STUDIA METODOLOGICZNE
NR 38 • 2017, 193-203
DOI: 10.14746/sm.2017.38.12

Nick Bostrom, *Superinteligencja. Scenariusze, strategie, zagrożenia*, tłum. Dorota Konowrocka-Sawa, Gliwice, Helion, 2016, 488 s.

PIOTR PRZYBYSZ

Czy należy obawiać się superinteligencji?

ABSTRACT. Should we be afraid of superintelligence? (review of the book by Nick Bostrom *Superintelligence. Paths, Dangers, Strategies*)
A review of Nick Bostrom's book "Superintelligence. Paths, Dangers, Strategies", translated by Dorota Konowrocka-Sawa. Published by Wydawnictwo Helion, Gliwice 2016, p. 488.

KEY WORDS: artificial intelligence, superintelligence, philosophy of informatics, humans vs. robots

1.

Przewidywania i spekulacje na temat koegzystencji ludzkiej i sztucznej inteligencji w mniej lub bardziej odległym świecie przyszłości budzą spore zainteresowanie i wywołują żywe dyskusje. Jednak do niedawna uchodziły one raczej za domenę literatury *science fiction* oraz filmu. Przykładowo w *Golemie XIV* Stanisław Lem przedstawił przewrotną wizję buntu – zbudowanych w celach wojskowych – inteligentnych maszyn cyfrowych, które zamiast zajmować się przeprowadzaniem symulacji konfliktów zbrojnych, lepiej odnajdywały się w roli komputerowej wyroczni wygłaszającej

filozoficzne pouczenia kierowane do ludzkości na temat ludzkiej natury, ewolucji oraz przyszłości Rozumu.

Ostatnio jesteśmy świadkami kolejnej dyskusji – tym razem jednak prowadzonej w obszarze filozofii i refleksji nad nauką – zainicjowanej pytaniem o przyszły rozwój technologii wykorzystujących rozwiązania z obszaru sztucznej inteligencji (SI). Liczne wdrożenia i perspektywa masowego użycia SI – na przykład w sferze technologii internetowych, pojazdów autonomicznych, robotów produkcyjnych i humanoidalnych, w medycynie i nanotechnologii – skłaniają do refleksji nad kierunkiem, tempem zachodzących zmian oraz nad ich krótko- i długoterminowymi skutkami mogącymi mieć wpływ na funkcjonowanie jednostki i społeczeństwa.

Szczególne zainteresowanie wywołuje w tym kontekście pytanie nie tyle o możliwość pojawienia się SI dorównującej poziomem inteligencji człowiekowi (por. problem testu Turinga), ale takiej, która znacznie przewyższałaby ludzką inteligencję (problem tak zwanej superinteligencji). Temu właśnie poświęcona jest wydana w 2016 roku w tłumaczeniu na język polski *Superinteligencja* Nicka Bostroma (oryg. 2014). Autor książki jest filozofem pracującym na uniwersytecie w Oksfordzie, gdzie kieruje placówką naukową Future of Humanity Institute, której celem jest przygotowywanie strategicznych analiz i ocen ryzyka globalnych trendów oraz technologicznych i cywilizacyjnych projektów przyszłości.

Ukazanie się książki Bostroma nadało przyśpieszenia debacie, która tliła się już od jakiegoś czasu i dotyczyła naukowej oraz filozoficznej oceny wiarygodności tych scenariuszy rozwoju technologii przyszłości, które zakładały przejęcie przez SI zarządzania i kontroli nad infrastrukturą bezpieczeństwa, organizacji społecznej, procesem produkcji, inwestycjami giełdowymi czy środkami komunikacji w hipotetycznym społeczeństwie przyszłości. Mapa stanowisk zaprezentowanych w tej dyskusji jest dość rozległa. Przewidywania i prognozy czynione na ten temat mają zwykle postać linearnych (na przykład R. Kurzweil) bądź multilinearnych (N. Bostrom) scenariuszy przyszłego rozwoju technologicznego. Jedne z nich wyrażają wiarę (na przykład N. Bostrom), inne zaś niewiarę (M. Boden, A. Ng) w to, że SI dogoni i prześcignie człowieka pod względem poziomu inteligencji oraz możliwości poznawczych. Niektóre z powyższych scenariuszy są raczej optymi-

styczne (na przykład L. Floridi), a inne bardziej pesymistyczne (S. Hawking, N. Bostrom), gdy chodzi o możliwość pokojowej i harmonijnej koegzystencji ludzkiej i sztucznej superinteligencji.

