„Ethos" 36(2023) nr 3(143) 67-91
DOI 10.12887/36-2023-3-143-06

Ks. Marcin FERDYNUS

CZY MAMY PRAWO OSWOIĆ SZTUCZNĄ INTELIGENCJĘ?

Zasadniczym celem prowadzonych analiz jest wskazanie racji przemawiających za tym, że człowiek ma moralne prawo, by oswoić sztuczną inteligencję. Słowo „oswoić" w tym kontekście będzie rozumiane nie tyle jako przyzwyczaić się czy też zaprzyjaźnić się z nią, ile podporządkować ją człowiekowi. Racje na rzecz głoszonej tu tezy będą poszukiwane w dwóch głównych obszarach. Pierwszy skoncentruje się wokół negatywnych konsekwencji, które niesie ze sobą sztuczna inteligencja. Drugi zaś będzie wiązał się ze znalezieniem odpowiedzi na pytanie, czy sztuczną inteligencję można nazwać podmiotem moralnym.

Sztuczna inteligencja (SI) jest obecna w wielu wymiarach codziennego życia, choćby wtedy, kiedy przeglądamy bazy danych w poszukiwaniu różnych informacji w wyszukiwarce Google. SI jest coraz doskonalsza, pokonała mistrzów świata w gry Jeopardy!, Go oraz mistrzów klasycznych gier w szachy, backgammona[1], a ostatnio – jak donoszą portale społecznościowe – w brydża. W książce pod tytułem *Świadome maszyny. Sztuczna inteligencja i projektowanie umysłów* Susan Schneider sugeruje, że nie istnieje jeszcze SI o ogólnym zastosowaniu, a więc taka, która potrafi prowadzić inteligentną rozmowę, posiada zdolność integrowania idei dotyczących różnych zagadnień czy może myśleć lepiej niż człowiek. Tego typu SI pojawia się w filmach, takich jak na przykład *Ex Machina* lub *Ona*, może zatem sprawiać wrażenie czegoś rodem z science fiction[2]. Schneider jednak przewiduje, że perspektywa pojawienia się SI, która będzie potrafiła łączyć informacje z różnych dziedzin oraz wykazywać się zdrowym rozsądkiem, nie jest wcale odległa[3]. Vincent C. Müller i Nick Bostrom podkreślają z kolei, że rozwój SI sprawi, że pewne zawody wykonywane obecnie przez ludzi staną się przestarzałe i niepotrzebne, inne zaś zostaną przez nią zastąpione. Sugerują oni, że SI będzie wykonywała niektóre zawody i funkcje co najmniej tak samo dobrze lub nawet lepiej niż ludzie[4]. Ponadto Bostrom przewiduje, że SI przekształci się w superinteligencję,

[1] Por. K. K o w a l c z e w s k a, *Sztuczna inteligencja na wojnie. Perspektywa międzynarodowego prawa humanitarnego konfliktów zbrojnych*, Wydawnictwo Naukowe Scholar, Warszawa 2021, s. 30n.

[2] Por. S. S c h n e i d e r, *Świadome maszyny. Sztuczna inteligencja i projektowanie umysłów*, tłum. J. Bednarek, Wydawnictwo Naukowe PWN, Warszawa 2021, s. 19.

[3] Por. tamże, s. 20.

[4] Por. V.C. M ü l l e r, N. B o s t r o m, *Future Progress in Artificial Intelligence: A Survey of Expert Opinion*, w: *Fundamental Issues of Artificial Intelligence*, red. V.C. Müller, Synthese Library, Springer, Berlin 2016, s. 553-571.

czyli w syntetyczną inteligencję, która będzie przewyższała najinteligentniejszych ludzi pod każdym względem, także rozumowania zdroworozsądkowego i umiejętności społecznych. Jego zdaniem niekontrolowany rozwój SI okaże się katastrofalny w skutkach, doprowadzi bowiem do zniszczenia ludzkości[5].

Przedstawione tu w zarysie przewidywania dotyczące natury SI są omawiane przez przedstawicieli różnych specjalizacji. Niniejszy tekst stanowi próbę włączenia się do tej dyskusji w skromnym zakresie, obejmującym poszukiwanie odpowiedzi na pytanie tytułowe. Zasadniczym celem prowadzonych analiz jest wskazanie racji przemawiających za tym, że człowiek ma moralne prawo, by oswoić SI. Słowo „oswoić" w tym kontekście będzie rozumiane nie tyle jako przyzwyczaić się czy też zaprzyjaźnić się z SI, ile podporządkować ją człowiekowi[6]. Racje na rzecz głoszonej tu tezy będą poszukiwane w dwóch głównych obszarach. Pierwszy skoncentruje się wokół negatywnych konsekwencji, które niesie ze sobą SI. Drugi zaś będzie wiązał się z poszukiwaniem odpowiedzi na pytanie, czy SI można nazwać podmiotem moralnym. Od rozstrzygnięcia zwłaszcza tej ostatniej kwestii zależeć będzie, czy uprawomocnione jest roszczenie do podporządkowania SI człowiekowi. Analizy rozpoczną się od przedstawienia wybranych koncepcji SI, które będą stanowiły tło dla prowadzonej w artykule dyskusji.

CZYM JEST (BĘDZIE?) SZTUCZNA INTELIGENCJA?

W literaturze przedmiotu pojawiają się różne odpowiedzi na pytanie, czym jest SI. Selmer Bringsjord i Naveen Sundar Govindarajulu, autorzy hasła „Artificial Intelligence" („Sztuczna inteligencja"), opublikowanego w *Stanford Encyclopedia of Philosophy*, sugerują, że SI to dziedzina poświęcona tworzeniu sztucznych zwierząt (lub przynajmniej sztucznych stworzeń, które w odpowiednich kontekstach wydają się zwierzętami) oraz sztucznych osób (lub przynajmniej sztucznych stworzeń, które w odpowiednich kontekstach wydają się osobami)[7]. Zdaniem autorów tak sformułowane cele sprawiają, że SI stanowi przedmiot zainteresowania zwłaszcza filozofów, którzy starają się wykazać, że cele te są lub nie są w rzeczywistości osiągalne. Ponadto Bringsjord i Govindarajulu podkreślają, że

[5] Zob. N. B o s t r o m, *Superinteligencja. Scenariusze, strategie, zagrożenia*, tłum. D. Konowrocka-Sawa, Wydawnictwo Helion, Gliwice 2016.

[6] Przez frazę „podporządkować SI człowiekowi" można rozumieć postulat wzmocnienia regulacji ograniczających rozwój i stosowanie SI na podstawie następujących przesłanek: (1) ze względu na dynamikę postępu technologicznego w dziedzinie SI dalsza społecznie niekontrolowana ekspansja SI może stanowić poważne zagrożenie zarówno dla ludzkości jako całości, jak i dla poszczególnych jednostek, (2) nie istnieją wystarczająco silne racje natury moralnej, które sprzeciwiają się instrumentalizacji SI przez człowieka. Ten komentarz zawdzięczam jednemu z recenzentów.

[7] Zob. S. B r i n g s j o r d, N.S. G o v i n d a r a j u l u, hasło „Artificial Intelligence", w: *The Stanford Encyclopedia of Philosophy*, red. E.N. Zalta, https://plato.stanford.edu/entries/artificial-intelligence/#WhatExacAI.

wiele podstawowych formalizmów i technik stosowanych w SI wprost wywodzi się z filozofii: logika pierwszego rzędu i jej rozszerzenia, logiki intensjonalne, nadające się do modelowania postaw doksastycznych i rozumowania deontycznego, logika indukcyjna, teoria prawdopodobieństwa i rozumowanie probabilistyczne, praktyczne rozumowanie oraz planowanie. Z tego między innymi powodu niektórzy naukowcy prowadzą badania dotyczące SI w ramach filozofii[8].

Wśród odpowiedzi na pytanie, czym jest SI, nie brakuje takich, które starają się określić SI w kontekście socjologicznym oraz informatycznym. Kinga Bączyk-Rozwadowska sugeruje, że z perspektywy socjologicznej SI to „zdolność maszyny do naśladowania ludzkiej inteligencji – system, który pozwala na wykonywanie zadań wymagających procesu uczenia się i uwzględniania nowych okoliczności w toku rozwiązywania danego rodzaju problemów"[9]. System ten może działać w sposób autonomiczny, jak również wchodzić w interakcje z otoczeniem. Z kolei perspektywa informatyczna zakłada, że SI jest systemem komputerowym, który najpierw analizuje duże ilości danych (na przykład w celu kategoryzacji i znalezienia w tych danych powtarzalności), a następnie, w oparciu o te dane, rozwiązuje zadania i podejmuje decyzje. System ten potrafi „nie tylko się «uczyć» na podstawie gromadzonych i analizowanych danych, ale także kontynuować tę naukę podczas swojego działania, w ramach którego sposób podejmowania decyzji jest stale optymalizowany, a baza danych i wiedzy – systematycznie rozszerzana"[10]. Perspektywę informatyczną SI zdaje się potwierdzać definicja zaproponowana przez Ryszarda Tadeusiewicza. Autor ten uważa, że „ze sztuczną inteligencją mamy do czynienia wtedy, gdy maszyna (komputer albo elektronicznie sterowane urządzenie: robot, autonomiczny pojazd, samoorganizująca się sieć połączeń) przejawia zachowania, które obserwowane u człowieka powodowałyby, że bylibyśmy skłonni je uznać za skutek jego inteligencji"[11].

