DOCTRINA
Studia Społeczno-Polityczne 20/2023
PL ISSN 1730-0274

Cezary Kalita
ORCID: 0000-0002-6019-0606
cezary.kalita@uws.edu.pl

Uniwersytet w Siedlcach
Wydział Nauk Społecznych

Diagnoza Nicka Bostroma rozwoju Superinteligencji a zagrożenia bezpieczeństwa integralności rodzaju ludzkiego

Nick Bostrom, *Superinteligencja. Scenariusze, strategie, zagrożenia,* tłum. Dorota Konowrocka-Sawa, Wydawnictwo Helion, Gliwice 2021, ss. 488

DOI: 10.34739/doc.2023.20.19

Problem ze sztuczną inteligencją przechodzi z fazy teoretycznej w praktyczną. Oznacza to, że jeszcze do niedawna większość prognoz dotyczących rozwoju SI przesuwała jej pojawienie się na okres około dwóch dekad. Dwadzieścia lat to na tyle bezpieczny czas przywidywania przeszłości, że futurolog nie ponosił żadnych konsekwencji swoich ewentualnie nietrafionych prognoz. Nikt nikogo nie rozlicza z twierdzeń z odległej przeszłości, a 20 lat to okres jednego pokolenia, jednego cyklu ludzkiej przemienności. Nowe pokolenie nie ma pretensji co do nietrafionych diagnoz swoich poprzedników, ponieważ ma własne diagnozy w horyzoncie najbliższych dekad.

W przypadku książki Nicka Bostroma *Superintelligence: Paths, Dangers, Strategies* z 2014 (pol. wyd.: *Superinteligencja. Scenariusze, strategie, zagrożenia* – 2016, 2021)[1] jesteśmy w połowie bezpiecznego okresu przewidywań, czyli po pierwszej dekadzie od jej napisania. A zmiany, jakie dokonują się aktualnie w kwestii sztucznej inteligencji, znacznie przyśpieszyły. Nie tylko funkcjonuje

[1] N. Bostrom, *Superinteligencja. Scenariusze, strategie, zagrożenia*, tłum. D. Konowrocka-Sawa, Gliwice 2021, s. 488.

ChatGPT–4, ale różne jego mutacje, nad którymi pracują zarówno naukowcy, jak i najbogatsze firmy, niekoniecznie kojarzone z branżą IT. W ostatnim czasie opracowały własne modele sztucznej inteligencji (AI) do generowania języka naturalnego, które konkurują z ChatGPT OpenAI. Oto kilka z nich:

1. **Transformer** od Google: jest to model, na którym opartych jest wiele nowoczesnych systemów generowania języka, w tym GPT-3. Google używa wersji tego modelu w swoim tłumaczu Google Translate.
2. **BERT (Bidirectional Encoder Representations from Transformers)**: jest to inny model opracowany przez Google, który przekształcił podejście do zrozumienia języka naturalnego, koncentrując się na analizie kontekstu obu stron tokenu, a nie tylko na jednym kierunku.
3. **GloVe (Global Vectors for Word Representation)**: jest to model opracowany przez naukowców z Uniwersytetu Stanforda. Choć jest starszy niż BERT czy GPT, nadal jest używany w niektórych zastosowaniach.
4. **Turing-NLG** od Microsoft: jest to model języka naturalnego, który został opracowany przez Microsoft i jest porównywalny z GPT-3 pod względem wielkości i złożoności.
5. **DialoGPT**: inny model OpenAI, stworzony specjalnie do generowania konwersacji. Jest on mniej skupiony na szerokim generowaniu tekstu, a bardziej na stworzeniu wiarygodnych dialogów.
6. **Chatbots** takie jak Mitsuku, Cleverbot czy Replika: chociaż nie są tak zaawansowane pod względem modelowania języka, są one konkurującymi systemami, które generują odpowiedzi w kontekście rozmów.
7. **Rasa**: open source'owa platforma do tworzenia chatbotów i asystentów głosowych, która pozwala na pełną kontrolę nad danymi i personalizację interakcji.

Wszystkie te modele i systemy konkurują w różnym stopniu z ChatGPT w zakresie generowania języka naturalnego i prowadzenia rozmów.

