Kultura i Wartości
ISSN 2299-7806
Nr 20/2016

http://dx.doi.org/10.17951/kw.2016.20.183

Czy powinniśmy się obawiać sztucznej inteligencji?

Recenzja: Nick Bostrom, *Superinteligencja. Scenariusze, strategie, zagrożenia*, tłum. D. Konowrocka-Sawa, Wydawnictwo Helion, Gliwice 2016 (s. 488)

Kamil Szymański

Prawdopodobnie większość z nas styka się na co dzień z algorytmami, które można uznać za przejawy sztucznej inteligencji (SI). Siadając rano do komputera i logując się do Facebooka, otrzymujemy dawkę posegregowanych informacji na temat naszych znajomych, a także związanych z polubionymi przez nas stronami. Nad tym, byśmy otrzymywali tylko te wiadomości, na których nam zależy, czuwają algorytmy, które analizują nasze upodobania i na ich podstawie wyświetlają nam odpowiednie treści. Innym przykładem sztucznej inteligencji może być Siri – inteligentny osobisty asystent, który znajduje się w smartfonach marki Apple. Dzięki funkcji rozpoznawania głosu mamy możliwość „kazać jej" wezwać taksówkę, zaś sygnał GPS skieruje do nas pożądany pojazd. Możemy poprosić ją o znalezienie restauracji w pobliżu nas, w których, przykładowo, serwują smaczne burgery. Możemy nawet posłuchać dowcipów, które opowiada, czy porozmawiać o pogodzie. Ilość sprzętów, jakie posiadają algorytmy, które możemy nazwać „prostą" sztuczną inteligencją jest ogromna i ciągle rośnie – autonomiczne samochody, lodówki, które „same robią zakupy", pociągi w metrze.

KAMIL SZYMALSKI, doktorant nauk o poznaniu i komunikacji społecznej na Wydziale Filozofii i Socjologii UMCS w Lublinie, doktorant filozofii na Wydziale Filozofii KUL; adres do korespondencji: Instytut Filozofii UMCS, Pl. M. Curie-Skłodowskiej 4, 20-031, Lublin; e-mail: szym.kamil@gmail.com

Technika staje się coraz bardziej „niezależna" od człowieka. Pojawienie się sztucznej inteligencji stanowić będzie dla człowieka ogromne ułatwienie w jego codziennych działaniach. Jednak wielu badaczy, a także osobistości związanych ze współczesnym rozwojem techniki, jak Elton Musk[1] czy Steven Hawking[2], przestrzegają przed bezkrytycznym, i pozbawionym nadzoru rozwoju SI, która może faktycznie w przyszłości zagrozić człowiekowi i jego istnieniu.

Kwestia sztucznej inteligencji poruszana jest przez Nicka Bostroma w jego najnowszej książce zatytułowanej *Superinteligencja. Scenariusze, strategie, zagrożenia* (oryginalne wydanie w roku 2014, przekład w języku polskim w roku 2016). Nick Bostom, profesor Uniwersytetu Oxfordzkiego, jest kierownikiem Instytutu Przyszłości Ludzkości[3]. Zajmuje się problemami rozwoju techniki, transhumanizmem oraz związanym z tym zagrożeniami egzystencjalnymi. Jego najnowsza książka poświęcona jest zagadnieniu „zaawansowanej" i w pełni autonomicznej sztucznej inteligencji: historii rozwoju owej koncepcji, jej możliwych sposobów funkcjonowania oraz zagrożeniom, które mogą wyniknąć z jej powstania. Książka zawiera piętnaście rozdziałów, które dodatkowo podzielone są na podrozdziały (od trzech do dziewięciu). Rozdziały zatytułowane są kolejno: *Dotychczasowe dokonania i obecne możliwości*; *Ścieżki wiodące ku superinteligencji*; *Formy superinteligencji*; *Dynamika eksplozji inteligencji*; *Decydująca przewaga strategiczna*; *Poznawcze supermoce*; *Pobudki superinteligencji*; *Czy czeka nas zagłada?*; *Problem kontroli*; *Wyrocznie*; *Dżiny, suwereni i narzędzia*; *Scenariusze wielobiegunowości*; *Zaszczepianie wartości*; *Wybór kryteriów wyboru*; *Perspektywa strategiczna*; *Moment krytyczny*.