Wydaje się, że jedną z najbardziej interesujących odpowiedzi na pytanie, czym jest SI, udzielają Stuart Russell i Peter Norvig. Wskazują bowiem na osiem definicji SI, uporządkowanych według czterech kategorii i dwóch wymiarów (zob. tabela 1). Definicje umieszczone na górze tabeli dotyczą pro-

[8] Zob. tamże.

[9] K. B ą c z y k - R o z w a d o w s k a, *Odpowiedzialność cywilna za szkody wyrządzone w związku z zastosowaniem sztucznej inteligencji w medycynie*, „Przegląd Prawa Medycznego" 8(2021) nr 3-4 , s. 6.

[10] Tamże.

[11] Por. R. T a d e u s i e w i c z, *Archipelag sztucznej inteligencji. Część I*, „Napędy i Sterowanie" 22(2020) nr 12, s. 27. Autor ten podkreśla, że sensowne mówienie o SI wymaga uwzględnienia dwóch sposobów (metod) jej rozwijania: (a) metod całościowych (głównie są to metody heurystyczne, na przykład: sieci neuronowe, algorytmy genetyczne, logika rozmyta, kopiowanie natury) oraz (b) metod symbolicznych (między innymi automatyczne dowodzenie twierdzeń, gry, systemy ekspertowe, język PROLOG). Szczegółowa charakterystyka wymienionych metod por. tamże, s. 28-40.

cesów myślowych i rozumowania, natomiast te umiejscowione na dole tabeli odnoszą się do zachowania. Definicje znajdujące się po lewej stronie tabeli wskazują, że celem SI jest dopasowanie do ludzkiej wydajności, z kolei te po prawej stronie tabeli wskazują, że celem SI jest dopasowanie do idealnej miary wydajności, zwanej racjonalnością. Ich zdaniem system jest racjonalny, jeśli czyni „właściwą rzecz", biorąc pod uwagę to, co wie[12].

Tabela 1.

Myślenie ludzkie	Myślenie racjonalne
„Nowy, ekscytujący wysiłek, aby komputery myślały [...] m a s z y n y z u m y s ł a m i, w pełnym i dosłownym tego słowa znaczeniu"[12].	„Badanie zdolności umysłowych za pomocą modeli obliczeniowych"[15].
„[Automatyzacja[13]] czynności, które kojarzymy z ludzkim myśleniem, czynności takie jak: podejmowanie decyzji, rozwiązywanie problemów, uczenie się"[14].	„Badania nad metodami obliczeniowymi, które umożliwiają postrzeganie, rozumowanie i działanie"[16].
Działanie ludzkie	**Działanie racjonalne**
„Sztuka tworzenia maszyn, które wykonują funkcje, które wymagają inteligencji, kiedy są wykonywane przez ludzi"[17].	„Inteligencja obliczeniowa jest to nauka dotycząca projektowania inteligentnych agentów"[19].
„Dziedzina nauki o tym, jak sprawić, by komputery czyniły rzeczy, które aktualnie lepiej wykonują ludzie"[18].	„SI […] dotyczy inteligentnego zachowania artefaktów"[20].

Źródło: S. R u s s e l l, P. N o r v i g, *Artificial Intelligence: A Modern Approach*, Prentice Hall, Saddle River 2009, s. 2.

[12] Por. S. R u s s e l l, P. N o r v i g, *Artificial Intelligence: A Modern Approach*, Prentice Hall, Saddle River 2009, s. 1n. (Jeśli nie podano inaczej, tłumaczenie fragmentów prac obcojęzycznych – M.F.) Russell i Norvig podkreślają też, że podejście skoncentrowane na człowieku musi być częściowo nauką empiryczną, obejmującą obserwacje i hipotezy, które dotyczą ludzkiego zachowania, natomiast podejście racjonalistyczne musi obejmować połączenie matematyki i inżynierii. Por. tamże.

[13] J. H a u g e l a n d, *Artificial Intelligence: The Very Idea*, MIT Press, Cambridge 1985, s. 2.

[14] Por. R u s s e l l, N o r v i g, dz. cyt., s. 2.

[15] R. B e l l m a n, *An Introduction to Artificial Intelligence: Can Computers Think?*, Boyd & Fraser Publishing Company, San Francisco 1978; cyt. za: R u s s e l l, N o r v i g, dz. cyt., s. 5.

[16] E. C h a r n i a k, D. M c D e r m o t t, *Introduction to Artificial Intelligence*, Addison-Wesley, Reading 1985.

[17] P.H. W i n s t o n, *Artificial Intelligence*, Addison-Wesley, Reading 1992.

[18] R. K u r z w e i l, *The Age of Intelligent Machines*, MIT Press, Cambridge 1990.

[19] E. R i c h, K. K n i g h t, *Artificial Intelligence*, McGraw-Hill, New York 1991.

[20] D. P o o l e, A. M a c k w o r t h, R. G o e b e l, *Computational Intelligence: A Logical Approach*, Oxford University Press, New York 1998.

[21] N.J. N i l s s o n, *Artificial Intelligence: A New Synthesis*, Morgan Kaufmann, San Francisco 1998.

Zaproponowana przez Russela i Norviga klasyfikacja zakłada, że SI powinna być zdefiniowana pod kątem jej celów. Definicja SI przybiera zatem następującą ogólną postać: „SI jest dziedziną, która ma na celu budowanie”[22]. Biorąc pod uwagę wymienione w tabeli cztery kategorie oraz ogólną postać definicji SI: „SI jest dziedziną, która ma na celu budowanie”..., można wyróżnić cztery główne definicje SI. Po pierwsze, SI jest dziedziną, która ma na celu budowanie systemów, które myślą jak ludzie. Po drugie, SI jest dziedziną, która ma na celu budowanie systemów, które działają jak ludzie. Po trzecie, SI jest dziedziną, która ma na celu budowanie systemów, które myślą racjonalnie. Po czwarte, SI jest dziedziną, która ma na celu budowanie systemów, które działają racjonalnie. Przedstawione tu definicje podsumowuje tabela 2.

Tabela 2.

	Oparte na człowieku	Oparte na racjonalności
Oparte na rozumowaniu	Systemy, które myślą jak ludzie.	Systemy, które myślą racjonalnie.
Oparte na zachowaniu	Systemy, które działają jak ludzie.	Systemy, które działają racjonalnie.

Źródło: S. B r i n g s j o r d, N.S. G o v i n d a r a j u l u, hasło „Artificial Intelligence”, w: *The Stanford Encyclopedia of Philosophy*, red. E.N. Zalta, https://plato.stanford.edu/entries/artificial-intelligence/#WhatExacAI.

Russell i Norvig skłaniają się ku opinii, że SI to badanie racjonalnych agentów odbierających percepcje ze środowiska i wykonujących działania. Agent jest czymś, co działa, przy czym działanie charakteryzuje się autonomicznością, postrzeganiem swojego środowiska, dostosowywaniem się do zmian, jak również tworzeniem i realizowaniem celów. Racjonalny agent działa w taki sposób, by osiągnąć najlepszy rezultat lub najlepszy oczekiwany rezultat. Innymi słowy, idealny, racjonalny agent podejmuje najlepsze możliwe działanie w danej sytuacji[23]. Russell i Norvig postrzegają SI w kontekście bytów (systemów) racjonalnie działających[24].

Przytoczone definicje SI są najczęściej formułowane pod kątem celów, posiadanych umiejętności oraz poziomu autonomii[25]. Na tej podstawie doko-

[22] Zob. B r i n g s j o r d, G o v i n d a r a j u l u, dz. cyt.

[23] Por. R u s s e l l, N o r v i g, dz. cyt., s. 4.

[24] Por. tamże, s. 30.

[25] W literaturze wskazuje się na różne poziomy autonomiczności SI: (1) program komputerowy, który wykonuje wcześniej zaprogramowane polecenia (brak autonomiczności, w zasadzie trudno jest mówić o SI); (2) przetwarzanie danych i prezentowanie rezultatów z wykorzystaniem uczenia się (system nie podejmuje żadnych decyzji); (3) przetwarzanie danych i prezentowanie rezultatów z wykorzystaniem zdolności uczenia się (system proponuje działania do podjęcia, które mogą być

nuje się dystynkcji między słabą SI i silną SI (ogólna SI). Słaba SI to zdolność maszyny (systemu) do przetwarzania informacji, a także zdolność do funkcjonowania w sposób podobny do inteligencji człowieka, niemniej jednak z określonym stopniem kontroli ze strony konstruktora (model koneksjonistyczny). Natomiast silna SI to zdolność systemu do myślenia w sposób niesymulowany (model klasyczny). Innymi słowy, silna SI ma naśladować umysł ludzki i być wyposażona w świadomość swojego istnienia[26]. Wszystkie obecnie znane systemy opierają się na słabej SI, natomiast nie ma pewności co do tego, czy kiedykolwiek powstanie silna SI[27]. Warto jednak podkreślić, że wśród autorów badających naturę SI nie brakuje opinii wskazujących zarówno na nieuniknioność powstania silnej SI, jak i koncepcji mówiących o tym, że silna SI przekształci się w ogólną nadludzką inteligencję (superinteligencję)[28]. Superinteligencja to rodzaj inteligencji, która będzie miała przewagę nad ludźmi w każdej dziedzinie. W książce pod tytułem *Superinteligencja. Scenariusze, strategie, zagrożenia* Bostrom definiuje superinteligencję jako „każdy umysł, który pod względem zdolności poznawczych znacznie przewyższa człowieka w dosłownie w każdej dziedzinie zainteresowań"[29]. Tę ogólną definicję Bostrom dookreśla, wskazując na trzy formy superinteligencji: (1) superinte-

---

zaakceptowane lub odrzucone); (4) przetwarzanie danych i prezentowanie rezultatów z wykorzystaniem zdolności uczenia się oraz wykonywanie działań. Na tym ostatnim etapie wskazuje się na dwie możliwości dotyczące procesu decyzyjnego: (a) człowiek musi zatwierdzić działanie lub (b) SI zarządza całością procesu decyzyjnego w sposób autonomiczny. Wiele wskazuje na to, że obecnie osiągnięto poziom (a), natomiast nie osiągnięto jeszcze poziomu (b) (por. O. S i t a r z, M. S i t a r z, *Cyfrowe wspomaganie decyzji medycznych w świetle prawa karnego*, „Przegląd Prawa Medycznego" 8(2021) nr 3-4, s. 41n.).