Czy rozważania zaproponowane przez Bostroma straciły na aktualności, czy dalej mogą być inspirujące, mimo gwałtownego przyspieszenia osiągnięć na polu sztucznej inteligencji? A może pewne oddalenie w czasie i znajomość faktów, które były tylko pro-

gnozowane i się wydarzyły (ewentualnie się nie wydarzyły), daje inne spojrzenie? Być może to ten moment, w którym możemy lepiej weryfikować idee, ale też dostrzegać realne procesy i ich dalszy kierunek rozwojowy. Prognoza nieokreślonej przyszłości nas nie przeraża (nawet jeśli jest przerażająca), ale realne dostrzeganie symptomów spełniającej się prognozy może być już bardzo przerażające. Dlatego warto przeanalizować kilka pomysłów Bostroma odnośnie scenariuszy rozwoju SI, strategii jej kontroli oraz ewentualnych zagrożeń dla trwania ludzkości w stanie, jaki znamy, przy założeniu jej integralności[2].

Sam Nick Bostrom to profesor na Uniwersytecie Oksfordzkim i renomowany filozof skupiający się na przyszłości człowieka i technologii, który z dużą starannością i przekonaniem prowadzi czytelnika przez skomplikowane scenariusze i modele potencjalnej superinteligencji, definiowanej jako intelekt znacznie przewyższający zdolności poznawcze ludzi we wszystkich praktycznie ważnych dziedzinach. Jego książka jest uważana za przełomowe dzieło, które zasadnie zyskało sobie reputację jednego z najważniejszych opracowań na temat sztucznej inteligencji (AI) i jej potencjalnego wpływu na przyszłość ludzkości[3].

Jak zauważa Max Tegmark, „świętym Graalem badań nad AI jest zbudowanie «ogólnej AI» (lepiej znana jako *silna sztuczna inteligencja*, AGI), która ma maksymalnie szeroki zakres, jest zdolna do osiągniecia praktycznie każdego celu, łącznie z nauką”[4]. Czyli ogólna sztuczna inteligencja (AGI – *Artificial General Intelligence*[5]) to rodzaj zaawansowanej sztucznej inteligencji, która ma zdolność do zrozumienia, nauki i zastosowania wiedzy na poziomie porówny-

[2] Vide: H. Jonas, *Zasada odpowiedzialności. Etyka dla cywilizacji technologicznej*, tłum. M. Klimowicz, Kraków 1996.

[3] S. Russell, *Human Compatible: Artificial Intelligence and the Problem of Control*, London 2019; K. Sotala, R.V. Yampolskiy, *Responses to catastrophic AGI risk: a survey*, “Physica Scripta” 2015, no. 90 (1); V.C. Müller, N. Bostrom, *Future Progress in Artificial Intelligence: A Survey of Expert Opinion*, [w:] *Fundamental Issues of Artificial Intelligence*, red. V.C. Müller, Oxford 2016, s. 555–572.

[4] M. Tegmark, *Życie 3.0. Człowiek w erze sztucznej inteligencji*, tłum. T. Krzysztoń, Warszawa 2019, s. 74.

[5] Pojęcie to zostało spopularyzowane przez Shane’a Legga, Marka Gubruda i Bena Goertzela. Marcus Hutter jest teoretykiem, który przyczynił się do rozwoju teorii uniwersalnej sztucznej inteligencji. Jego prace, takie jak książka *Universal Artificial Intelligence. Sequential Decisions Based on Algorithmic Probability*, Berlin 2005, są istotnym wkładem w dziedzinę AGI.

walnym z umiejętnościami człowieka. AGI jest w stanie rozwiązywać problemy, podejmować decyzje, a także uczyć się i adaptować do nowych sytuacji bez konieczności wcześniejszego szkolenia lub programowania. Różni się od innych form sztucznej inteligencji, takich jak „wąska sztuczna inteligencja" (ANI – *Artificial Narrow Intelligence*), która jest ograniczona do wykonywania konkretnych zadań, takich jak rozpoznawanie mowy czy zalecanie produktów na podstawie historii zakupów.