Warto podkreślić, że dyskusja na temat superinteligencji, jaką proponuje w swojej książce Bostrom, różni się nieco od wcześniejszej o kilkanaście lat burzliwej debaty dotyczącej *singularity* (osobliwości), wywołanej głównie słynną książką Raya Kurzweila. Autor *Nadchodzi osobliwość* rozważał w swojej pracy przede wszystkim futurystyczny scenariusz stopniowego zespalania biologicznego życia i myślenia z technologiami cyfrowymi, co „umożliwiłoby ludziom przekroczenie ograniczeń naszych biologicznych ciał i umysłów" i uzyskanie „władzy nad własnym losem" [por. 2013, s. 24]. Z kolei autor *Superinteligencji* relacje między człowiekiem a sztuczną inteligencją postrzega nieco inaczej – jako pole przyszłej rywalizacji o zasoby i informacje, w której stawką może stać się przetrwania gatunku ludzkiego.

2.

Tematyka książki Bostroma skoncentrowana jest wokół trzech przecinających się osi tematycznych: (1) *możliwych scenariuszy dochodzenia do superinteligencji*, z których każdy opisuje nieco inną, autonomiczną ścieżkę powstania zaawansowanej SI, (2) *szybkości, z jaką zaawansowana SI nabierze mocy obliczeniowej równej superinteligencji*, czyli tego, jak dużo czasu zabierze maszynom inteligentnym osiągnięcie mocy przewyższającej inteligencję człowieka, a także (3) *możliwości ludzkiej kontroli i sterowania tak rozumianą superinteligencją*.

Pierwszemu z wymienionych tematów – tj. różnym scenariuszom wiodącym do powstania wyrafinowanej sztucznej inteligencji – poświęcone są początkowe rozdziały książki. Znajdujemy w nich również charakterystykę różnych postaci, jakie może ona przybrać w przyszłości (por. rozdz. 1–3 i rozdz. 10). I tak, Bostrom wymienia pięć głównych ścieżek, które, według niego, doprowadzić mogą naukowców i inżynierów do skonstruowania zaawansowanej SI. Są to: (1) rozwój maszynowej sztucznej inteligen-

cji, (2) projekty skupione na emulacji ludzkiego mózgu, (3) rozwój i ulepszanie poznania biologicznego, (4) interfejsy mózg–komputer, a także (5) udoskonalenie sieci komunikacji i organizacji łączących ludzi między sobą oraz łączących ich z inteligentnymi artefaktami i robotami (tak zwana „superinteligencja zbiorowa"). Według Bostroma już sam fakt, że istnieje wiele równoległych dróg budowy sztucznej superinteligencji, zwiększa prawdopodobieństwo jej pojawienia się w przyszłości i psychologicznie utwierdza „nas w przekonaniu, że w końcu do niej dotrzemy" (s. 83). Jednak z różnych względów uważa on, że najbardziej obiecującą ścieżką na zbudowanie zaawansowanej SI jest pierwsza z wymienionych dróg, czyli konstruowanie autonomicznych i inteligentnych maszyn (por. przewaga sprzętowa i związana z oprogramowaniem, s. 97–98). Zarazem nie wyklucza jednak realizacji w przyszłości scenariusza hybrydowego, na przykład łączącego udoskonalenia w zakresie inteligencji zbiorowej z sukcesami w konstruowaniu maszyn sterowanych za pomocą SI.