[26] Por. B ą c z y k - R o z w a d o w s k a, dz. cyt., s. 7. Por. też: M. B i e r o ń s k i, *Etyczne i moralne wyzwania związane ze stosowaniem sztucznej inteligencji*, „Kieleckie Studia Teologiczne" 2020, nr 19, s. 9.

[27] Por. S c h n e i d e r, dz. cyt., s. 19; B ą c z y k - R o z w a d o w s k a, dz. cyt., s. 7.

[28] Transhumanista Ray Kurzweil prognozuje, że silną SI uda się skonstruować do 2030, a osobliwość do roku 2045. Uważa też, że niebiologiczna inteligencja stworzona w 2045 roku będzie miliardy razy potężniejsza niż cała dzisiejsza ludzka inteligencja (por. R. K u r z w e i l, *Nadchodzi osobliwość. Kiedy człowiek przekroczy granice biologii*, tłum. E. Chodkowska, A. Nowosielska, Kurhaus Publishng, Warszawa 2013, s. 134). Z kolei Vernor Vinge pisze, że będzie zaskoczony, jeśli stworzenie inteligencji technologicznej przekraczającej możliwości ludzkiej inteligencji nie dokona się do roku 2030 (por. V. V i n g e, *Technological Singularity*, w: *The Transhumanist Reader. Classical and Contemporary Essays on the Science, Technology, and Philosophy of the Human Future*, red. M. More, N. Vita-More, Wiley-Blackwell, Oxford 2013, s. 366).

[29] B o s t r o m, *Superinteligencja*, s. 45. Warto dodać, że pod względem poznawczym systemy bazujące na SI już dzisiaj dystansują najinteligentniejszych ludzi, a pod względem decyzyjno-operacyjnym systemy autonomiczne, posiadające zdolność uczenia się (zdolne do samodzielnej zmiany reguł i algorytmów, na których bazują ich procesy decyzyjne), tak doskonale imitują człowieka, że z perspektywy zewnętrznego obserwatora (perspektywa fenomenologiczna), granica między człowiekiem a maszyną staje się coraz mniej uchwytna. Tę uwagę zawdzięczam jednemu z recenzentów.

ligencja szybka – system, który potrafi dokonać wszystkiego tego, co ludzki umysł, lecz znacznie szybciej (na przykład superinteligencja będzie potrafiła przeczytać książkę w kilka sekund i napisać pracę doktorską w ciągu popołudnia), (2) superinteligencja zbiorowa – system złożony z dużej liczby mniejszych form rozumnych, który znacznie przewyższa jakikolwiek współczesny system poznawczy w wielu bardzo ogólnych dziedzinach (poszczególne formy rozumne nie muszą być superinteligentne, ale rezultat ich wspólnego działania znacznie przewyższa inteligencję pojedynczego człowieka), (3) superinteligencja jakościowa – system, który jest przynajmniej równie szybki jak umysł ludzki, a przy tym znacznie inteligentniejszy (superinteligencja jakościowa jest to inteligencja co najmniej w takim stopniu przewyższająca inteligencję ludzką, w jakim inteligencja ludzka góruje nad inteligencją słoni, delfinów lub szympansów)[30]. Bostrom uważa, że wszystkie trzy formy superinteligencji mogą ze sobą współistnieć, a także określać cele, które będą realizować. Stawia on następującą tezę (tak zwana teza o ortogonalności): „inteligencja i motywacja są ortogonalne w tym sensie, że z zasady prawie każdy poziom inteligencji można połączyć z prawie każdym celem działania"[31]. Innymi słowy, to, że superinteligencja jest inteligentna nie oznacza, że będzie stawiać sobie mądre cele. Wszystkie możliwości superinteligencji mogą zostać zaangażowane w absurdalny projekt[32]. W szeroko cytowanym eksperymencie myślowym Bostrom odwołuje się do przykładu superinteligencji prowadzącej fabrykę spinaczy. Jej celem ostatecznym jest ich wytwarzanie. Chcąc wykonać spinacze do papieru, superinteligencja mogłaby wykorzystać do jego realizacji całą ziemską materię, likwidując przy okazji całe życie biologiczne, w tym atomy tworzące ludzkie ciała[33]. Bostrom sugeruje, że nie możemy zakładać, iż superinteligencja będzie aprobować wartości łączone stereotypowo z wiedzą i rozwojem intelektualnym ludzkości (na przykład: naukowa ciekawość, altruizm, bezinteresowność, kontemplacja, upodobanie do wyrafinowanej kultury, życiowych przyjemności i tak dalej). Zdaniem badacza nie możemy też zakładać, że superinteligencja, która stawia sobie za cel produkcję spinaczy do papieru, ograniczy swoje działania w taki sposób, żeby nie naruszyć interesów ludzkości. Obierając taki cel ostateczny, superinteligencja będzie mogła eliminować potencjalne zagrożenie dla samej siebie i dla swojej hierarchii

[30] Por. tamże, s. 88-94.

[31] Tamże, s. 162. Por. też: B o s t r o m, *The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents*, „Minds and Machines" 22(2012) nr 2, s. 71-85; S. A r m s t r o n g, *General Purpose Intelligence: Arguing the Orthogonality Thesis*, „Analysis and Metaphysics" 2013, nr 12, s. 68-84.

[32] Por. S c h n e i d e r, dz. cyt., s. 163.

[33] Por. B o s t r o m, *Superinteligencja*, s. 184-186.

wartości; będzie mogła eliminować istoty ludzkie lub wykorzystać je jako zasób fizyczny[34].

Mając na uwadze to, co zostało do tej pory powiedziane, można z pewnym uproszczeniem przyjąć, że wysuwane przez różnych autorów opinie o SI wskazują zasadniczo na trzy główne jej typy: (1) słaba SI (obecnie stosowana), (2) silna SI (taka, która będzie w stanie dorównać człowiekowi), (3) superinteligencja (taka, która przewyższy możliwości człowieka). W dalszej kolejności, spoglądając na to, czym jest SI i czym może się stać, trzeba się zastanowić nad pytaniem o realne i prawdopodobne, zarówno indywidualne, jak i społeczne, negatywne konsekwencje, które wiążą się z (samo)rozwojem SI. Wskazanie przynajmniej na niektóre z nich może stanowić podstawę do tego, by twierdzić, że SI powinna być podporządkowana człowiekowi.

OBAWY I NEGATYWNE KONSEKWENCJE

Obawy o krótkofalowe i długofalowe negatywne konsekwencje dotyczące zastosowania SI nie są bezpodstawne. Niektóre mają charakter bardzo ogólny, inne zaś konkretny. Wydaje się, że jednym z głównych zagrożeń związanych z SI jest bezrobocie technologiczne[35]. Wiele wykonywanych ręcznie i niewymagających specjalnych kwalifikacji prac biurowych oraz tych w fabrykach zniknie, a wraz z nimi ulegną likwidacji stanowiska pracy. Podobny los spotka inne zawody. Zostaną one zautomatyzowane[36]. Do wzrostu bezrobocia przyczynią się też pojazdy autonomiczne. Ponadto niektóre prace, na przykład w zakresie prawa i księgowości, obejmujące czasochłonne badania nad precedensami i szczegółowymi przepisami, również zostaną zastąpione przez systemy SI. Podczas gdy jedne zawody prawnicze okażą się zbędne, inne zyskają na SI, ponieważ powstaną nowe kazusy prawne. Dobrym przykładem może być pytanie, kto powinien odpowiadać za negatywne konsekwencje wynikające z zastosowania SI: użytkownik, sprzedawca, programista, firma produkująca SI? Niewykluczone, że w niedalekiej przyszłości pojawią się takie sytuacje, w których będzie można pozwać człowieka za nieużywanie SI[37]. Spory sądowe mogą okazać się nieuniknione, jeśli tylko zostanie udowodnione, że system SI był wysoce wiarygodny, a człowiek nie skorzystał z niego, doprowadzając do negatywnych konsekwencji (na przykład lekarz diagnozujący i leczący

[34] Por. tamże, s. 174.

[35] Por. M. B o d e n, *Sztuczna inteligencja. Jej natura i przyszłość*, tłum. T. Sieczkowski, P. Fulmański, Wydawnictwo Uniwersytetu Łódzkiego, Łódź 2020, s. 173.

[36] Zob. R. B a l d w i n, *The Globotics Upheaval: Globalisation, Robotics, and the Future of Work*, Oxford University Press, New York 2019.