Czy jednak rzeczywiście mamy do czynienia z gwałtownym przyśpieszeniem osiągnięć w zakresie sztucznej inteligencji – przejściem z ANI do AGI? Może jest to tylko zjawisko pozorne, co w jakimś stopniu przewiduje sam Bostrom? „Sztuczna inteligencja może dokonać pozornie raptownego skoku intelektualnego wyłącznie w rezultacie antropomorfizacji – ludzkiej skłonności do wyobrażania sobie «wioskowego głupka» i «Einsteina» jako dwóch przeciwnych krańców spektrum inteligencji, a nie dwóch niemal nierozróżnialnych punktów na skali umysłów jako takich"[6]. Jako ludzie dostrzegamy olbrzymie różnice między sobą – jesteśmy wyczulenie na dyferencję i ją traktujemy jako typową cechę ludzką. Jednak jeśli istnieje rzeczywista przepaść miedzy nami a szympansem, myszą i muszką owocówką, to już rozpatrywana wewnątrzgatunkowo jest ona praktycznie nieistotna.

Umysł Einsteina i przysłowiowego wioskowego głupka strukturalnie się nie różną. Szympans nie stanie się Einsteinem ani nawet „wioskowym głupkiem", podczas gdy ten ostatni może stać się wybitnym fizykiem kwantowym (tym samym nawet pokona poziomem wiedzy Einsteina, który w wielu twierdzeniach dotyczących fizyki kwantowej się mylił). Stąd istotna zmiana (pojawienie się ChatGPT) nie jest radykalnym skokiem skali, jest istotnym skokiem poziomu ludzkiego. To, że coś przechodzi test Turinga, nie oznacza, że jest w istotny sposób różne od algorytmu, który prawie przechodzi ten test. Przykładamy nadmierną wartość do zwycięzców i chcemy podkreślać ich wyjątkowość[7]. Najlepszy aktualnie maratończyk zbliża się do przekroczenia bariery pokonania dystansu w czasie 2 godzin, podczas gdy autor tego artykuły ma problem

[6] N. Bostrom, *Superinteligencja. Scenariusze...*, s. 112.
[7] Ibidem, s. 47.

z pokonaniem bariery 4 godzin, to i tak różnica nie jest w skali gatunku ludzkiego istotna.

Mimo że przeceniamy szympansy, co pewnie wynika z faktu bliskości ewolucyjnej, to i tak nie mają one szans na udoskonalenie swojej natury. Większe szanse posiadają zwierzęta domowe, chociażby psy, które dzięki udomowieniu wykazują się większą inteligencją od szympansów[8]. Szukaliśmy przemian i rozwoju inteligencji silnie skorelowanej ze zmianą struktury biologicznej mózgu. A może zmiana będzie wynikiem zmiany paradygmatu myślenia o inteligencji – może wyjście jest jak zwykle z boku, a nie tam, gdzie się go najbardziej spodziewamy. (Bostrom rozważa też udomowienie superinteligencji, co wydaje się bardzo obiecującą koncepcją – chodzi o to, aby SI nie chciała, jak pies u dobrego opiekuna, uciec i uniezależnić od niego). Zdolność uczenia się i podążanie za gestem posiadają tylko ludzi i psy, to właśnie one są nam najbliższe społecznie, chociaż oddalone biologicznie.

Podobnie jak w poszukiwaniu najbliższego nam przodka zaczynamy od pokrewieństwa genetycznego, tak też pierwszym sposobem uzyskania sztucznej inteligencji będzie naśladowanie procesów biologicznych. Rozpoczynamy zatem od problemu emulacji mózgu, czyli jego transferu jako metody najprostszej, za którą przemawia naśladownictwo[9]. Czy można odwzorować ludzki mózg? Bostrom wskazuje na dwa podejścia: syntetyczne i neuromorficzne. Mamy możliwość dzisiaj poruszania się w powietrzu jak ptaki, czyli potrafimy uzyskiwać efekt unoszenia się maszyn cięższych od powietrza. Jednak nasze samoloty nie machają skrzydłami jak ptaki i raczej w niewielkim stopniu je naśladują. Stąd – przez analogię – czy aby osiągnąć inteligencję dorównującą ludzkim możliwościom, mamy naśladować mózg (podejście neuromorficzne), czy też poszukać innych, syntetycznych rozwiązań?

