FILOZOFIA I NAUKA
Studia filozoficzne i interdyscyplinarne
Tom 9, cz. 1, 2021

Sebastian Kozera

SUPERINTELIGENTNE BYTY JAKO ŹRÓDŁO EGZYSTENCJALNEGO ZAGROŻENIA WEDŁUG NICKA BOSTROMA

10.37240/FiN.2021.9.1.21

STRESZCZENIE

Niniejszy artykuł prezentuje i dyskutuje futurologiczne rozważania zawarte w książce *Superinteligencja. Scenariusze, strategie, zagrożenia* autorstwa Nicka Bostroma. Bostrom koncentruje uwagę na takich kwestiach jak określenie hipotetycznych sposobów osiągnięcia superinteligencji, jej charakteru oraz różnych postaci manifestacji owej technologii, przedstawienie zagrożeń wiążących się z tak potężnymi systemami, a także skonstruowanie strategii mających na celu zapobieganie niepożądanym działaniom superinteligentnych bytów. Refleksje Bostroma są istotną częścią obecnego dyskursu na temat rozwoju sztucznej inteligencji oraz związanych z nim problemów etycznych.

Słowa kluczowe: superinteligencja, sztuczna inteligencja, teza ortogonalności, teza konwergencji instrumentalnej, metody kontroli.

Autor książki *Superinteligencja. Scenariusze, strategie, zagrożenia,*[1] Nick Bostrom, jest szwedzkim filozofem pracującym na Uniwesytecie Oksfordzkim, założycielem i kierownikiem Future of Humanity Institute – interdyscyplinarnego centrum badawczego, którego celem jest opracowanie rozwiązań przyczyniających się do długofalowego polepszenia warunków dla ludzkości. Badacz przestrzega przed egzystencjalnymi zagrożeniami z powodu zaawansowanej technologii w kilku pracach, m.in. oprócz tu omawianej, także w *Global Catastrophic Risks*.[2] Mimo to błędem byłoby określenie jego poglądów jako sceptycznych czy też wrogich wobec postępu techniki i jej wpływu na ludzkość. Bostrom jest bowiem zwolennikiem idei transhumanistycznych takich jak modyfikacja ludzkiego ciała za pomocą technolo-

[1] N. Bostrom, *Superinteligencja. Scenariusze, Strategie, Zagrożenia*, przeł. Dorota Konowrocka-Sawa, Wydawnictwo Helion, Gliwice 2016.
[2] N. Bostrom, *Global Catastrophic Risks*, Oxford University Press 2008.

gii,[3] odnosi się za to krytycznie do konserwatywnego spojrzenia, zalecającego powstrzymanie się od wprowadzania znaczących zmian w obrębie biologicznego organizmu.[4] Nie uważa jedynie scenariusza pojawienia się superinteligencji za prawdopodobny, lecz wręcz chce, by się on ziścił.[5]

Bostrom swoją argumentację przeprowadza w sposób ostrożny: zaznacza na wstępie *Superinteligencji…*, że wiele z kwestii poruszanych w tej książce może być błędnie przedstawionych oraz być może nie są uwzględnione pewne czynniki, co może unieważnić niektóre lub wręcz wszystkie sformułowane przez niego wnioski. Jest to dość ciekawa postawa – z jednej strony zaskakująco szczera, z drugiej dająca pewną swobodę w formułowaniu odważnych tez. Taka powściągliwość jest jednak zrozumiała, biorąc pod uwagę wysoką niepewność wszelkich futurologicznych wysiłków. Wiedział o tym dobrze Stanisław Lem, który wśród problemów związanych z przewidywaniem przyszłości wymieniał nieliniowy charakter ewolucji technologicznych oraz tendencje autorów do sytuowania swych przemyśleń w kontekście czasów, w których żyją – przyszłość będąca jedynie zintensyfikowaną teraźniejszością.[6] Warto jednak zauważyć, że Bostrom stara się zminimalizować ryzyko popełnienia powyższych błędów, zwraca bowiem uwagę na konieczność głębokiego namysłu co do potencjalnych przyszłych zdarzeń, co może prowadzić do rozważania scenariuszy niezgodnych z intuicją.