Jak sądzę, pewien niedosyt u czytelnika może budzić charakterystyka docelowej formy SI, czyli tytułowej superinteligencji. W terminach generalnych Bostrom definiuje ją początkowo jako „każdy umysł, który pod względem zdolności poznawczych znacznie przewyższa człowieka w każdej dziedzinie zainteresowań" (s. 45) i wskazuje, że przewaga ta dotyczyć może szybkości przetwarzania informacji (tak zwana superinteligencja szybka), zdolności do agregacji dużej liczby jednostek obliczeniowych (superinteligencja zbiorowa) oraz posiadania dodatkowych możliwości poznawczych obcych obecnemu człowiekowi (superinteligencja jakościowa). Dodatkowo z rozproszonej opisowej charakterystyki pojawiającej na kartach książki można domyślić się kilku innych generalnych cech przypisywanych superinteligencji przyszłości, a mianowicie tego, że będzie ona miała postać (1) inteligencji ogólnej, czyli systemów o charakterze wielozadaniowym, szerokiego zastosowania i, być może, świadomych, (2) będzie obdarzona autonomią, czyli możliwością podejmowania decyzji pozostających poza bieżącą kontrolą programisty lub operatora, a także (3) będzie wyposażona w zdolność samoulepszania, czyli projektowania urządzeń i wprowadzania nowych wersji własnego oprogramowania [por. równ. na przykład Amstrong, 2014].

Widoczny już na pierwszy rzut oka definicyjny minimalizm powyższych określeń i ich porównawczo-relacyjny charakter nie powinny zbytnio dziwić, jeśli się pamięta, że powyższe postulaty znaczeniowe odnoszą się do czegoś, co w zasadzie jeszcze w opisywanej formie nie istnieje i czego jedynym obecnie uchwytnym i miarodajnym punktem odniesienia pozostaje inteligencja współczesnego człowieka.

Nieco bardziej konkretne przewidywania dotyczące możliwego kształtu przyszłej superinteligencji pojawiają się w książce Bostroma w rozdziale 10, gdy autor – tym razem językiem wysoce metaforycznym i obrazowym – proponuje zgrabną i intuicyjną typologię „systemów superinteligentnych". Typologia ta jest interesująca głównie dlatego, że pozwala uwypuklić, w jakiej relacji pozostają one z człowiekiem, jak się z nim komunikują, a także – stopień posiadanej przez nie autonomii. I tak, pierwszym typem superinteligencji mają być tak zwane „wyrocznie", czyli maszyny, których główną funkcją ma być odpowiadanie na pytania człowieka i rozwiązywanie postawionych przez niego w konwersacji z komputerem problemów. Jako drugi rodzaj superinteligentnych systemów wymienione są „dżiny", czyli systemy usłużnie wykonujące praktyczne zadania zlecone przez człowieka, które po wykonaniu rozkazu przerywają działanie „w oczekiwaniu na kolejne polecenie" (s. 219). Trzecim wyróżnionym typem są „suwereni", czyli takie systemy SI, które – jak sama nazwa wskazuje – w największym stopniu obdarzone zostały przez programistów autonomią, możliwością uczenia się oraz automodyfikacji sposobów osiągania zaprogramowanych celów i rozwiązywania problemów.

Bostrom wyróżnia w tym rozdziale jeszcze jeden dodatkowy typ superinteligentnego systemu, którą nazywa „narzędziową SI". Jednak wydaje się, że narzędziowa SI została przez niego przywołana jedynie w celu przedyskutowania pewnego zasadniczego problemu, a mianowicie – czy jest możliwe stworzenie sytemu sztucznej superinteligencji, który cechowałby się „krzepiącą biernością banalnego narzędzia"? (s. 224). W tym ważnym i kluczowym fragmencie książki autor formułuje przekonującą – moim zdaniem – argumentację za tym, że zasadniczo nie będzie możliwe (albo będzie ekstremalnie trudne) zaprojektowanie i wyprodukowanie takiego systemu sztucznej superinteligencji, który byłby układem autono-

micznym, uczącym się, posiadał cechy inteligencji ogólnej, i zarazem pozostawałby jedynie biernym narzędziem i wykonawcą poleceń człowieka. Wywód przedstawiony w tym punkcie przez Bostroma przypomina nieco swoją strukturą znany argument o ciastku, którego nie można zarazem mieć i go zjeść. Sądzę, że wywód ten można też odczytać jako interesujący głos w dyskusji na temat tego, dlaczego tak trudno będzie człowiekowi utrzymać skonstruowaną przez siebie superinteligencję na smyczy racjonalnej, odpowiedzialnej kontroli, oraz jako uzasadnienie pesymizmu w tej sprawie.