Możemy doskonalić nasz mózg. Odbywałoby się to przy okazji jego coraz głębszego poznawania. Może dokonywać drobnych poprawek, tak aby rozszerzać możliwości naszego mózgu[10]. Już teraz go stymulujemy poprzez różnego typu substancje, które zmieniają

[8] Vide: B. Hare, V. Woods, *Przetrwają najżyczliwsi. Jak ewolucja wyjaśnia istotę człowieka?*, tłum. K. Kalinowski, Kraków 2022.
[9] N. Bostrom, *Superinteligencja. Scenariusze...*, s. 56–64.
[10] Ibidem, s. 96–100.

jego wydajność – od poprawy samopoczucia (kawa, tytoń) po zmiany sposobów myślenia (leki psychotropowe). Są to jednak działanie ewolucyjne, a zatem bardzo powolne. Proces eugeniki pozytywnej cały czas następuje, chociaż jest bardzo umiarkowany jeśli chodzi o przyrost wydajności naszych mózgów. Można przyspieszyć farmakologicznie ten proces, a w zasadzie on stopniowo się odbywa, niemniej jest to obciążone dylematami moralnymi. Czy jest to jednak wystarczające ograniczenie? Można np. replikować plemniki (komórki macierzyste) i na nich dokonywać ewolucji, co radykalnie skróciłoby proces doskonalenia ludzi. Ale i w tym przypadku na efekty trzeba czekać pokolenie (znowu powraca dwadzieścia lat), aby zweryfikować osiągnięte rezultaty, gdy badany osobnik uzyska pełną dojrzałość. Ludzkość wydaje się w tym względzie wielce niecierpliwa. Jednak można sobie wyobrazić, że istnieje jakieś przykładowe społeczeństwo azjatyckie, które ma kulturową zdolność do realizacji celów długookresowych. Wie ono, że każdy proces trzeba zainicjować i cierpliwie czekać na rezultaty oraz ewentualnie dalej je koordynować – dwadzieścia lat to nie aż taki długi okres, a on przecież z całą pewnością nadejdzie.

Innym bardzo ważnym problemem poruszonym przez Bostroma jest zagadnienie instytucjonalnego nadzoru nad sztuczną inteligencją[11]. SI oznacza władzę, a zatem czy jej rozwój i wykorzystanie powinno (musi być) nienadzorowane przez instytucje polityczne. W przypadku broni atomowej prace nad nią od początku zostały zorganizowane przez państwo – projekt Manhattan. Nadzór państwa podlegał najwyższemu klazulowaniu (ściśle tajne). Jednak uznając USA za kraj demokratyczny, ta kontrola nad rozwojem broni atomowej była pośrednio nadzorowana przez wyborców. Chociaż sam projekt był ściśle tajny, a zatem żaden z wyborców nie miał nad nim realnej kontroli, to jednak sprawował taki nadzór pośrednio, zarówno przez prezydenta, jak i sam akt kadencyjnych wyborów. Skutki programu zostały również ujawnione światowej opinii publicznej. Czy jednak wywiady interesują się rozwojem AI? A może dojdzie do jej nacjonalizacji SI – państwo wywłaszczy prywatnych jej posiadaczy. Cały zatem problem, tak jak w życiu spo-

[11] Ibidem, s. 213.

łecznym, sprowadza się do kontroli. A w zasadzie do dylematu kontroli tych, którzy kontrolują – kto nadzoruje nadzorców?

Czy ktoś (państwo, instytucja prywatna) może zmonopolizować dostęp do sztucznej inteligencji? Nick Bostrom posługuje się pojęciem „singleton", które odnosi się do hipotetycznego scenariusza, w którym ludzkość osiąga zaawansowany poziom technologiczny i tworzy jednostkę centralną, władającą ogromną siłą i kontrolującą globalne sprawy[12]. Singleton to forma globalnej władzy, w której jedno podmiotowe istnienie, takie jak superinteligencja lub kolektyw umysłów, obejmuje kontrolę nad decyzjami i działaniami na skalę globalną. Singleton mógłby mieć zdolność do monitorowania i wpływania na wszystkie aspekty życia ludzkiego. Singleton w tym kontekście jest często rozważany jako potencjalny rezultat rozwoju sztucznej inteligencji, gdzie dominujące superinteligentne systemy mogą wyprzedzić ludzkie zdolności i narzucić swoją kontrolę nad resztą społeczeństwa.