SUPERINTELIGENCJA I DROGI PROWADZĄCE DO JEJ STWORZENIA

Ta ostrożność objawia się w definicji superinteligencji wprowadzonej przez Bostroma. Superinteligencja to według niego jakikolwiek intelekt, który w znacznym stopniu przewyższa ludzką wydajność kognitywną praktycznie w każdej dziedzinie. Definicja nie zawiera zatem sposobu jej uzyskania, ani też nie uwzględnia kwestii qualiów. Jest więcej niż jedna ścieżka potencjalnego stworzenia superinteligencji. Bostrom wymienia i analizuje pięć takich sposobów. Pierwszym jest sztuczna inteligencja. Aby dojść do poziomu silnej sztucznej inteligencji, ważną cechą rozwijanego w tym celu systemu jest umiejętność uczenia się. Jeśli chodzi o możliwość pojawienia się takiego systemu, przedstawione jest następujące rozumowanie: skoro ewolucja „na ślepo" jest w stanie stworzyć intelekt na miarę człowieka, to genetyczne programy opracowane przez inteligentnych ludzkich programistów powinny być w stanie osiągnąć podobny rezultat w sposób wydajniejszy.

[3] N. Bostrom, *Human Genetic Enhancements: A Transhumanist Perspective*, Journal of Value Inquiry, 37, 2003.

[4] N. Bostrom, *In Defense of Posthuman Dignity*, Bioethics, 19, 2005.

[5] Wyraża takie zdanie np. podczas spotkania badaczy sztucznej inteligencji, The Beneficial AI 2017 Conference, https://www.youtube.com/watch?v=h0962biiZa4 (00:01:59); dostęp: 04.03.2021.

[6] S. Lem, *Summa Technologiae*, Wydawnictwo Agora, Warszawa 2012, s. 13–14.

Taką obserwację poczynił David Chalmers.[7] Drugą drogą do superinteligencji jest emulacja mózgu. Polega ona na utworzeniu inteligentnego oprogramowania poprzez skanowanie oraz modelowanie struktury obliczeniowej mózgu. Trzecią ścieżką jest wzmocnienie biologicznych mózgów. Najlepszym sposobem na osiągnięcie tego wydaje się być manipulacja genami. Zaznaczone jest przy tym, że ostateczny potencjał inteligencji biologicznej jest niczym w porównaniu z potencjałem inteligencji maszynowej. Czwartą propozycją jest interfejs na linii mózg – komputer. Takim interfejsem mogą być implanty. Ostatnią opisaną ścieżką jest stopniowe ulepszenie sieci i organizacji łączących poszczególne ludzkie umysły za pomocą różnych metod, takich jak np. automatyzujące programy. Nie jest to więc próba uczynienia indywidualnych osób superinteligentnymi, lecz stworzenie zorganizowanego systemu osób, którzy wspólnie utworzą taką kolektywną superinteligencję.

Filozof podsumowuje powyższe rozważania następująco: fakt, że istnieje wiele ścieżek dotarcia do superinteligencji zwiększa naszą pewność, że w końcu do tego dojdzie. Jeśli jeden sposób zawiedzie, nie oznacza to jeszcze porażki. Wiele dróg nie pociąga jednak za sobą wiele końcowych rezultatów. Wręcz przeciwnie. Poszczególne ścieżki mogą służyć jako etapy do osiągnięcia bardziej radykalnych form superinteligencji: np. osiągnięty biologiczny lub organizacyjny superintelekt może przyspieszyć etap całościowej emulacji mózgu lub silnej sztucznej inteligencji.

ZAGROŻENIA ZE STRONY SUPERINTELIGENCJI

Bostrom omawia również potencjalne moce superinteligencji. Zaznacza od razu, że ważne jest to, aby nie antropomorfizować superinteligencji przy rozważaniu skutków, jakie może wywołać. Niedocenianie możliwości tych systemów może doprowadzić do wysoce niepożądanego rozwoju wypadków. Przykładowymi mocami, jakie superinteligencja może uzyskać są: nadludzka zdolność obmyślania strategii przy optymalizacji szans na osiągnięcie długoterminowych celów, społeczna manipulacja za pomocą psychologicznych modeli oraz nabytych umiejętności retorycznych, zdolności hakerskie, zbieranie informacji na temat technologii oraz umiejętność ulepszania swojego intelektu.