Poszczególne rozdziały są podzielone równomiernie, od około 30 do 40 stron, dając łącznie z wstępem, zakończeniem i obszerną bibliografią – 488 stron. Największą wadą polskiego wydania jest umieszczenie przypisów na końcu książki. Liczba przypisów waha się od 40 do 60 dla każdego rozdziału, a całość przypisów zawiera się na 78 stronach, co znacznie utrudnia wygodne korzystanie z książki.

---

[1] https://www.theguardian.com/technology/2014/oct/27/elon-musk-artificial-intelligence-ai-biggest-existential-threat [dostęp: 10.02.2017].

[2] http://www.bbc.com/news/technology-30290540 [dostęp: 10.02.2017].

[3] http://www.philosophy.ox.ac.uk/members/other_faculty/bostrom_nick [dostęp: 10.02.2017].

Pierwszy rozdział to wprowadzeni do problematyki sztucznej inteligencji. Autor dokonuje spojrzenia wstecz na rozwój gatunku ludzkiego, wskazując na wciąż przyśpieszający rozwój naukowo-techniczny, nawiązując przy tym do koncepcji osobliwości (ang. *Singularity*) zaproponowanej przez Raya Kurtzweila. Wychodząc z przesłanki przyśpieszenia postępu technicznego, sugeruje, że sztuczna inteligencja może objawić się znacznie szybciej, niż prognozuje to wielu naukowców. Przypomina jednak okres połowy XX wieku, kiedy to w nawiązaniu do rozwoju komputerów pojawiły się pomysły, by uczynić maszynę „inteligentną". Prognozowano, że okres od wynalezienia komputerów do pojawienia się SI wyniesie około 20 lat. Jednak, jak świadczy dotychczasowa historia, wraz z rozwojem badań nad sztuczną inteligencją w ośrodkach badawczych USA okazywało się, że te nad wyraz optymistyczne prognozy nie mogą być urzeczywistnione. Dodatkowo Bostrom przywołuje wydarzenie, które zapoczątkowało badania oraz prace nad opracowaniem SI – sześciotygodniowe warsztaty w Dartmouth College pod nazwą Dartmouth Summer Research on Artifical Inteligence w 1956 roku. Był to okres, gdy starano się udowodnić, że sztuczna inteligencja jest w stanie działać w świecie, np. dokonywać obliczeń czy wymyślać dowody logiczne[4]. Okazało się, że wraz z rozwojem komputeryzacji oraz techniki, komputer może pełnić funkcję „psychoterapeuty" czy wykorzystywać elementy, podłączone do niego jak sztuczne ramię, do działania w przestrzeni, w której się znajduje. Bostrom wskazuje, że wraz z rozwojem badań nad SI, jej możliwości stale się rozwijają. Już teraz wiele algorytmów góruje nad człowiekiem w takich grach jak warcaby, szachy, Scrabble[5], czy Azjatycka gra GO[6]. Choć algorytmy te odnoszą sukcesy w swych dyscyplinach, to jednak, jak zauważa autor, są one wyspecjalizowane do działania jedynie w obrębie swojej dziedziny, przez co nie można przypisać im inteligencji. Przeprowadzono jednak ankiety z udziałem ekspertów w zakresie kognitywistki i badań nad sztuczną inteligencją, pytając ich o szacunkową datę zaistnienia maszyny, która dorówna inteligencją człowiekowi. Według tych ekspertów szanse na zaistnienie SI o możliwościach człowieka w 2050 roku będą wynosiły 50%, zaś około 2080 roku aż 90%[7].

[4] N. Bostrom, *Superinteligencja: scenariusze, strategie, zagrożenia*, tłum. D. Konowrocka-Sawa, Wydawnictwo Helion, Gliwice 2016, s. 24.

[5] Tamże, s. 32.

[6] https://qz.com/639952/googles-ai-won-the-game-go-by-defying-millennia-of-basic-human-instinct/ [dostęp: 10.02.2017].

[7] N. Bostrom, *Superinteligencja*, dz. cyt., s. 42.