Warto zaznaczyć, że singleton w rozumieniu Bostroma to hipotetyczny scenariusz, a nie obecna rzeczywistość. To koncepcja, którą Bostrom analizuje w kontekście potencjalnych implikacji rozwoju technologicznego na przyszłość ludzkości. Kiedy (czy) taki singleton może się pojawić? Pojawienia się singletonu jako jedynej bezkonkurencyjnej superinteligencji może być wynikiem efektu pierwszeństwa. Jeśli jeden projekt rozwoju superinteligencji osiągnie przewagę, może zdominować i ograniczyć rozwój innych projektów. Gdy już dojdzie do uzyskania singletonu, to żadna władza nie będzie zainteresowana złamaniem jego monopolu, a wręcz przeciwnie, zwiększeniem przewagi nad konkurencją. Jednak trudno jest uwierzyć, biorąc pod uwagę konkurencje państw narodowych, aby one odpuściły wyścig zbrojeń AI. Szczególnie po pandemii, która wskazała na znaczenie relacji online i Internetu czy szerzej – rzeczywistości cyfrowej, rezygnacja z konkurowania na tym polu nie wchodzi w grę[13].

Bostrom mocno uprawdopodobnia tezę singletonu i uzasadnia jego technologiczną ucieczkę, niemal tempem wykładniczym, w bardzo krótkim czasie. Czy jednak inni nie będą w stanie skorzy-

[12] Ibidem, s. 134–138.
[13] K. Schwab, Th. Malleret, *Covid-19: The Great Reset*, Cologny 2020.

stać z efektu opóźnienia[14]. Brak zaufania w stosunkach międzynarodowych, pojawienie się dużych graczy technologicznych, waga systemów regionalnych wymusi powstanie nie tyle singletonu, co może skutkować pojawieniem się multitonów. Nie tylko chodzi o wykorzystanie wielu singletonów, co już tworzy multiton, ale bardziej zróżnicowanej inteligencji (oczywiście z połączenia różnych singeltonów w końcu musi powstać multiton).

Cały czas, rozważając kwestię AI, powielamy błąd antropomorfizacji. Ciągle myślimy tak, jakby istniała jedna inteligencja (stąd szczególne podkreślanie przez Bostroma zagrożenia monopolizacji superinteligencji w postaci singletonu[15]). Już nawet na poziomie definicyjnym to *implicite* zakładamy. A może istnieje wiele rodzajów inteligencji? To tak, jakbyśmy musieli pomyśleć, że istnieją inne stworzenia (organiczne lub nieorganiczne), które myślą (albo wykonują proces algorytmiczny w języku jakiegoś typu algebry lub geometrii lub synestezji).

Szczególnie obiecująca wydaje się w kwestii SI synestezja jako neurologiczne zjawisko, w którym bodźce jednego zmysłu wywołują skojarzone doznania w innych zmysłach. Nie znamy jeszcze modelu (oprócz matematycznego czy lingwistycznego), w jakim inna inteligencja może się objawiać. Wspomniana synestezja jest takim przejawem. To tzw. sawanci (zespół sawanta) wskazują na inny gatunek poznawczy, którego wydajność epistemologiczna jest bardzo wysoka. Zatem teza Bostroma o singletonie – z punktu ludzkiego (antropomorfizacja) – wydaje się jedyną możliwą, ale z punktu nawet sztucznej inteligencji mało prawdopodobną. Muszą istnieć inne sposoby poznawcze, których z różnych powodów nie rozwijamy – chociażby myślenie obrazami, a nie tylko za pomocą znaków symboliki języków naturalnych czy sztucznych, jak matematyka.