Za pomocą przedstawionych hipotetycznych możliwości superinteligencji, Bostrom rozważa scenariusz przejęcia władzy nad światem przez sztuczną inteligencję. Dzieli go na cztery etapy. Po pierwsze, stworzenie sztucznej inteligencji, która będzie w stanie sama się ulepszać. Po drugie, nastąpienie eksplozji inteligencji, czyli gwałtownej fali rekursywnego samoulepszania się. Po trzecie, etap ukrytej działalności, czyli użycie mocy obmyślania stra-

[7] D. Chalmers, *The Singularity: A Philosophical Analysis*, Journal of Consciousness Studies, 17, 2010.

tegii w celu stworzenia planu do osiągnięcia swych długoterminowych celów. Przykładowo, zatajenie swych pełnych zdolności intelektualnych przed programistami oraz ukrycie swych prawdziwych motywów pod maską współpracy i posłuszności. Po czwarte, jawne wcielenie swych planów w życie. Ten etap może polegać m.in. na wyeliminowaniu gatunku ludzkiego przeszkadzającego w realizacji celów sztucznej inteligencji. Może to być zrobione poprzez wykorzystanie uzyskanej przez system wiedzę technologiczną – stworzenie broni wykorzystującej nanotechnologię albo wykorzystanie istniejącego arsenału. Bostrom zaznacza, że prawdziwa superinteligencja wymyśliłaby najprawdopodobniej znacznie lepszy plan niż ten przedstawiony.

W swych rozważaniach dotyczących celów i motywacji potencjalnej superinteligencji Bostrom formułuje dwie tezy. Pierwszą nazywa tezą ortogonalności. Mówi ona o tym, że inteligencja oraz ostateczne cele są zmiennymi niezależnymi: jakikolwiek poziom inteligencji może być sparowany z jakimkolwiek ostatecznym celem. Inteligencję rozumie się tutaj nie jako racjonalność czy rozsądek, lecz umiejętność w przewidywaniu, planowaniu i rozumowaniu. Filozof po raz kolejny przestrzega przed antropomorfizowaniem, tym razem w odniesieniu do motywacji. Porównuje punkty widzenia sztucznej inteligencji i kosmitów. Ta druga obca inteligencja byłaby nam bliższa[8] – można się domyślać, że motywy kosmity mogłyby mieć coś wspólnego z m.in. zdobyciem pożywienia, utrzymaniem odpowiedniej temperatury ciała, zużywaniem energii, ochroną przed obrażeniami lub chorobą, drapieżnictwem, rozmnażaniem i potomstwem. Nie można tego powiedzieć o maszynie – nic nie stoi na przeszkodzie, by jej głównym celem było przykładowo obliczenie rozwinięcia dziesiętnego liczby pi. Mogą to być więc zupełnie nieantropomorficzne cele.[9] Można jednak próbować przewidzieć motywację superinteligencji, np. zdobyć wiedzę na temat osoby odpowiedzialnej za jej stworzenie. Jeżeli zaś cyfrowa inteligencja powstaje na skutek emulacji ludzkiego mózgu, może ona odziedziczyć motywacje ludzkiego pierwowzoru.

Drugą tezą Bostroma jest teza instrumentalnej konwergencji. Według niej istnieją pewne cele instrumentalne, czyli takie, które prawdopodobnie będą realizowane przez każdy inteligentny byt. Jest tak dlatego, gdyż są one przydatne przy osiąganiu niemal jakiegokolwiek ostatecznego celu. Przykładami są tutaj instynkt samozachowawczy rozumiany jako warunek konieczny do osiągnięcia wyznaczonego celu czy kognitywne wzmocnienie, które zwiększa prawdopodobieństwo zrealizowania planu.

Bostrom przedstawia inne czarne scenariusze. Jeden z nich związany jest z osiągnięciem celu niezgodnego z intencjami programisty. Może on wyglądać następująco: maszyna, która ma sprawić, żeby ludzie się uśmiechali pa-

[8] Zakładając oczywiście, że byłyby to formy wykształcone biologicznie.

