

Adam Głowacki
Wydział Filozofii

Świadomość maszyn w obliczu trudnego problemu świadomości

Wraz ze wzrostem mocy obliczeniowej komputerów oraz rozwojem technologii powstają coraz skuteczniejsze modele sztucznej inteligencji, które przejawiają umiejętności mówienia, pisania, wnioskowania i rysowania na poziomach porównywalnych z kompetencją ludzi. Współcześnie podejmowane są dyskusje nad tym, gdzie stoi granica kompetencji maszyn i programów pokroju GPT-4 (OpenAI, 2023) czy Gemini (Team, G.T.M.D.P.H.C.-I. *et al.*, 2023). W niniejszej pracy przedstawiam problemy granic poznania multimodalnych modeli sztucznej inteligencji oraz ich świadomości, korzystając z ram teoretycznych nauk o świadomości, a w szczególności, z konsekwencji trudnego problemu świadomości.

Problem świadomości można podzielić na dwa podproblemy, szerzej znane jako łatwy oraz trudny problem świadomości (Chalmers, 1997). Łatwy problem świadomości wynika z psychologicznego charakteru umysłu i dotyczy neuronalnych korelatów świadomości, podstawowych mechanizmów, które umożliwiają doświadczanie przytomności przez podmiot w danej sytuacji. Współczesne metody obrazowania pozwalają na coraz dokładniejszy wgląd w mechanizm mózgu, a wraz z rozwojem technologii będziemy w stanie opisać jego funkcjonowanie z dokładnością do pojedynczego potencjału wywołanego. Trudny problem świadomości dotyczy subiektywnego doświadczenia i fenomenalnych¹ aspektów umysłu, które tajemniczo wyłaniają się z procesów poznawczych i nie są obserwowalne (jak dotąd) przez żadną metodę obrazowania.

Jeżeli subiektywne, wewnętrzne, świadome przeżycia nie są obserwowalne z poziomu funkcjonowania mózgu, to również nie będą obserwowalne z poziomu funkcjonowania systemu, według którego stworzony jest model sztucznej inteligencji. Zatem jedynymi narzędziami dostępnymi do badania wewnętrznych przeżyć maszyn są narzędzia oparte na zachowaniu, takie jak test Turinga (Turing, 1950). Słynny przypadek Blake’a Lemoine’a, inżyniera pracującego nad modelem LaMDA² pokazuje jednak, że są to narzędzia zawodne: model językowy mógł być zwyczajnie wytrenowany na danych, w których występują teksty o świadomości. Podobny problem ilustruje eksperyment myślowy o chińskim pokoju (Searle, 1980), w którym wykorzystywane są tylko instrukcje do tłumaczenia języka, bez jego rozumienia (w sensie rozumienia świadomego).

Istnieje jednak jedna, obiecująca (lecz kontrowersyjna) metoda, która nie opiera się ani na danych behawiorystycznych, ani na oddolnych procesach systemowych. Teoria zintegrowanej informacji³ zakłada istnienie świadomości i używa matematycznych narzędzi do analizy złożoności systemu, które pozwala na ilościowe wnioskowanie o procesach świadomych. Niestety, posiada wiele wad, jednakże wyznacza pewien kierunek w analizie systemów multimodalnych, które, przede wszystkim, przetwarzają ogromne ilości informacji.

Brak solidnych metod naukowych do opisywania i wyjaśniania jakościowych stanów mentalnych zmusza do wyrażania sądów na temat świadomości maszyn jedynie w obrębie filozofii umysłu, co nie przyczynia się w znacznym stopniu do empirycznych odkryć w kwestii świadomości. Daje to nam jedynie ramy teoretyczne do dyskusji oraz podstawy w tworzeniu metod naukowych. W tym momencie rozwoju technologicznego, trudny problem świadomości oraz problem świadomości maszyn nadal pozostają nierozwiązywalne.

1 Jakościowych stanów mentalnych takich jak *odczuwanie* emocji.

2 Transkrypcja rozmowy dostępna pod linkiem: <https://www.aidataanalytics.network/data-science-ai/news-trends/full-transcript-google-engineer-talks-to-sentient-artificial-intelligence-2>

3 Po więcej szczegółów: http://www.scholarpedia.org/article/Integrated_information_theory

Bibliografia

Chalmers, D.J. (1997) *The conscious mind: In Search of a Fundamental Theory*. Oxford Paperbacks.

OpenAI (2023) 'GPT-4 Technical Report,' *arXiv (Cornell University)* [Preprint]. <https://doi.org/10.48550/arxiv.2303.08774>.

Searle, J.R. (1980) 'Minds, brains, and programs,' *Behavioral and Brain Sciences*, 3(3), pp. 417–424. <https://doi.org/10.1017/s0140525x00005756>.

Team, G.T.M.D.P.H.C.-I. *et al.* (2023) 'Gemini: a family of highly capable multimodal models,' *arXiv (Cornell University)* [Preprint]. <https://doi.org/10.48550/arxiv.2312.11805>.

Turing, A. (1950) 'I.—COMPUTING MACHINERY AND INTELLIGENCE,' *Mind*, LIX(236), pp. 433–460. <https://doi.org/10.1093/mind/lix.236.433>.