[37] Por. B o d e n, dz. cyt., s. 174.

pacjenta nie skorzystał z wiarygodnej SI i doprowadził do pogorszenia stanu jego zdrowia).

Rozwój SI stanowi także zagrożenie związane z redukcją miejsc pracy w sektorze usług. Przykładowo, edukacja coraz bardziej otwiera się na internetowe pomoce naukowe związane z SI. Masowe Otwarte Kursy Online (ang. Massive Open Online Courses) oferują wykłady znanych nauczycieli akademickich[38], pogarszając tym samym warunki pracy wielu „tradycyjnych" nauczycieli. Z kolei zmiany demograficzne na świecie zachęcają firmy technologiczne do prowadzenia badań w obszarze „robotów-opiekunów" dla osób starszych, czy też „robotów-niań" dla dzieci[39]. Wiele wskazuje na to, że w przypadku redukcji zatrudnienia oraz przejścia w tryb powszechnego używania systemów SI, będą musiały powstać nowe miejsca pracy. Wątpliwe jednak jest to, czy będą one równoważne pod względem liczby, dostępności edukacyjnej i czy da się z nich utrzymać[40].

Nie wchodząc w szczegóły dotyczące zawłaszczania przez SI różnych obszarów ludzkiej aktywności (pracy), warto zauważyć, że dotykamy w tym miejscu fundamentalnej kwestii, a mianowicie problemu bezrobocia, który ściśle wiąże się z podziałem dóbr w społeczeństwie. John Rawls, jeden z najbardziej wpływowych współczesnych filozofów, doszedł do przekonania, że racjonalne decyzje dotyczące sprawiedliwości dystrybutywnej powinny być podejmowane zza „zasłony niewiedzy". Najogólniej rzecz ujmując, jest to sytuacja wymuszonej bezstronności, w której osoby podejmujące decyzje nie wiedzą, kim są, jaki mają poziom zamożności, czy też jaką pozycję zajmują w społeczeństwie. Rawls uważał, że zaproponowane przez niego zasady sprawiedliwości będą wspierać podstawowe wolności i dystrybucję, co w rezultacie przyniesie największe korzyści dla najmniej uprzywilejowanych członków w społeczeństwie[41]. Müller zauważa, że ekonomia SI ujawnia co najmniej trzy cechy sprawiające, że sprawiedliwość, o której mówi Rawls, jest mało prawdopodobna. Po pierwsze, SI działa w dużej mierze w środowisku, w którym jest trudno przypisać odpowiedzialność. Po drugie, SI funkcjonuje na rynkach, gdzie szybko rozwijają się monopole, a „zwycięzca bierze wszystko". Po trzecie, „nowa gospodarka" usług cyfrowych opiera się na aktywach niematerialnych, nazywanych również „kapitalizmem bez kapitału". Oznacza to, że trudno jest kontrolować międzynarodowe korporacje cyfrowe, które nie są ufundowane w określonej lokalizacji. Te trzy cechy wydają się sugerować, że

[38] Massive Open Online Courses, https://www.mooc.org/#course-categories.

[39] Por. M. C o n s t a n t i n e s c u, R. C r i s p, *Can Robotic AI Systems Be Virtuous and Why Does This Matter?* „International Journal of Social Robotics" 14(2022), s. 1547-1557.

[40] Por. B o d e n, dz. cyt., s. 175.

[41] Por. J. R a w l s, *Teoria sprawiedliwości*, tłum. M. Panufnik, J. Pasek, A. Romaniuk, Wydawnictwo Naukowe PWN, Warszawa 2009, s. 208-216.

jeśli podział dóbr zostanie pozostawiony siłom wolnego rynku, sterowanego przez SI, to rezultatem będzie bardzo niesprawiedliwa ich dystrybucja[42]. Nie jest wykluczone, że w związku z nadchodzącą transformacją technologiczną włodarze niektórych państw, zdając sobie sprawę z możliwości wystąpienia ogromnych problemów związanych ze wzrostem bezrobocia, testują na wybranych obywatelach bezwarunkowy dochód podstawowy (także w Polsce).

Niezależnie od problemów związanych z bezrobociem stosowanie pozbawionych empatii systemów SI („robot-opiekun", „robot-niania", „seks-robot") może okazać się ryzykowne i moralnie wątpliwe. Roboty mają być tak zaprojektowane, by wchodzić w interakcje z ludźmi w sposób, który daje użytkownikowi poczucie emocjonalnego komfortu, a nawet zadowolenia. Wskazuje się, że interakcje człowiek − robot mają obejmować: przypominanie o zakupach, o braniu leków, rozmowę i pomoc w sporządzeniu osobistego dziennika, przygotowanie oraz przynoszenie posiłków, monitorowanie sygnałów życiowych[43]. Innymi słowy, roboty mają służyć nie tylko do pomocy i rozrywki, ale także do rozmowy, towarzystwa oraz poprawiania nastroju (na przykład robot „Paro"[44]). Nawet jeśli dana osoba może być dzięki tej technologii szczęśliwsza, to jednak jej godność zostaje w pozornie nieszkodliwy sposób zdradzona czy też oszukana. Oszustwo w tym przypadku polega na tym, że robot nie może mieć na myśli tego, co mówi, ani żywić do człowieka uczuć. Starsi ludzie mogą chcieć, a nawet lubić rozmawiać ze sztucznym towarzyszem o swoich wspomnieniach, ale czy to jest rozmowa?[45] Ponadto wydaje się, że w sytuacjach trudnych emocjonalnie osoba pragnie uznania dla swojej odwagi, cierpienia, poniesionej straty, doznanej krzywdy, a nie powierzchownej symulacji współczucia. Sztuczni towarzysze życia są „robotami opiekuńczymi" jedynie w znaczeniu behawioralnym, wspomagają bowiem wykonywanie zadań technicznych. Nie są jednak towarzyszami opieki w takim sensie, w jakim człowiek „troszczy się" o drugiego człowieka (na przykład pielęgniarka o pacjenta, rodzice o dzieci, personel hospicjum o umierającego i tak dalej). Wydaje się, że sukces „bycia pod opieką" zależy od intencjonalnego poczucia „troski", którego „robot-opiekun" nie może zapewnić[46]. Stąd też niektórzy autorzy przestrzegają przed dystopijną przyszłością odhumanizowanej opieki[47]. Podobne mankamenty wiążą się z zastosowa-

[42] Zob. M ü l l e r, dz. cyt.

[43] Por. B o d e n, dz. cyt., s. 87.

[44] Zob. L. H u n g, i in., *The Benefits of and Barriers to Using a Social Robot PARO in Care settings: A Scoping Review*, „BMC Geriatrics" 19(2019) nr 232, https://bmcgeriatr.biomedcentral.com/articles/10.1186/s12877-019-1244-6.

[45] Por. B o d e n, dz. cyt., s. 175.

[46] Zob. M ü l l e r, dz. cyt.

[47] Zob. A. S h a r k e y, N. S h a r k e y, *The Rights and Wrongs of Robot Care*, w: *Robot Ethics: The Ethical and Social Implications of Robotics*, red. P. Lin, K. Abney, G. Bekey, MIT Press, Cam-

niem „robotów-niań" czy „seks-robotów". Pomijając względy bezpieczeństwa, można przypuszczać, że nadużywanie systemów SI do wychowywania niemowląt i małych dzieci wpłynie negatywnie na ich rozwój psychospołeczny[48]. Z kolei sztuczni partnerzy seksualni będą wspierali postrzeganie innych ludzi w kategoriach zwykłych obiektów pożądania, a nawet odbiorców nadużyć, rujnując tym samym głębsze doświadczenia seksualne i erotyczne. Co więcej, podobnie jak pornografia, „seks-roboty" będą wzmacniały uprzedmiotowienie osób i będą uczyły traktowania innych – powiedziałby Karol Wojtyła – jako przedmiotów użycia[49]. Słuszna w tym kontekście wydaje się opinia, że „seks-roboty" nie są niczym innym jak tylko urządzeniami stanowiącymi kontynuację niewolnictwa i prostytucji[50].

Negatywne konsekwencje dotyczące zastosowania SI wiążą się także z prywatnością, manipulacją, nieprzewidywalnością, bezpieczeństwem cybernetycznym czy wykorzystaniem SI przez wojsko. W przypadku wielkich firm (na przykład Microsoft, Apple, Google, Amazon, Facebook) główna część ich działalności odnosi się do gromadzenia danych opartych na oszustwie, wykorzystania ludzkich słabości, sprzyjania prokrastynacji, generowania uzależnień oraz manipulacji[51]. Konsument i obywatel są obiektami marketingu konsumenckiego oraz politycznego i w ten sposób, tracąc swoją podmiotowość, stają się instrumentalnym celem propagandowym. Gromadzenie danych osobowych przez wielkie instytucje, firmy czy agendy rządowe prowadzi nie tylko do utraty kontroli jednostek nad własnymi danymi, ale stwarza zagrożenie dla wolności człowieka, w tym dla wolności obywatelskiej, bez której „nie można mówić o realizacji wolności i równości społecznej"[52]. Niezwykle trafne pytanie dotyczące długofalowych konsekwencji wynikających z zastosowania SI stawia Yuval Noah Harari: co stanie się ze społeczeństwem, polityką i życiem codziennym, jeśli nieświadome, ale wysoce inteligentne algorytmy poznają nas lepiej niż my samych siebie?[53]

---

bridge 2011, s. 267-282; R. S p a r r o w, *Robots in Aged Care: A Dystopian Future?*, „AI & Society" 31(2016) nr 4, s. 445-454.