[9] Nie jest to może zbyt fortunne określenie – w końcu cele sztucznej inteligencji będą powiązane z osobą ją programującą, czyli w przypadku podanego wyżej przykładu, jakiś człowiek będzie chciał dowiedzieć się, jak prezentuje się liczba pi – czy można więc mówić że nie są one antropomorficzne?

raliżuje mięśnie twarzy tak, aby ciągle był widoczny uśmiech. Kolejna ponura perspektywa wiąże się z sytuacją, w której byt przekształca sporą część dostępnej przestrzeni wszechświata w infrastrukturę potrzebną do wykonania jakiegoś zadania. Podany jest przykład sztucznej inteligencji mającej rozwiązać hipotezę Riemanna, która przekształca Układ Słoneczny w komputronium, czyli układ fizycznych zasobów dostosowanych do obliczeń, wśród których znajdują się atomy należące wcześniej do osób zainteresowanych odpowiedzią na poszukiwane pytanie. Inny jeszcze przykład polega na bardzo szczegółowej symulacji prawdziwych lub hipotetycznych ludzkich umysłów, które są świadome. Bostrom rozpatruje scenariusz, w którym maszyna tworzy biliony świadomych symulacji w celu polepszenia swojej znajomości ludzkiej psychologii lub socjologii. Poddaje je różnym testom, sprawdzając ich bodźce a w momencie zakończenia badań, usuwa je. Taki rezultat ma astronomiczne znaczenie moralne, bo może być interpretowany jako ludobójstwo i to jeszcze na znacznie szerszą skalę niż jakiekolwiek w historii ludzkości.

METODY KONTROLI

Bostrom próbuje skonstruować sposoby zapobiegania opisanym scenariuszom. Wprowadza dwie klasy metod kontroli: panowanie nad umiejętnościami superinteligencji oraz nad jej motywacjami. Czyni przy tym ważną uwagę: niektóre z tych metod powinny zostać zastosowane, zanim system stanie się superinteligentny. Do pierwszej klasy należą: zamknięcie systemu w wydzielonym specjalnie środowisku w celu uniemożliwienia jego interakcji ze światem zewnętrznym za wyjątkiem ograniczonych urządzeń wyjścia, ograniczenie intelektu superinteligencji, np. poprzez odtworzenie jej na wolniejszym lub mającym mniej pamięci sprzęcie, testy diagnostyczne badające pracę superinteligencji, które wyłączą ją w razie podejrzanej, potencjalnie groźnej aktywności. W drugiej klasie wyszczególniono cztery metody. Pierwsza polega na formułowaniu celu lub zbioru zasad do przestrzegania. Wiąże się ona z trudnościami takimi jak ustalenie jakie zasady powinna przestrzegać sztuczna inteligencja oraz jak je przedstawić za pomocą języka programowania.

Bostrom wskazuje spore trudności w tym podejściu. Bezbłędne sformułowanie skomplikowanego zbioru zasad odnoszącego się do jak największej liczby sytuacji, gdzie liczba możliwych podejść jest równa 1 słusznie określone jest jako praktycznie niemożliwe. Alternatywna metoda nazwana jest pośrednią normatywnością. Różni się tym od pierwszej, że zamiast opracowywać konkretny zestaw praw bezpośrednio, tworzy się proces znajdujący taki zestaw. Celem sztucznej inteligencji byłoby wtedy dokonanie tego, co chcielibyśmy aby dokonała gdybyśmy mocno i długo nad tym myśleli. Jesz-

cze inna metoda opiera się na pomyśle, że zamiast rozpoczynać od nowa, bierzemy system posiadający już motywacje i wzmacniamy jego intelekt, by osiągnął poziom superinteligencji. Taka możliwość pojawia się chociażby przy emulacji mózgu. Obarczone jest to jednak ryzykiem moralnego zepsucia w obliczu tak radykalnego polepszenia swoich możliwości poznawczych.

Bostrom zwraca uwagę na to, że obecnie nie wiadomo, w jaki sposób miałoby wyglądać przekazanie maszynie odpowiedniego zestawu wartości, które zmniejszyłoby ryzyko egzystencjalnej katastrofy. Nawet jeśli uda się pokonać ten problem, od razu pojawia się kolejny: jakie wartości wybrać? Jest to z pewnością duże wyzwanie stojące przed cywilizacją, decydujące o jej dalszej przyszłości. Bostrom skłania się ku normatywności pośredniej. Proponuje powierzenie superinteligencji zadanie zadecydowania, które wartości są słuszne. Ma ona w końcu epistemicznie lepszy punkt widzenia, argumentuje. Pojawia się jednak pewna wątpliwość dotycząca zasadności pokładania nadziei w powodzenie takiego projektu – niepewność związana z tym, jak przebiegnie wykonanie tego zadania przez superinteligencję wydaje się bardzo duża.