Drugi rozdział książki zajmuje się sposobami zaistnienie „autonomicznej" SI. Według Bostroma, możliwe tu metody to: rozwój sztucznej inteligencji, emulacja mózgu, poznanie biologiczne, interfejsy mózg-komputer oraz budowa sieci i organizacji. Pomysł oparty o rozwój SI zakłada, że w najbliższym czasie dzięki wzrostowi mocy obliczeniowej oraz udoskonalaniu emulacji mózgu, komputer będzie w stanie wykształcić coś w rodzaju „świadomości", która będzie podobna do świadomości człowieka. Emulacja mózgu opiera się z kolei na zeskanowaniu i odwzorowaniu struktur obliczeniowych ludzkiego mózgu w przystosowanym ku temu urządzeniu, np. komputerze. Dzięki temu urządzenie uzyskałoby „świadomość", ponieważ funkcjonowałby jako dokładna kopia „matematyczna" danej osoby, której mózg został zeskanowany. Biorąc pod uwagę niemalże nieograniczone możliwości przyśpieszenia mocy obliczeniowej danej maszyny, tego typu świadomość prawdopodobnie byłaby w stanie przewyższać intelektualnie „zwykłych" ludzi. Trzecim sposobem na uzyskanie superinteligencji jest „ulepszenie" zdolności myślnych człowieka. Można tego dokonać poprzez eugenikę, ale, jak zauważa Bostrom, budzi to poważne kontrowersje natury moralnej i politycznej. Inna możliwość to wykorzystanie leków poprawiających zdolności mózgu czy manipulacje na poziomie genów w zarodkach lub nawet u dorosłych osób. Inny wariant to wykorzystanie możliwości obecnej techniki – jako syntezy ludzkiego mózgu oraz artefaktów, np. instalowanie chirurgiczne implantów czy procesorów, poprawiających pamięć lub zdolności analityczne. Ostatnią metodą na stworzenie superinteligencji jest zbudowanie ogromnej sieci i organizacji, która w jedną całość łączyłaby możliwości kognitywne pojedynczych ludzi, komputerów oraz innych urządzeń. Jednym z przejawów tego rodzaju działania jest powstanie i wykorzystanie Internetu do wręcz nieograniczonej wymiany informacji oraz idei.

W trzecim rozdziale autor przedstawia możliwe formy, jakie może przybrać superinteligencja. Pierwszą jest superinteligencja szybka, posiadająca możliwości równe człowiekowi, jednakże znacznie szybsza. Drugą jest superinteligencja zbiorowa, która, dzięki temu, że składa się z wielu jednostek, jako całość przewyższa wszelkie inne inteligentne formy życia. Jest to nawiązanie do superinteligencji opartej na sieci i organizacji. Trzecią forma jest superinteligencja jakościowa, która jest równie „szybka" jak ludzki umysł, jednakże, właśnie co do inteligencji, go przewyższa.

W rozdziale czwartym Bostrom analizuje, ile czasu potrzeba, by superinteligencja mogła się pojawić. Zauważa, że przyśpieszenie naukowo-tech-

niczne znacznie zwiększyło możliwości cywilizacyjne, a tempo, w jakim wyłaniają się coraz to nowe możliwości, stale rośnie. W związku z tym, czas do zaistnienia superinteligencji stale się „kurczy"; każdy kolejny etap znacząco przybliża nas do tego momentu.

W piątym rozdziale Bostrom zastanawia się nad tym, czy w świecie pojawi się tylko „jedna" superinteligencja, czy może będzie ich „kilka"? Biorąc pod uwagę, że jest obecnie wiele konkurujących ze sobą ośrodków naukowych, jak też państw, których priorytetem jest przewaga techniczna, jak USA, Rosja czy Chiny, to jest prawdopodobne, że w mniej więcej tym samym czasie może pojawić się kilka różnych superinteligencji – podobnie jak było to z badaniami nad skonstruowaniem bomby atomowej.

Rozdział szósty to analiza teoretycznych możliwości, które może posiąść superinteligencja. Bostrom wymienia to następujące możliwości: potęgowanie własnej inteligencji, zdolność myślenia strategicznego, zdolność manipulowania ludźmi, umiejętność hakowania, możliwość prowadzenia badań technologicznych oraz produktywność gospodarcza. Bostrom rozważa również scenariusze, w ramach których superinteligencja z nieokreślonych pobudek starałaby się przejąć „władzę nad światem".

Rozdział siódmy zawiera analizę potencjalnych pobudek, według których miałaby działać superinteligencja. Takie pobudki to: kierowane instynktem samozachowawczym działanie na rzecz samoprzetrwania, realizacja jakichś nieokreślonych celów społecznych (w przypadku powstania superinteligencji opartej na systemie organizacji społecznej), ciągłe podnoszenie zdolności poznawczych, doskonalenie na poziomie techniki lub wystąpienie u sztucznej inteligencji „egoizmu konsumpcyjnego", tj. pozyskiwanie i gromadzenie przez nią zasobów naturalnych.