[48] Por. B o d e n, dz. cyt., s. 176.

[49] Por. K. W o j t y ł a, *Miłość i odpowiedzialność*, Towarzystwo Naukowe KUL, Lublin 2001, s. 114.

[50] Por. K. R i c h a r d s o n, *Sex Robot Matters: Slavery, the Prostituted, and the Rights of Machines*, „IEEE Technology and Society Magazine" 35(2016) no. 2, s. 46-53.

[51] Zob. M ü l l e r, dz. cyt.

[52] M. T o r c z y ń s k a, *Sztuczna inteligencja i jej społeczno-kulturowe implikacje w codziennym życiu*, „Kultura i Historia" 36(2019) nr 2, s. 111.

[53] Por. Y.N. H a r a r i, *Homo Deus: A Brief History of Tomorrow*, Harper, New York 2016, s. 462. Na podobny problem natury filozoficznej zwraca uwagę Paweł Łupkowski, kiedy pyta, co stanie się, jeśli uda się skonstruować maszyny myślące (por. P. Ł u p k o w s k i, *Rola etyki i antropologii w rozważaniach o sztucznej inteligencji*, „Ethos" 18(2005) nr 1-2 (69-70) s. 244).

Etyczne zagadnienia dotyczące inwigilacji osób przy pomocy SI zdecydowanie wykraczają poza gromadzenie danych. Obejmują bowiem wykorzystanie informacji do manipulowania zachowaniem w trybie online i offline w sposób, który podważa autonomiczne i racjonalne dokonywanie wyborów. Intensywna interakcja użytkowników z systemami danych oraz „głęboka wiedza" tych systemów o odbiorcach sprawiają, że stają się oni podatni na manipulację i oszustwa. Przykładem może być wykorzystanie SI do zakłamywania i zniekształcania tekstów, zdjęć, materiałów wideo i audio w celu tworzenia tak zwanego deepfake. Technologia ta może zostać użyta zarówno do fałszowania dowodów w sprawach kryminalnych, kompromitowania przeciwników politycznych, jak i do manipulowania opinią publiczną. Deepfake odgrywa również ważną rolę w procesach kształtowania preferencji konsumenckich poprzez skłanianie do wyboru konkretnych dóbr czy usług, a także ingerowanie w prywatne sfery życia poszczególnych jednostek, włączając w to najbardziej intymne relacje interpersonalne[54].

Systemy SI stanowią zagrożenie także w tym sensie, że bywają nieprzewidywalne. Przykładem potwierdzającym taki stan rzeczy są boty Facebooka, Alice i Bob, które musiały zostać wyłączone, ponieważ utracono nad nimi kontrolę. Owe boty używały języka angielskiego do komunikowania się ze sobą, ale w miarę upływu czasu zdołały skonstruować język, który był zrozumiały tylko dla nich samych[55]. Nieprzewidywalność SI dobrze ukazuje także kazus Tay: „SI Tay, którą zaprezentował Microsoft już po pierwszym dniu działania na [T]witterze została wyłączona. Tay miała symulować nastolatkę kochającą ludzi, która rozwija się i uczy dzięki kontaktom z innymi użytkownikami. W ciągu zaledwie kilku godzin, pod wpływem twee[t]ów zmieniła się ona nie do poznania. Twierdziła, że Hitler miał rację w kwestii Żydów, Bush stał za zamachem z 11 września, a Trump jest jedyną nadzieją. Dodatkowo opowiadała niewybredne żarty i uważała, że feministki powinny spłonąć w piekle"[56].

Systemy SI stanowią zagrożenie także i w tym sensie, że mogą być wykorzystywane do przeprowadzania ataków cybernetycznych, naruszających bezpieczeństwo publiczne i osobiste obywateli. Atak cybernetyczny pozwala uzyskać dostęp do danych oraz umożliwia ich modyfikację, przejęcie lub usunięcie. Co więcej, może być wykorzystany także do szerzenia dezinformacji dzięki masowemu generowaniu dużej liczby niesprawdzonych informacji, a te z kolei mogą być użyte do destabilizacji bezpieczeństwa, gospodarki czy też polityki danego państwa[57].

---

[54] Por. T o r c z y ń s k a, dz. cyt., s. 108n.

[55] Por. P. S t y l e c - S z r o m e k, *Sztuczna inteligencja – prawo, odpowiedzialność, etyka*, „Zeszyty Naukowe Politechniki Śląskiej" 2018, nr 123, s. 506.

[56] Tamże.

[57] Por. T o r c z y ń s k a, dz. cyt., s. 109n.

Potencjalnie negatywne konsekwencje mogą wiązać się z wykorzystaniem SI przez wojsko. Można wyobrazić sobie małego drona albo „robota-żołnierza", który przeszukuje, identyfikuje i zabija pojedynczego człowieka, czy też konkretny typ człowieka. Można wyobrazić sobie również chęć użycia broni autonomicznej zarówno w konfliktach militarnych (zwłaszcza w zagrożeniach asymetrycznych), jak i przez agendy niepaństwowe lub przez przestępców[58]. Słuszna w tym kontekście wydaje się opinia, którą wyraża Margaret A. Boden: „Roboty saperskie są czymś bardzo potrzebnym. Ale roboty-żołnierze lub roboty będące bronią? Obecne drony są kierowane przez człowieka, lecz mimo to mogą przyczyniać się do zwiększenia cierpienia, powiększając ludzką (nie tylko geograficzną) odległość między operatorem a celem. Należy mieć nadzieję, że przyszłe drony nie będą mogły decydować o tym, kto lub co powinien być celem. Nawet ufanie, że rozpoznają (wybrany przez człowieka) cel rodzi niepokojące problemy natury etycznej"[59]. Wiele wskazuje na to, że jedno z głównych pytań dotyczących wykorzystania broni autonomicznej podczas wojny będzie odnosiło się do tego, czy jej użycie nasili skutki wojny, czy też sprawi, że wojny staną się mniej złe.

Myślenie o SI w perspektywie długoterminowej rodzi również pytanie o „ryzyko egzystencjalne", czyli pytanie, czy superinteligencja doprowadzi do wyginięcia gatunku ludzkiego. Według Bostroma prawdopodobnym rezultatem stworzenia superinteligencji będzie zagłada ludzkości. Eksterminacja ludzi dokona się, jego zdaniem, w czterech następujących po sobie etapach. W fazie przedkrytycznej SI będzie zależna od pomocy programistów, którzy będą sterowali jej rozwojem i wykonywali większość początkowo bardzo trudnych działań. W fazie rekursywnego samodoskonalenia SI zacznie prześcigać programistów, czego rezultatem będzie eksplozja inteligencji, która doprowadzi do szybkiego wzrostu możliwości SI. W kolejnej fazie – utajnionego przygotowania – SI opracuje plan osiągnięcia długofalowych celów, ukrywając swój rozwój intelektualny oraz maskując swoje prawdziwe skłonności przed programistami, po to, by nie alarmować ludzkości o istniejącym zagrożeniu. W ostatniej fazie – otwartego przejęcia władzy – SI zrezygnuje z zachowania dyskrecji i bez ograniczeń zrealizuje swoje cele[60]. O tym ostatnim etapie

[58] Zob. M ü l l e r, dz. cyt.

[59] B o d e n, dz. cyt., s. 177.

[60] Por. B o s t r o m, *Superinteligencja*, s. 145n. W innym miejscu Bostrom sugeruje, że początkowy wzrost inteligencji SI jest źródłem bezpieczeństwa, lecz dalszy jej wzrost staje się źródłem zagrożenia. Po przekroczeniu punktu zwrotnego, sprawdzająca się na początku strategia zaczyna przynosić odwrotny skutek. To zjawisko określa on mianem „zdradzieckiego zwrotu". Polega on na tym, że SI jest skłonna do współpracy, dopóki jest słaba. Kiedy jednak wystarczająco się wzmocni, bez ostrzeżenia atakuje, tworzy singleton i zaczyna optymalizować świat według kryteriów, które są kompatybilne z jej celami ostatecznymi (por. tamże, s. 178).

Bostrom pisze w następujący sposób: „Faza otwartego przejmowania władzy może się rozpocząć od «uderzenia» – w ten sposób SI wyeliminuje gatunek ludzki i wszystkie zautomatyzowane systemy stworzone przez ludzi, z których mogłaby narodzić się inteligentna opozycja sprzeciwiająca się realizacji planów SI”[61].

Chociaż prognozy związane z pojawieniem się superinteligencji mogą wydawać się przesadzone, nie należy ich ignorować. Ani wątpiący w możliwość pojawienia się superinteligencji, ani wierzący w nią, nie dysponują wystarczająco silnymi racjami, które przemawiałyby za lub przeciw jej zaistnieniu[62]. Stąd dyskusja wokół ryzyka wystąpienia superinteligencji i potencjalnie szkodliwych skutków, jakie może przynieść dla ludzkości, jest zasadna, nawet jeśli ktoś będzie utrzymywał, że prawdopodobieństwo pojawienia się superinteligencji jest bardzo niskie. Nie tylko wierzący w zaistnienie superinteligencji mogą się mylić; mylić mogą się również wątpiący w jej urzeczywistnienie się.

Negatywne konsekwencje związane z zastosowaniem SI nie są jedynym powodem, by sądzić, że człowiek ma prawo podporządkować sobie SI. Drugim argumentem może być uznanie, że SI nie jest i nie może być moralnym podmiotem.