3.

Kolejnym kluczowym tematem *Superinteligencji* jest dalsza ewolucja SI, po tym, jak osiągnie ona moc obliczeniową równą sile inteligencji człowieka (por. rozdz. 4–8). Jak już wcześniej wspomniałem, Bostroma nie interesuje klasyczny problem SI – formułowany najczęściej pod postacią „testu Turinga", w którym chodzi o rozstrzygnięcie, czy maszyna dorównała już swoją inteligencją człowiekowi – ale raczej to, co będzie po tym, jak człowiek zbuduje wreszcie taką sztuczną inteligencję zdolną prześcignąć jego samego „pod względem ogólnej zdolności rozumowania" (s. 101).

W tej najciekawszej, ale i najbardziej kontrowersyjnej, moim zdaniem, partii książki, poświęconej dynamice dochodzenia do sztucznej superinteligencji, jej autor stawia przykładowo pytanie, jak szybkie będzie tempo odejścia SI od poziomu inteligencji człowieka (tzw. problem odejścia) i po jakim czasie osiągnie ona zaawansowanie charakterystyczne dla superinteligencji (por. rozdz. 4). Analiza trzech różnych możliwych wariantów takiego odejścia – tj. odejścia wolnego, umiarkowanego i szybkiego – ma przekonać czytelnika, że przyrost potencjału obliczeniowego i intelektualnego zaawansowanej SI będzie dramatycznie szybki, czyli będzie miała miejsce tzw. „eksplozja inteligencji".

Jeszcze inne kluczowe pytanie, jakie stawia Bostrom, odnośnie dynamiki rozwoju superinteligencji, dotyczy tego, czy w trakcie własnej ewolucji przybierze ona postać „singletonu", czyli pojedynczego systemu, który uzyska „przewagę strategiczną" i podporządkuje sobie wszystkie pozostałe

inteligentne projekty (por. rozdz. 5). Ze względu na przypisanie superinteligencji szerokich uzdolnień w zakresie potęgowania własnej mocy, budowania przez nią samą coraz doskonalszych wersji inteligentnych maszyn, zdolności myślenia strategicznego i prognozowania, umiejętności manipulowania społecznego i perswazji czy hakowania (tj. znajdywania i wykorzystywania luk w systemach bezpieczeństwa, por. s. 143) – Bostrom pokazuje, że istnieje możliwość spełnienia się „czarnego scenariusza" w sprawie przyszłej koegzystencji człowieka i sztucznej inteligencji. Możliwe jest mianowicie podporządkowanie świata ludzi światu maszyn.

W interesującym rozdziale 7 autor zastanawia się przykładowo nad „pobudkami superinteligencji" i utrzymuje, że, wbrew obecnym naszym przekonaniom o niemożliwości przypisania maszynom cyfrowym ostatecznych „pobudek i motywów", możliwe jest określenie zbioru pobudek o charakterze instrumentalnym i związanych z zaprogramowanym celem, którymi mogłaby kierować się superinteligencja. Takimi pobudkami i motywami mogłyby być: odpowiednik instynktu samozachowawczego (na przykład superinteligentnej maszynie może „zależeć instrumentalnie" na przetrwaniu aż do momentu wykonania zadania), niezmienność zaprogramowanego celu, reguła ciągłego podnoszenia poziomu zdolności poznawczych, zasada perfekcji technologicznej czy zasada pozyskiwania zasobów.

Skrajny i karykaturalny przykład działania inteligentnej maszyny kierującej się sztywno powyższymi regułami przedstawił Bostrom w postaci eksperymentu myślowego. W jego ramach mamy wyobrazić sobie superinteligentną maszynę przeznaczoną do produkcji maksymalnej liczby spinaczy do papieru. Taka idealna superinteligentna maszyna po wyczerpaniu dostępnych jej zasobów i surowcowych nadających się do produkcji dążyłaby dalej za wszelką cenę do realizacji powierzonego jej zadania, i w tym celu zaczęłaby przerabiać na spinacze każdy atom fizycznego świata napotkany na swej drodze, w tym również i ludzi.