Rozdział ósmy jest poświęcony analizie możliwej zagłady ludzkości, która mogłaby stanowić rezultat powstania niezależnej superinteligencji. Bostrom sugeruje, że nawet uzyskanie maksymalnej kontroli nad superinteligencją nie gwarantuje, że nie uzna ona ludzi za „przeszkodę", którą należy usunąć[8]. Dodatkowo, projekty oparte o SI mogą nieść niebezpieczeństwo pojawienia się niespodziewanych błędów, które mogą zmienić sposób działania SI. Dlatego Bostrom analizuje, czy możliwe byłoby określenie celów, których SI nigdy nie może realizować; mogłoby to blokować podejmowanie przez nią samodzielnych decyzji i działania bez nadzoru.

[8] Tamże, s. 174–175.

Rozdział dziewiąty to rozważania na temat kontroli superinteligencji. Bostrom twierdzi, że skoro pojawienie się SI może rodzić egzystencjalne zagrożeniem dla ludzkości, jak najszybciej należy rozpocząć prace nad zminimalizowaniem takiego ryzyka[9]. Można by tego dokonać poprzez coś w rodzaju uwięzienia fizycznego lub informacyjnego, czyli umieszczenie superinteligencji w zamkniętym ośrodku lub stworzenie dla niej odrębnej sieci komunikacji, bez dostępu do Internetu. Innym sposobem mogłoby być stworzenie metody zachęty, czyli skonstruowanie praw, określających działanie superinteligencji; przy czym SI nie mogłaby tych praw złamać. Podobnie do praw Asimova[10]: 1) robot nie może skrzywdzić człowieka, ani też przez zaniechanie działania dopuścić, by człowiek doznał krzywdy; 2) robot musi być posłuszny rozkazom człowieka, chyba że stoją one w sprzeczności z Pierwszym Prawem; 3) robot musi chronić sam siebie, jeśli tylko nie stoi to w sprzeczności z Pierwszym lub Drugim Prawem.

Kolejna możliwość kontroli to stopniowe „upośledzanie" SI, czyli ograniczanie jej możliwości poznawczych i dostępu do informacji, przez co człowiek wciąż posiadałby przewagę i byłby w stanie się bronić. Ostatnia możliwość to metoda wyzwalaczy, które działałyby podobnie do praw robotów, jednakże byłyby nieuświadomione i ukryte dla SI. W przypadku działania niezgodnego z prawami, sztuczna inteligencja uległaby zablokowaniu, a człowiek mógłby odpowiednio zareagować: naprawić błąd, poddać kontroli lub nawet ją „wyłączyć".

W rozdziale dziesiątym autor przedstawia trzy możliwe typy superinteligencji. Pierwszy to *wyrocznia* – celem superinteligencji jest udzielanie odpowiedzi oraz rozwiązywania kwestii, których nie może rozwiązać człowiek, np. pytań filozoficznych. Drugi to *dżin* – urządzenie, które „spełnia życzenia". Taka superinteligencja wykonywałaby skomplikowane logistycznie zadania, np. wykorzystując w tym celu podległe jej roboty. System ten mógłby np. przystosowywać inne planety do ludzkich potrzeb. Trzeci typ to *SI narzędziowa* – superinteligencja to „doskonały" program, który nie posiada świadomości i „wolnej woli", lecz jedynie wykonuje z idealną precyzją określone zadania, np. zarządza globalnym ruchem lotniczym.

Rozdział jedenasty przedstawia skutki, jakie w ludzkim świecie mogłaby wywołać sztuczna inteligencja. Bostrom analizuje możliwe zmiany społeczne i związane z nimi problemy, jak np. zastąpienie pracy fizycznej

[9] Tamże, s. 193.

[10] https://pl.wikipedia.org/wiki/Etyka_robot%C3%B3w#Prawa_robot.C3.B3w_Asimova [dostęp: 10.02.2017].

człowieka przez roboty oraz wzrost bezrobocia, problemy z rozdziałem kapitału wytworzonego przez SI czy wytworzenie singletonu, czyli globalnego państwa zarządzanego przez superinteligencję.