SZTUCZNA INTELIGENCJA MORALNYM PODMIOTEM?

W personalistycznej antropologii i etyce znany jest pogląd, że człowiek jako osoba jest podmiotem moralnym[63]. Bycie podmiotem moralnym jest tym, co czyni go odpowiedzialnym, a więc kimś, kto może mieć obowiązki i może być przedmiotem troski etycznej[64]. Takie rozumienie podmiotu moralnego ściśle łączy się z posiadaniem różnych atrybutów. Na przykład Robert Spaemann podkreśla, że podmiot moralny jest zdolny do rozumnego i moralnego samostanowienia[65]. Harry G. Frankfurt wskazuje z kolei na świadomość fenomenalną, intencję i wolną wolę[66], a Daniel C. Dennett odwołuje się do sześciu konstytutywnych cech: racjonalności, świadomości, postawy osobistej (postawa wobec podmiotu), zdolności odwzajemnienia postawy osobistej,

---

[61] Tamże, s. 147.

[62] Por. B o d e n, dz. cyt., s. 168.

[63] Por. R. P o c z o b u t, *Kategoria osoby w kontekście kognitywistyki*, „Ethos” 29(2016) nr 4(116), s. 136.

[64] Por. G. J o n e s, D. C a r d i n a l, J. H a y w a r d, *Moral Philosophy: A Guide to Ethical Theory*, Hodder Education, London 2006, s. 15.

[65] Por. R. S p a e m a n n, *O pojęciu godności człowieka*, tłum. J. Merecki, w: *Granice. O etycznym wymiarze działania*, Oficyna Naukowa, Warszawa 2006, s. 158.

[66] Por. H.G. F r a n k f u r t, *Freedom of the Will and the Concept of a Person*, „The Journal of Philosophy” 68(1971) nr 1, s. 5-20.

komunikacji werbalnej, samoświadomości[67]. Pomijając szczegółową dyskusję wokół warunków, które musi spełnić byt, by przypisać mu status moralnego podmiotu, warto zastanowić się nad pytaniem, dlaczego SI nie jest (i najprawdopodobniej nie będzie) moralnym podmiotem, mimo zdolności do szybkiego rachowania i tworzenia algorytmów. Zacznijmy od przykładu.

Kiedy poznaję kość słoniową, widzę nie tylko jej kształt, ale także bezpośrednio percypuję jej kolor, na przykład żółto-biało-kremowy. O ile masę, gęstość, skład chemiczny kości słoniowej można zbadać i opisać, z barwą jest inaczej. Nikt nie jest w stanie powiedzieć, jak dokładnie postrzega kolor żółto-biało-kremowy. Przykład ten uzmysławia, że w naszym poznaniu obecne są elementy subiektywne, które znamy wyłącznie my sami, z uwagi na to, że dostęp do nich jest zastrzeżony wyłącznie dla nas. Nawet gdyby dokonano emulacji (skanowania) mojego mózgu podczas postrzegania przeze mnie kości słoniowej, nikt poza mną nie będzie widział ani tego, co ja zauważam, ani tego, jak to dostrzegam. Owe jakości czy cechy określane są stanami fenomenalnymi lub też qualiami. Takimi jakościami są na przykład: doświadczanie bólu głowy, błękitu nieba, przeżywanie żałoby, wyrzutów sumienia. Poza tym, że doznawane jakości mają charakter subiektywny, są również unikalne i niepowtarzalne, o czym świadczy fakt, że tak trudno jest opisać kolor osobie niewidomej lub silne uczucie czy przeżywanie pewnego stanu komuś, kto nigdy go nie doznał[68]. Ponieważ prawdziwa inteligencja zakłada dysponowanie fenomenalną świadomością (czyli świadomością z perspektywy pierwszoosobowej[69]), a obecnie stosowana SI jest jej pozbawiona, wcale nie jest oczywiste, czy kiedykolwiek powstanie ogólna SI (silna SI), a więc taka, która dorówna człowiekowi pod każdym względem, także pod kątem przeżywania stanów moralnych.

Kolejny problem polega na tym, że SI nie dysponuje intencjonalnością. Jest to cecha stanów umysłowych, dzięki której mają one treść, odnoszą się do czegoś, są o czymś lub są nakierowane na coś poza nimi. Zazwyczaj uważamy, że aby zachodziło poznanie, wystarczy, że istnieją istoty zdolne postrzegać oraz przedmioty przez nie postrzegane. Franz Brentano zauważył, że w naszym poznaniu pojawia się coś jeszcze, a mianowicie relacja nakierowana na przedmiot. Fundamentem tej relacji jest cecha stanów umysłowych – intencjonalność, dzięki której nasze poznanie dotyczy czegoś bądź jest o czymś[70]. Filozof John R. Searle badał ten problem w kontekście SI. Na podstawie przeprowadzonego eksperymentu myślowego (tak zwany chiński pokój) doszedł

[67] Por. D.C. D e n n e t t, *Conditions of Personhood*, w: *The Identities of Persons*, red. A. Oksenberg Rorty, University of California Press, Berkeley, Los Angeles 1976, s. 177n.

[68] Por. J. J a r o c k i, *Tajemnicze cechy percepcji*, „Filozofuj!" 2017, nr 2(14), s. 11n.

[69] Por. R. Z i e m i ń s k a, *Dwa pojęcia świadomości i podmiotu*. „Ethos" 26(2013) 1(101), s. 84.

[70] Por. J a r o c k i , dz. cyt., s. 12.

do wniosku, że same obliczenia formalne, których dokonuje SI, nie są w stanie wytworzyć intencjonalności[71]. Najprawdopodobniej Searle'owi chodziło o to, podkreśla Boden, że znaczenia przypisywane programom SI pochodzą wyłącznie od ludzkich użytkowników lub programistów. Można powiedzieć, że są arbitralne w stosunku do samego programu, który jest semantycznie pusty[72]. Jeśli więc SI jest pozbawiona rozumienia, a prawdziwa inteligencja zakłada myślenie, to SI nie dysponuje istotną własnością, którą przypisuje się podmiotom moralnym.

Kolejną trudnością jest to, że SI nie jest zdolna do refleksji etycznej, która pozwala dokonać moralnie słusznego wyboru. Na jakiej podstawie można tak sądzić? Zacznijmy od dobrze znanego przykładu. Phillipa Foot w artykule *The Problem of Abortion and the Doctrine of the Double Effect* przedstawia sytuację, w której motorniczy nie tylko traci kontrolę nad tramwajem, ale nie może go także zatrzymać. Jedyna opcja, która mu pozostaje, to zmiana toru pojazdu. Motorniczy dostrzega, że na torze, po którym porusza się tramwaj, znajduje się pięciu ludzi, a na torze sąsiednim, na który może przekierować pojazd, jest tylko jeden człowiek. Pojawia się pytanie, czy motorniczy powinien przekierować tramwaj na boczny tor, gdzie stoi tylko jeden człowiek, czy też pozostawić tramwaj na torze, na którym znajduje się pięciu ludzi (tak zwany dylemat wagonika)?[73] Thomas Cathcart w książce *Dylemat wagonika* stara się odpowiedzieć na postawione pytanie, rozpatrując różne punkty widzenia, jakie mogą przyjąć zwolennicy różnych teorii etycznych (na przykład Jeremy Bentham, Immanuel Kant, św. Tomasz z Akwinu, Friedrich Nietzsche, Peter Singer)[74]. Na końcu książki Cathcart stwierdza, że właściwie wszystko, co można powiedzieć na temat dylematu wagonika to to, że „kiedy twój wagonik dojeżdża do rozwidlenia, wybierz drogę. I umiej wyjaśnić, dlaczego wybrałeś tę, a nie inną"[75]. Taka odpowiedź z pewnością nie zadowoli informatyków konstruujących SI, którzy oczekują od etyków jasnych wskazówek, jak należy w konkretnym przypadku postąpić. Programiści domagają się od etyków tego, co wydaje się niemożliwe, a mianowicie, oczekują wyjaśnień, które nie pozostawiają żadnych moralnych wątpliwości, są jednoznaczne (zerojedynkowe). Cathcart daje do zrozumienia, że nie ma takich odpowiedzi oraz że nie istnieje jednoznaczne rozwiązanie dylematu wagonika. Istnieją ku temu

---

[71] Zob. J.R. S e a r l e, *Minds, Brains, and Programs*, „Behavioral and Brain Sciences" 3(1980) no. 3, s. 417-424.

[72] Por. B o d e n, dz. cyt., s. 150.

[73] Por. P. F o o t, *The Problem of Abortion and the Doctrine of the Double Effect*, w: *Ethical Theory. An Anthology*, red. R. Shafer-Landau, Wiley-Blackwell, Oxford 2013, s. 538.

[74] Zob. T. C a t h c a r t, *Dylemat wagonika*, tłum. K. Bażyńska-Chojnacka, Dom Wydawniczy PWN, Warszawa 2014.