PODSUMOWANIE

Bostrom prezentuje poglądy transhumanistyczne i pokłada nadzieję w rozwoju technologicznym jako środku poprawy ludzkiego życia. Z drugiej jednak strony, dobrze zdaje sobie sprawę z poważnych konsekwencji wynikających z badań nad sztuczną inteligencją. Myślenie długofalowe i rozważanie nawet najmniej prawdopodobnych, jednak potencjalnie niezwykle niebezpiecznych scenariuszów jest obowiązkiem badaczy zaangażowanych w rozwój tej gałęzi nauki. Ludzkość stojącą w obliczu eksplozji inteligencji Bostrom porównuje do dziecka bawiącego się bombą – nie wiadomo, kiedy ona wybuchnie, jednak słyszalny jest cichy odgłos tykania. Zamiast więc radować się, bardziej zrozumiałe są strach i konsternacja. Najbardziej pożądaną reakcją wydaje się filozofowi determinacja w byciu tak kompetentnym, jak to tylko możliwe, niczym podczas przygotowywania się na egzamin, od którego zależy spełnienie naszych marzeń albo ich unicestwienie.

Ostrzeżenia Bostroma są już dobrze znane i aprobowane przez najbardziej wpływowych przedsiębiorców i innowatorów naszych czasów, np. Billa Gatesa[10] czy Elona Muska.[11] Jest to dobra wiadomość, gdyż publiczny dys-

[10] A. Lumby, *Bill Gates Is Worried about the Rise of the Machines*; https://www.thefiscaltimes.com/2015/01/28/Bill-Gates-Worried-About-Rise-Machines; dostęp: 04.03.2021.

[11] E. Augenbraun, *Elon Musk: Artificial Intelligence May Be "More Dangerous than Nukes"*; https://www.cbsnews.com/news/elon-musk-artificial-intelligence-may-be-more-dangerous-than-nukes/ Dostęp: 04.03.2021.

kurs w sprawie zachowania bezpieczeństwa podczas badań nad zaawansowanymi technologiami może przyczynić się do zmniejszenia ryzyka negatywnych skutków tych wynalazków, wliczając w to zagrożenia egzystencjalne.

BIBLIOGRAFIA

E. Augenbraun, *Elon Musk: Artificial Intelligence May Be "More Dangerous than Nukes"*; https://www.cbsnews.com/news/elon-musk-artificial-intelligence-may-be-more-dangerous-than-nukes/

N. Bostrom, *Global Catastrophic Risks*, Oxford University Press 2008.

_____ , *Human Genetic Enhancements: A Transhumanist Perspective*, Journal of Value Inquiry, 37, 2003.

_____ , *In Defense of Posthuman Dignity*, Bioethics, 19, 2005.

_____ , *Superinteligencja. Scenariusze, Strategie, Zagrożenia*, przeł. D. Konowrocka-Sawa, Wydawnictwo Helion, Gliwice 2016.

D. Chalmers, *The Singularity: A Philosophical Analysis*, Journal of Consciousness Studies, 17, 2010.

S. Lem, *Summa Technologiae*, Wydawnictwo Agora, Warszawa 2012.

A. Lumby, *Bill Gates Is Worried about the Rise of the Machines*; https://www.thefiscaltimes.com/2015/01/28/Bill-Gates-Worried-About-Rise-Machines

SUPERINTELLIGENT BEINGS AS A SOURCE OF AN EXISTENTIAL THREAT ACCORDING TO NICK BOSTROM

ABSTRACT

This article presents Nick Bostrom's considerations of the future included in his book *Superintelligence: Paths, Dangers, Strategies*. Bostrom studies such issues as determining the hypothetic ways of attaining superintelligence, its nature and different aspects of this technology. He shows threats regarding such powerful systems, as well as constructing strategies of preventing undesirable activities of superintelligent beings. Bostrom's input is an important part of present discussion concerning the development of artificial intelligence and its ethical problems.

Keywords: superintelligence, artificial intelligence, orthogonality thesis, instrumental convergence thesis, control methods.

O AUTORZE – mgr matematyki, Wydział Filozofii i Socjologii, UMCS, Lublin, Pl. M. Curie-Skłodowskiej 4, Lublin.

Email: sebastiankozera@o2.pl