W rozdziale dwunastym i trzynastym Bostrom rozważa pytanie, czy Sztucznej Inteligencji można będzie zaszczepić wartości. Dzięki ewolucji pojawiły się istoty zdolne do poznawania i rozróżniania wartości (czego dowodem jest człowiek[11]). Bostrom sądzi więc, że ten mechanizm można zastosować w ewolucyjnym procesie uczenia wartości sztucznej inteligencji – zaczepiając jej kilka podstawowych prawd, które następnie sama będzie rozwijać. Innym sposobem na zaszczepianie wartości miałoby być wykorzystanie mechanizmu uczenia się. Gdy superinteligencja postępuje w sposób niebudzący wątpliwości moralnych, otrzymuje „nagrodę", co wytwarza w niej nawyk postepowania w sposób „właściwy". Trzecia możliwość to umieszczenie inteligencji w spreparowanym środowisku, w którym przebywanie wymusza „naukę" obowiązujących w nim wartości. Ostatnią metodą mogłoby być najpierw nauczenie SI wartości podstawowych, a następnie stopniowe podnoszenie kierowanych wobec niej wymagań.

Kolejny ważny problem to wybór wartości i zachowań, których chcielibyśmy nauczyć superinteligencję. Bostrom powołuje się na koncepcję Eliezera Yudkowsky'ego nazwaną CEV – coherent extrapolated volition (spójna, ekstrapolowalna wola). Zgodnie z nią, superinteligencja powinna realizować takie cele, które realizowałby człowiek, lecz musi mieć rzeczywiście pewność, że ludzie chcieliby je zrealizować. Jeśli SI nie jest tego pewna lub ma się domyślać, musi zaprzestać działania. Dodatkowo, sztuczna inteligencja powinna posiadać odpowiedni zestaw komponentów w postaci koncepcji celu, teorii decyzji, teorii poznania oraz ratyfikacji.

W rozdziale czternastym autor pyta o to, w jakim kierunku powinniśmy zmierzać: analizuje koncepcje związane ze zróżnicowanym rozwojem technologicznym, preferowaną kolejnością nadejścia, tempem nadchodzących przemian, oraz stara się odpowiedzieć na pytanie, czy powinniśmy dalej prowadzić badania nad emulacją mózgu i sztuczną inteligencją. W ostatnim, bardzo krótkim rozdziale Bostrom wskazuje cele, na których ludzkość powinna się obecnie skupić, by doprowadzić do wytworzenia superinteligencji oraz by ograniczyć ryzyka związane z rozwojem technologicznym.

[11] N. Bostrom, *Superinteligencja*, dz. cyt., s. 274.

Superinteligencja Nicka Bostroma to książka niezwykle ważna i intelektualnie prowokująca. Pozwala poszerzyć wiedzę o problemach sztucznej inteligencji tym osobom, które dopiero zaczynają zajmować się tą problematyką. Dodatkowo, niezwykle bogata bibliografia ułatwia czytelnikom dalsze poszukiwania. Zaletą książki jest również to, że analizując możliwe problemy i zagrożenia, Bostrom bierze pod uwagę różne możliwe formy sztucznej inteligencji, nie spłycając jej do ogólnego „programu", np. superkomputera – co niestety można zauważyć w częstokroć niezbyt głębokich analizach medialnych. Bostrom zwraca uwagę na to, że jest kilka możliwych dróg, na jakich może powstać Superinteligencja. Zauważa również, że rożne mogą być też przyczyny jej ewentualnego „buntu". To sprawia, że analizy Nicka Bostroma stanowią obecnie analiz najbardziej pełne zestawienie zagrożeń związanych z rozwojem SI.

Bibliografia

Bostrom N., Superinteligencja: scenariusze, strategie, zagrożenia, tłum. D. Konowrocka-Sawa, Wydawnictwo Helion, Gliwice 2016.

https://pl.wikipedia.org/wiki/Etyka_robot%C3%B3w#Prawa_robot.C3.B3w_Asimova [dostęp: 10.02.2017].

https://qz.com/639952/googles-ai-won-the-game-go-by-defying-millennia-of-basic-human-instinct/ [dostęp: 10.02.2017].

https://www.theguardian.com/technology/2014/oct/27/elon-musk-artificial-intelligence-ai-biggest-existential-threat [dostęp: 10.02.2017].

http://www.bbc.com/news/technology-30290540 [dostęp: 10.02.2017].

http://www.philosophy.ox.ac.uk/members/other_faculty/bostrom_nick[dostęp: 10.02.2017].

Information about Author:

KAMIL SZYMAŃSKI, PhD student in cognitive science and social communication at the Faculty of Philosophy and Sociology, UMCS, Lublin, PhD student in Philosophy at the Faculty of Philosophy, KUL, Lublin; address for correspondence: Pl. Marii Curie-Skłodowskiej 4, PL 20-031 Lublin; e-mail: szym.kamil@gmail.com