Przywołany eksperyment myślowy ilustrujący ideę maksymalizacji instrumentalnego celu superinteligntnej maszyny dość łatwo jest konceptualnie zneutralizować lub osłabić, choćby poprzez uwzględnienie dodatkowych warunków i ograniczeń, na przykład znanych powszechnie trzech „praw robotów" postulowanych przez Isaaca Asimova. Tego samego nie da się już,

niestety, tak samo łatwo osiągnąć odnośnie całej – złożonej z wielu kroków i wielopoziomowej – argumentacji opisującej hipotetyczną dynamikę rozwoju sztucznej superinteligencji. Kontrargument w postaci pokazania, że wiarygodność apokaliptycznego rozumowania Bostroma jest wątpliwa, gdyż opiera się na długim i niepewnym łańcuchu powiązanych wielopiętrowo ze sobą okresów warunkowych, jest poprawny, lecz nie wystarcza do całkowitego zakwestionowania możliwości wystąpienia tej jednej hipotetycznej ścieżki rozwoju wydarzeń. Jak ujął to jeden z krytyków: wprawdzie opisywane w tym scenariuszu zagrożenie jest „skrajnie mało prawdopodobne" to jednak nie można go całkowicie wykluczyć [za: Shermer, 2017, s. 74].

4.

W obliczu przedstawionej przez siebie pesymistycznej wizji ewolucji SI Bostrom zastanawia się nad możliwością przeciwdziałania hipotetycznemu splotowi wydarzeń wiodącemu do realizacji czarnego scenariusza w rozwoju technologicznym. Jest to trzecia oś tematyczna, która zaprząta uwagę autora w ostatniej części książki, gdzie szkicuje on dwa różne scenariusze (por. rozdz. 9–14). Pierwszy z tych scenariuszy przedstawiony został w rozdziale dziesiątym – tym samym, w którym pojawia się odróżnienie wyroczni, dżinów i suwerenów. Autor rozważa tam pomysł kontroli i ograniczenia możliwości superinteligencji poprzez fizyczne lub informatyczne jej „uwięzienie" (przypadek wyroczni) lub poprzez „udomowienie" (jak w przypadku dżina).

Jednak to drugi z proponowanych scenariuszy traktuje autor jako bardziej obiecujący sposób zapanowania nad superinteligencją. Chodzi mianowicie o próbę „zaszczepienia" superinteligencji określonego systemu wartości ostatecznych (rozdz. 12). Prace w tym kierunku powinny odbywać się również wielościeżkowo: między innymi poprzez projekty zmierzające do zaprogramowania z góry maszyny w odpowiedni sposób, albo na przykład poprzez prace nad wszczepieniem jej kilku podstawowych prawd, które następnie inteligentna maszyna (tak zwana „zalążkowa SI") mogłaby pielęgnować i rozwijać w środowisku dzięki mechanizmowi

uczenia się ze wzmacnianiem. To jednak nie kończy problemu, gdyż jeśli nawet takie zaszczepianie wartości okazałoby się technicznie wykonalne, to problematyczne pozostaje ciągle, który z systemów wartości moralnych należałoby zaszczepić inteligentnym maszynom. W tym kontekście Bostrom szczególnie wiele uwagi poświęca propozycji Eliezera Yudkowsky'ego, aby wyposażyć zalążkową SI w taki spójny system wartości, który chcieliby realizować w swoim życiu bezstronni i wolni od uprzedzeń ludzie (nazywa go za Yudkowskim doktryną „spójnej ekstrapolowanej woli", por. Yudkowsky, 2004). Tak zaprogramowana sztuczna superinteligencja mogłaby w sprzyjających okolicznościach nabrać cech podmiotu moralnego, choć technicznym warunkiem realizacji tego projektu byłoby wyposażenie jej w trudne dziś do zaprojektowania moduły kierowania się celami ostatecznymi, teorii decyzji, złożonej teorii poznania oraz w gotowość do dynamicznego uzgadniania jej własnych celów z celami ludzi.

5.