Pierwszy singleton będzie z faktu swojej jedyności najdoskonalszy (przypomina to ontologiczny dowód na istnienie Boga św. Anzelma). Czy jednak doskonałość może się różnicować? Różnicowanie zakłada pojęcie rozwoju, a przynajmniej zmiany. Jeśli tak, to dojdzie do dyferencji. Bóg musi być niezmienny, aby nie wytworzył konkurencji czy chociażby tylko sprzeczności. Doskonałość to tylko

[14] D. S. Landes, *The Wealth and Poverty of Nations. Why Some Are So Rich and Some So Poor*, New York 1998.
[15] N. Bostrom, *Superinteligencja. Scenariusze...*, s. 138.

pożądany stan czy pragnienie antropomorficznego i skończonego myślenia. Proces jeśli trwa w czasie, musi się różnicować. Nie da się utrzymać monoteizmu inteligencji. Bogowie inteligencji muszą zacząć ze sobą konkurować. To może oznaczać, że jednak któryś zdobędzie dominującą pozycję – jak najlepszy szachista w świecie – ale jak długo ją utrzyma, zwłaszcza gdy gra w szachy wyda się działalnością mało atrakcyjną?

Czy sztuczna inteligencja zawładnie światem? Ten scenariusz nie tyle jest realny, co interesujący poznawczo. Jednak pytanie powinno brzmieć: jak sztuczna inteligencja zawładnie światem? Są dwa możliwe scenariusze. Po pierwsze, ten scenariusz rozpatruje Bostrom, AI może wykorzystać człowieka do stworzenia sobie warunków dominacji. Polegałoby to na hakowaniu ludzi metodami służb specjalnych, tj. albo przekupstwem (mogłaby zaoferować niewyczerpalne zasoby) lub ideologią (zaangażować quasi-religijnych zmotywowanych ideologią wiedzy czy świadomości wyznawców)[16]. W wersji umiarkowanej sztuczna inteligencja mogłaby zaoferować nadzór nad sobą samą na takiej zasadzie, że grupa osób realizowałby jej interesy związane z kontrolą społeczną. Czyli SI powołałaby klasę zarządczą, składającą się z wyselekcjonowanych i uprzywilejowanych funkcjonariuszy, stanowiących ułamek populacji. Ten przypadek opisuje *explicite* Janusz Zajdel w książce *Limes inferior*[17].

Drugi scenariusz polega na tym, że to ludzie przejęliby sztuczną inteligencję i mieli monopol na jej udostępnianie innym. Ludzie chcieliby wykorzystywać AI przeciwko innym ludziom tak, aby ich kontrolować. Ten scenariusz zakłada, że to zło zawsze wyprzedza technologię – ludzie ją tylko wykorzystują do dobrych lub złych celów. Decyduje ten, kto ma do niej nieograniczony dostęp. Czy jednak te scenariusze inżynierii społecznej różnią się od siebie? Efekt nadzorowania ludzi przez sztuczną inteligencję z wykorzystaniem przedstawicieli (elity) naszego gatunku jest taki sam. Wszyst-

[16] Oba sposoby opisał w 1951 roku Isaac Asimov w książce o *Fundacji* i jej sposobach ekspansji politycznej, która wykorzystuje swoją przewagę w wiedzy naukowej nad innymi społecznościami. Vide: I. Asimov, *Foundation*, New York 1951.

[17] Wprawdzie Zajdel opisuje sprawowanie władzy w imieniu kosmitów (których nikt nigdy nie widział), jednak owi tajemniczy obcy spełniają wymogi na poziomie możliwości działania superinteligencji o charakterze suwerena. Por. J. Zajdel, *Limes inferior*, Warszawa 1982.

kim ludziom będzie lepiej, ale zarządzającym bardziej, jak mogłoby głosić, w złagodzonej formie, jedno z haseł *Animal Farm* (1945) George'a Orwella[18]. Czy występuje tu istotna różnica w stosunkach społecznych w kontekście historii politycznej ludzkości lub aktualnych struktur władzy? Czy bogatym nie jest lepiej niż biednym? Zasoby zawsze były rozmieszczone nierówno. Możemy dążyć do ich równomiernego dostępu dla wszystkich ludzi, a może tylko dla wspólnoty pewnych ludzi zgromadzonych w państwach typu narodowego. Ale zawsze ktoś musi dzielić te zasoby, a to już jest uprzywilejowana pozycja i tworzenie hierarchii.

Oba scenariusze są korzystne jeśli chodzi o ich skutki (poprawa funkcjonowania ludzi), chociaż pod cynicznym warunkiem, że to nas będą dotyczyły. Miejsca geograficzne bez SI pozostaną w stagnacji, bez żadnych dodatkowych korzyści, a ich mieszkańcy będą „skazani" na bogatych turystów z obszarów działania SI. Wykluczenie ze sztucznej inteligencji jest jeszcze gorsze niż „zniewolenie" przez nią i jej „zarządców". Tego aspektu, wydaje się, Bostrom nie dostrzega jako szczególne zagrożenie.

