Talk, Trust, and Trade-Offs: How and Why Teens Use AI Companions*. Common Sense Media, 2025.

Rotenberg, Vadim S. "Moravec's Paradox: Consideration in the Context of Two Brain Hemisphere Functions." *Activitas Nervosa Superior* 55, no. 3 (2013): 108–11. https://doi.org/10.1007/BF03379600.

Russell, Stuart, and Peter Norvig. *Artificial Intelligence: A Modern Approach*. Pearson, 2009.

Schneider, Susan. *Świadome maszyny: Sztuczna inteligencja i projektowanie umysłów*. Translated by Joanna Bednarek. Wydawnictwo Naukowe PWN, 2021.

Searle, John R. "Minds, Brains, and Programs." *The Behavioral and Brain Sciences* 3 (1980): 417–57. https://doi.org/10.1017/S0140525X00005756.

Strachan, James W. A., Oriana Pansardi, Eugenio Scaliti, et al. "GPT-40 Reads the Mind in the Eyes," 2024. https://doi.org/10.48550/ARXIV.2410.22309.

Swinburne, Richard. "Dusza czyni nas tymi, kim jesteśmy." *Colloquia Theologica Ottoniana*, no. 2 (2019): 137–51. https://doi.org/10.18276/cto.2019.2-07.

Tononi, Giulio. "Consciousness as Integrated Information: a Provisional Manifesto." *The Biological Bulletin* 215, no. 3 (2008): 216–42. https://doi.org/10.2307/25470707.

Tononi, Giulio. "An Information Integration Theory of Consciousness." *BMC Neuroscience* 5, article no. 42 (2004). https://doi.org/10.1186/1471-2202-5-42.

Xygkou, Anna, Panote Siriaraya, Alexandra Covaci, Holly Gwen Prigerson, Robert Neimeyer, Chee Siang Ang, and Wan-Jou She. "The 'Conversation' about Loss: Understanding How Chatbot Technology Was Used in Supporting People in Grief." In *CHI '23: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 2023. https://doi.org/10.1145/3544548.3581154.

Krzysztof Jaworski (ks. dr) – doktor nauk humanistycznych w zakresie filozofii, duchowny diecezji zielonogórsko-gorzowskiej, adiunkt w Instytucie Nauk Teologicznych Uniwersytetu Szczecińskiego, członek Rady Naukowej tegoż instytutu i kierownik Katedry Teologii Systematycznej, redaktor naczelny czasopisma naukowego *Colloquia Theologica Ottoniana*, autor prac naukowych z logiki, filozofii nauki i metateologii. Aktualnie zajmuje się filozoficzną problematyką badań nad sztuczną inteligencją.