W podsumowaniu niniejszego omówienia warto zwrócić uwagę na kilka spraw. Po pierwsze, można chyba zaryzykować stwierdzenie, że analizy i rozważania zaprezentowane w *Superinteligencji* wyznaczają granice nowego, niezagospodarowanego jeszcze pola badań naukowych i filozoficznych, na które coraz częściej zapuszczać się będą badacze różnych profesji, w ten lub inny sposób powiązani z rozwijaniem technologii SI i z refleksją nad jej rozwojem. Powodów tego jest kilka, a najważniejszy jest chyba taki, że wprowadzanie w życie technologii wykorzystujących SI przestało być wyłączną domeną i specjalnością samych programistów, a stało się „problemem egzystencjalnym" dla dużej części ludzi zamieszkujących obecnie naszą planetę, gdyż wybory dokonywane w tym obszarze „mogą potencjalnie wpłynąć na całe nasze przyszłe życie" [Tegmark, 2017, s. 36]. Po drugie, ujęciu zaprezentowanemu przez Nicka Bostroma można wprawdzie zarzucić pewien rodzaj stronniczego pesymizmu w sprawie przyszłych relacji człowieka z budowaną przez niego sztuczną inteligencją, lecz w żadnym razie nie wydaje się, aby ten pogląd był wyrazem antytechnologicznych uprze-

dzeń. Jego stanowisko jest dość dobrze uzasadnione, racjonalnie uargumentowane, obudowane dużą ilością zastrzeżeń oraz wzbogacone dyskusją ograniczeń własnej koncepcji. Źródeł jego podejścia należy upatrywać raczej w przeświadczeniu, że negatywne i pesymistyczne scenariusze przyszłości wymagają większej uwagi i dogłębniejszej analizy niż scenariusze pozytywne i optymistyczne. Po trzecie, nie da się ukryć, że „wąskim gardłem" rozumowań zaprezentowanych przez Bostroma w omawianej książce pozostają spekulatywne prognozy i przewidywania dotyczące przyszłego rozwoju technologicznego. Z punktu widzenia metodologii badań rozwoju historycznego podejściu takiemu jak to, zaprezentowane w jego książce, łatwo jest postawić zarzut braku wiedzy na temat warunków przyszłego rozwoju oraz posługiwania się pustymi proroctwami [Popper, 1989, s. 71–76]. W obronie Bostroma należy zauważyć, że autor *Superinteligencji* nie zakłada linearnego przebiegu rozwoju dziejów i nie projektuje jego dalszych etapów na podstawie znajomości praw historii. Przywołuje on raczej odmienną, bo multilinearną, wizję rozwoju historycznego oraz metodę polegającą na analizie trendów i zestawianiu różnych wariantów i scenariuszy *możliwego* przebiegu historycznego. Po czwarte wreszcie, nawet jeśli przewidywania Bostroma dotyczące powstania oraz dalszego rozwoju superinteligencji okażą się nietrafione, to nie przekreśla to zarazem intelektualnego pożytku płynącego z lektury jego książki. Filozoficzne prace poświęcone zagadnieniom SI, koniec końców, dotykają problemów odnoszących się do naszej własnej biologicznej inteligencji i pozwalają lepiej zrozumieć naturę i ograniczenia naszych własnych umysłów.

Bibliografia

Amstrong S., (2014), *Smarther Than Us. The Rise of Machine Intelligence*, Berkeley, MIRI.

Kurzweil R., (2013), *Nadchodzi osobliwość. Kiedy człowiek przekroczy granice biologii*, Warszawa, Kurhaus.

Popper K.R., (1989), *Nędza historycyzmu*, Warszawa, Krąg.

Shermer M., (2017), „Apokalipsa AI. Sztuczna inteligencja jako zagrożenie egzystencjalne", *Świat Nauki*, 4 (308), s. 74.

Tegmark M., (2017), *Life 3.0. Being Human in the Age of Artificial Intelligence*, New York, Random House.
Yudkowsky E., (2004), *Coherent Extrapolated Volition, San Francisco*, The Singularity Institute.

Piotr Przybysz
Instytut Filozofii UAM
ul Szamarzewskiego 89 C
60-568 Poznań
e-mail: przybysz@amu.edu.pl