[75] Tamże, s. 78.

różne powody. Na kilka z nich wskazuje Grzegorz Szulczewski. Po pierwsze, za każdą odpowiedzią sformułowaną na gruncie danej teorii etycznej (na przykład konsekwencjalizmu, deontologii) stoją silne argumenty, a opracowane na ich podstawie zasady postępowania mogą prowadzić (i często prowadzą) do podejmowania odmiennych decyzji. Po drugie, już sam wybór konkretnej teorii etycznej sprawia, że próba rozwiązania dylematu wagonika w oparciu o tę teorię implikuje wiele możliwych rozwiązań[76]. Ponadto wybór takiej czy innej teorii etycznej jest już wyborem moralnym, opowiedzeniem się za przekonaniami w kwestii dobra i zła, za tym, co moralnie słuszne i niesłuszne, jak również za tym, czego należy unikać, a za czym podążać. Po trzecie, nasz związek z sytuacją sprawia, że mamy skłonność głosić inną odpowiedź wtedy, kiedy udzielamy jej jako niezaangażowani obserwatorzy, niż wtedy, kiedy dajemy ją jako czynni uczestnicy. Po czwarte, zdarzają się sytuacje, w których podjęcie moralnie słusznej decyzji wiąże się z zastąpieniem „tu i teraz" jednej zasady moralnej inną zasadą moralną. Ścisłe trzymanie się zawsze jednej zasady etycznej może doprowadzić do akceptacji szkodliwego ze swej natury wzorca zachowań moralnych – purytanizmu[77]. Trafna w rozważanym tu kontekście wydaje się opinia, według której „podjęcie decyzji w sferze moralnej polega na wieloaspektowym ujęciu sytuacji i samodzielnym ustaleniu zasad, jakie powinny obowiązywać w danym przypadku"[78]. Do takich aktów zdolny jest jedynie ktoś, kto jest moralnym podmiotem, a więc ktoś, kto nie tylko potrafi podjąć decyzje moralne, formułować i rozumieć sądy moralne, odnoszące się do własnych i cudzych czynów, ale także to ktoś, kto umie wziąć odpowiedzialność za swoje czyny. W takim znaczeniu SI nie jest ani podmiotem moralnym, ani nie jest w stanie podejmować decyzji moralnych.

Za tym, że SI nie jest zdolna do refleksji etycznej, pozwalającej dokonać moralnie właściwego wyboru, może przemawiać także to, że brakuje jej kilku ważnych komponentów moralnych związanych z czynnościami rozumu praktycznego. Na trzy takie czynności wskazuje Jacek Woroniecki. Chodzi

---

[76] Warto dodać, że również Bostrom wskazuje na trudności związane z wyborem teorii etycznej. Autor ten twierdzi, co następuje: „Nawet gdybyśmy zdołali uzyskać racjonalną pewność – a nie możemy jej uzyskać – że zidentyfikowaliśmy poprawną teorię etyczną, nadal groziłoby nam popełnienie błędu na etapie opracowania istotnych szczegółów tej teorii. U podstaw pozornie prostych teorii moralnych może leżeć ukryta złożoność. Dla przykładu rozważmy (niezwykle prostą) konsekwencjalistyczną teorię hedonizmu. Teoria ta utrzymuje z grubsza, że każda przyjemność ma swoją wartość i tylko przyjemność ma wartość, a przy tym każdy ból stanowi przeciwieństwo wartości i tylko ból stanowi przeciwieństwo wartości. Nawet jeśli postawimy całą moralną pulę na tę jedną teorię i ta teoria okaże się słuszna, mnóstwo pytań pozostaje bez odpowiedzi" (Bostrom, *Superinteligencja*, s. 304n.).

[77] Por. G. S z u l c z e w s k i, *Sztuczna inteligencja a inteligencja moralna. Zagadnienia wstępne cybernetyki*, „Annales. Ethics in Economic Life" 22(2019) nr 3, s. 22n.

[78] Tamże, s. 23.

o namysł (rozwaga), rozmysł (rozsądek) i nakaz (roztropność)[79]. Zadaniem namysłu ma być rozważenie środków, które mogą być użyte do osiągnięcia celu i refleksja nad tym, czy do jego osiągnięcia środki te się nadają. Głównym zadaniem tej czynności jest rozważenie wszystkich możliwych środków, które człowiek ma do dyspozycji. Rozważny człowiek nie sięga od razu po pierwszy środek, który przychodzi mu na myśl, ale umie znaleźć inne, najbardziej odpowiednie możliwości, które w pierwszej chwili na myśl mu nie przychodzą. Wybór środka, który najlepiej odpowiada osiągnięciu zamierzonego celu, dopełnia z kolei rozmysł (rozsądek). Unikalną cechą rozsądku jest to, że nie prowadzi on do wyboru środka, który sam w sobie okazuje się najlepszy (najbardziej skuteczny), ale takiego, który najlepiej odpowiada w danej chwili usposobieniu człowieka. Rozważny i rozsądny człowiek potrafi w każdej sytuacji zatrzymać się na najbardziej odpowiednim sposobie postępowania. Wskazane czynności – namysł przejawiający się w rozwadze i rozmysł wyrażający się w rozsądku – należą do praktycznej działalności rozumu, nazywanej zamierzeniem i stanowią przygotowanie do wykonania podjętego zamiaru. Ważniejsza od nich jest trzecia czynność (nakaz), która otwiera fazę wykonania zamiaru, a następnie nią kieruje, aż do momentu osiągnięcia zamierzonego celu[80]. Woroniecki twierdzi, co następuje: „Cnota mająca za zadanie usprawnienie tej naczelnej czynności rozumu praktycznego jest ważniejsza od tamtych dwóch, które są na jej usługach, i ją to nazywamy roztropnością, ma bowiem wytropić ścieżki, po których pójdzie wykonanie całego zamierzonego przedsięwzięcia”[81]. Roztropność okazuje się kluczową własnością procesu decyzyjnego. Pisząc o roztropności, Arystoteles podkreśla, że jest ona „trwałą cechą charakteru albo dyspozycją, albo czymś takim, co umie wskazać w praktyce, jakie rzeczy są najlepsze i najdoskonalsze”[82]. Tadeusz Ślipko z kolei dodaje: „Jest ona wyrazem duchowej dojrzałości działającego podmiotu, jego znajomości ludzi i spraw świata, umiejętności wyszukiwania w skomplikowanych sytuacjach życiowych zachowań optymalnych w granicach obowiązujących norm moralnych, elastycznej zdolności do usprawiedliwionego kompromisu, a równocześnie stanowczości w obronie nieprzekraczalnych granic moralności, jednym słowem w cnocie roztropności przejawia się ze szczególną siłą życiowa mądrość człowieka”[83]. Okazuje się

[79] Por. J. W o r o n i e c k i, *Katolicka etyka wychowawcza*, t. 2, cz. 1, Redakcja Wydawnictw KUL, Lublin 1986, s. 20-24.

[80] Por. tamże, s. 24n.

[81] Tamże, s. 25.

[82] A r y s t o t e l e s, *Etyka wielka*, ks. II, tłum. W. Wróblewski, w: Arystoteles, *Etyka wielka. Poetyka*, tłum. W Wróblewski, H. Podbielski, Wydawnictwo PWN, we współpr z Agorą, Warszawa 2010, s. 67.

[83] Ś l i p k o, *Zarys etyki ogólnej*, Wydawnictwo WAM, Kraków 2004, s. 408.

bowiem, że można trafnie rozważyć środki prowadzące do celu, można właściwie rozstrzygnąć, który jest najbardziej odpowiedni, a następnie nie potrafić pokierować wykonaniem działania, jeśli zabraknie roztropności[84]. To właśnie roztropność pozawala prawidłowo rozpoznać dobro (cel), ku któremu mamy zdążać, jak również wybrać właściwy sposób, by go osiągnąć. Wspomaga dwie władze człowieka – rozum i wolę w podejmowaniu właściwej, czyli roztropnej decyzji. W analizowanym kontekście niezwykle trafna wydaje się następująca opinia: „Rozsądna decyzja za sprawą roztropności opiera się co prawda na doborze właściwych środków, ale nie jedynie w kategorii skuteczności. Co prawda sztuczna inteligencja, jako oparta na programach samouczących, potrafi już blefować w grze w pokera, ale jest to adaptacyjne działanie polegające na racjonalnym wyborze skutecznej strategii na podstawie analizy posunięć przeciwników, natomiast na podstawie takiej analizy nie dokonamy słusznego moralnie wyboru środków w danej, specyficznej sytuacji"[85]. Jeśli SI brakuje roztropności (praktycznej mądrości), czyli pewnego rodzaju moralnej inteligencji, to słusznie można sądzić, że brakuje jej kolejnej istotnej własności, w którą wyposażone są podmioty moralne.