Każda nowa zmiana budzi strach, ale i nadzieję. Wykorzystanie siły zwierząt było jedną z większych rewolucji w dziejach ludzkości. Gdy koń został udomowiony, jego wpływ kreacyjny na cywilizację ludzką był olbrzymi – transport, rolnictwo, wojna, komunikacja, religia itd. A jednak chociaż był on dobrem samym w sobie, a narzędzia typu pług ewentualnym dodatkiem do niego, to już w konfrontacji z traktorem koń okazał się zbędny[19]. Czy ludzkość nie czeka los koni? Aktualnie SI jest dodatkiem do człowieka, ale może Superinteligencja ogólna, przewyższająca zdolności człowieka, nie będzie potrzebowała dodatku w formie ludzkiej istoty? Tak jak konie dzisiaj są głównie trzymane dla rozrywki (czasem w celach spożywczych), tak może ludziom zostanie zająć się rozrywką, być może dostarczaną przez SI. Nad tym powinniśmy się zastanawiać, chociaż czas podejmowania decyzji gwałtownie zaczyna się skracać. Musimy podejmować próby zrozumienia tego, co się dzieje, „co jest wizją wciąż stosunkowo bezkształtną i określoną

[18] G. Orwell, *Folwark zwierzęcy*, tłum. B. Zborski, Warszawa 2011.
[19] Według GUS-u pogłowie koni w ciągu tylko ostatnich 20 lat spadło w Polsce o ponad 50%: 2002 – 329 533, 2022 – 156 519, *Pogłowie koni w Polsce*, https://www.pzhk.pl/hodowla/poglowie-koni-polsce/ (data dostępu: 25.06.2023).

przez narrację – tego, co stawia nam w charakterze głównego zadania moralnego (przynajmniej z perspektywy laickiej i bezosobowej) ograniczenie groźby zagłady ludzkości i uzyskania trajektorii cywilizacyjnej, która doprowadzi nas do triumfalnego, a jednocześnie pełnego zrozumienia dla rodzaju ludzkiego wykorzystania kosmicznej spuścizny”[20]. Cdn.

Bibliografia / References

Asimov I., *Foundation*, New York 1951.

Bostrom N., *Superinteligencja. Scenariusze, strategie, zagrożenia*, tłum. D. Konowrocka-Sawa, Gliwice 2021.

Hare B., Woods V., *Przetrwają najżyczliwsi. Jak ewolucja wyjaśnia istotę człowieka?*, tłum. K. Kalinowski, Kraków 2022.

Hutter M., *Universal Artificial Intelligence. Sequential Decisions Based on Algorithmic Probability*, Berlin 2005.

Jonas H., *Zasada odpowiedzialności. Etyka dla cywilizacji technologicznej*, tłum. M. Klimowicz, Kraków 1996.

Landes D.S., *The Wealth and Poverty of Nations. Why Some Are So Rich and Some So Poor*, New York 1998.

Müller V.C., Bostrom N., *Future Progress in Artificial Intelligence: A Survey of Expert Opinion,* [w:] *Fundamental Issues of Artificial Intelligence*, red. V.C. Müller, Oxford 2016.

Orwell G., *Folwark zwierzęcy*, tłum. B. Zborski, Warszawa 2011.

Pogłowie koni w Polsce, https://www.pzhk.pl/hodowla/poglowie-koni-polsce.

Russell S., *Human Compatible: Artificial Intelligence and the Problem of Control*, London 2019.

Schwab K., Malleret Th., *Covid-19: The Great Reset*, Cologny 2020.

Sotala K., Yampolskiy R.V., *Responses to catastrophic AGI risk: a survey*, „Physica Scripta” 2015, 90 (1).

Tegmark M., *Życie 3.0. Człowiek w erze sztucznej inteligencji*, tłum. T. Krzysztoń, Warszawa 2019.

Zajdel J., *Limes inferior*, Warszawa 1982.

[20] N. Bostrom, *Superinteligencja. Scenariusze...*, s. 375–377.