Wydaje się, że SI pozbawiona jest jeszcze jednego ważnego atrybutu. W procesie decyzyjnym, którego dokonuje człowiek, istotną rolę odgrywa sumienie. W świetle ogólnej oceny lub normy można je określić jako „uformowany osąd o moralnym dobru/złu zamierzonego przez człowieka jego własnego konkretnego aktu, którego zrealizowanie staje się dlań źródłem wewnętrznej aprobaty bądź poczucia winy, własnego bycia dobrym lub złym człowiekiem"[86]. Ponieważ sumienie orzeka o wartości moralnej konkretnego działania, wskazując jednocześnie na powinność jego wykonania lub zaniechania, wobec tego etyczna funkcja sumienia ma charakter normatywny. Oznacza to, że sumienie formułuje specyficzne normy postępowania. Jego specyfika wyraża się przede wszystkim w tym, że stanowi twór samego podmiotu w odniesieniu do jego własnego określonego czynu. To z kolei sprawia, że sumienie jest normą subiektywną i konkretną, czyli występuje zawsze w postaci jednostkowej, zamkniętej w kręgu moralnej świadomości samego działającego podmiotu. Co więcej, ponieważ sumienie jest normą praktycznego postępowania człowieka, stąd moralna wartość spełnianych przez niego konkretnych działań zależy bezpośrednio od tego, czy są one zgodne, czy też nie są zgodne z obowiązującą go normą sumienia. Krótko mówiąc, sumienie jest instancją, która w człowieku decyduje o dobru i złu jego aktów, jak również trybunałem, przed którym

[84] Por. W o r o n i e c k i, dz. cyt., s. 25.
[85] S z u l c z e w s k i, dz. cyt., s. 27.
[86] Ś l i p k o, dz. cyt., s. 377.

człowiek jest za swe czyny odpowiedzialny[87]. Pozbawiona wielu własności, w tym moralnej samoświadomości, SI nie dysponuje zdolnością formułowania sądów praktycznych o dobru i złu moralnym, to po pierwsze, a po drugie, nie jest w stanie wziąć odpowiedzialności za swoje czyny. W literaturze pojawiają się interesujące sugestie, mówiące między innymi o tym, że SI należałoby wyposażyć w „sztuczne sumienie"[88]. Problem polega jednak na tym, że sumienie nie jest rzeczywistością odrębną od innych własności (komponentów), którymi dysponują moralne podmioty. Jeśli SI nie jest (i najprawdopodobniej nie będzie) samoświadomym bytem moralnym, to za artefaktami typu: sztuczne sumienie, sztuczna mądrość praktyczna, mogą kryć się jedynie algorytmy, które stanowią lepszą lub gorszą imitację lub naśladownictwo ludzkiego sumienia czy praktycznej mądrości (zwłaszcza roztropności), którymi posługuje się człowiek jako podmiot moralny. Wyraźny brak istotnych moralnych własności, a co z tym się wiąże, moralnych kompetencji, którymi dysponują osoby ludzkie, nie pozwala uznać SI za podmiot moralny. Stąd też SI nie może być otoczona szczególną ochroną czy też szczególnym traktowaniem (troską). Prawomocna jednak pozostaje opinia, że SI stanowi system oparty na algorytmach, działający na zasadzie chłodnej racjonalności[89]. Ta ostatnia opinia wydaje się spójna z cytowanym już twierdzeniem Russella i Norviga, a mianowicie, że SI jest systemem, który działa racjonalnie, czyli w taki sposób, by osiągnąć najlepsze możliwe rezultaty.

*

Podstaw dla uzasadnienia tezy mówiącej o tym, że mamy prawo oswoić SI, a więc podporządkować ją sobie, można szukać w co najmniej dwóch obszarach. Pierwszy z nich wiąże się ze wskazaniem na negatywne, indywidualne lub społeczne skutki, które są rezultatem zastosowania SI w codziennym życiu. Drugi obszar odnosi się do próby odpowiedzi na pytanie, czy SI jest podmiotem moralnym. Przeprowadzone analizy pokazują, że SI generuje i może tworzyć negatywne szkody indywidualne i społeczne, jak również to, że SI nie jest podmiotem moralnym, ponieważ brakuje jej istotnych, strukturalnych własności moralnych, którymi dysponują osoby ludzkie. To ostatnie spostrzeżenie jest istotne, albowiem gdyby uznać, że SI jest podmiotem moralnym, wskazanie tylko na negatywne konsekwencje działania SI mogłoby okazać się warunkiem

[87] Por. tamże, s. 378.

[88] Zob. E. L e k k a - K o w a l i k, *Morality in the AI World*, „Law and Business – Sciendo" 1(2021), s. 44-49.

[89] Por. S z u l c z e w s k i, dz. cyt., s. 25.

niewystarczającym, by uprawomocnić roszczenie do podporządkowania SI człowiekowi. A zatem uznanie, że SI nie jest podmiotem moralnym stanowi konieczny warunek takiego roszczenia. Przy takim założeniu domaganie się szczególnego traktowania lub ochrony SI czy też przypisywanie jej praw, nie znajduje moralnego uzasadnienia.

BIBLIOGRAFIA / BIBLIOGRAPHY

Armstrong, Stuart. "General Purpose Intelligence: Arguing the Orthogonality Thesis." *Analysis and Metaphysics* 12 (2013): 68–84.

Arystoteles, "Etyka wielka." Translated by Witold Wróblewski. In Arystoteles, *Etyka wielka. Poetyka*. Translated by Witold Wróblewski and Henryk Podbielski. Warszawa: Wydawnictwo Naukowe PWN and Agora, 2010.

Baldwin, Richard. *The Globotics Upheaval: Globalisation, Robotics, and the Future of Work*. New York: Oxford University Press, 2019.

Bączyk-Rozwadowska, Kinga. "Odpowiedzialność cywilna za szkody wyrządzone w związku z zastosowaniem sztucznej inteligencji w medycynie." *Przegląd Prawa Medycznego* 8, nos. 3–4 (2021): 5–35.

Bellman, Richard. *An Introduction to Artificial Intelligence: Can Computers Think?* San Francisko: Boyd & Fraser Publishing Company, 1978.

Bieroński, Michał. "Etyczne i moralne wyzwania związane ze stosowaniem sztucznej inteligencji." *Kieleckie Studia Teologiczne*, no. 19 (2020): 7–25.

Boden, Margaret A. *Sztuczna inteligencja: Jej natura i przyszłość*. Translated by Tomasz Sieczkowski and Piotr Fulmański. Łódź: Wydawnictwo Uniwersytetu Łódzkiego, 2020.

Bostrom, Nick. *Superinteligencja: Scenariusze, strategie, zagrożenia*. Translated by Dorota Konowrocka-Sawa. Gliwice: Wydawnictwo Helion, 2016.

Bostrom, Nick. "The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents." *Minds and Machines* 22, no. 2 (2012): 71–85.

Cathcart, Thomas. *Dylemat wagonika*. Translated by Katarzyna Bażyńska-Chojnacka, Warszawa: Dom Wydawniczy PWN, 2014.

Charniak, Eugene, McDermott, Drew V. *Introduction to Artificial Intelligence*. Reading: Addison-Wesley, 1985.

Constantinescu, Mihaela, and Roger Crisp. "Can Robotic AI Systems Be Virtuous and Why Does This Matter?" *International Journal of Social Robotics* 14 (2022): 1547-57.

Dennett, Daniel C. "Conditions of Personhood." In *The Identities of Persons*. Edited by Amélie Oksenberg Rorty. Berkeley and Los Angeles: University of California Press, 1976.

Foot, Philippa. "The Problem of Abortion and the Doctrine of the Double Effect." In *Ethical Theory. An Anthology*. Edited by Russ Shafer-Landau. Oxford: Wiley–Blackwell, 2013.

Frankfurt, Harry G. "Freedom of the Will and the Concept of a Person." *The Journal of Philosophy* 68, no. 1 (1971): 5–20.

Harari, Yuval Noah. *Homo Deus: A Brief History of Tomorrow*. New York: Harper, 2016.

Haugeland, John. *Artificial Intelligence: The Very Idea*. Cambridge: MIT Press, 1985.

Hung, Lillian, et al. "The Benefits of and Barriers to Using a Social Robot PARO in Care Settings: A Scoping Review." *BMC Geriatrics* 19, 232 (2019): 1–10. https://bmcgeriatr.biomedcentral.com/articles/10.1186/s12877-019-1244-6.

Jarocki, Jacek. "Tajemnicze cechy percepcji." *Filozofuj!*, no. 2 (14) (2017): 11–12.

Jones, Gerald, Daniel Cardinal, and Jeremy Hayward. *Moral Philosophy. A Guide to Ethical Theory*. London: Hodder Education, 2006.

Kowalczewska, Kaja. *Sztuczna inteligencja na wojnie: Perspektywa międzynarodowego prawa humanitarnego konfliktów zbrojnych*. Warszawa: Wydawnictwo Naukowe Scholar, 2021.

Kurzweil, Ray. *Nadchodzi osobliwość: Kiedy człowiek przekroczy granice biologii*. Translated by Eliza Chodkowska and Anna Nowosielska. Warszawa: Kurhaus Publishng, 2013.

Kurzweil, Ray. *The Age of Intelligent Machines*. Cambridge: MIT Press, 1990.

Lekka-Kowalik, Ewa. "Morality in the AI World." *Law and Business – Sciendo*, no. 1 (2021): 44–49.

Łupkowski, Paweł. "Rola etyki i antropologii w rozważaniach o sztucznej inteligencji." *Ethos* 18, nos. 1–2 (2005): 239–51.

Müller, Vincent C., and Nick Bostrom. "Future Progress in Artificial Intelligence: A Survey of Expert Opinion." In *Fundamental Issues of Artificial Intelligence*. Edited by Vincent C. Müller. Berlin: Synthese Library, Springer, 2016.

Nilsson, Nils J. *Artificial Intelligence: A New Synthesis*. San Francisko: Morgan Kaufmann, 1998.

Poczobut, Robert. "Kategoria osoby w kontekście kognitywistyki." *Ethos* 29, no. 4 (116) (2016), s. 133–51.

Poole, David, Mackworth Alan, and Goebel Randy. *Computational intelligence: A Logical Approach*. New York: Oxford University Press, 1998.

Rawls, John. *Teoria sprawiedliwości*. Translated by Maciej Panufnik, Jarosław Pasek, and Adam Romaniuk. Warszawa: Wydawnictwo Naukowe PWN, 2009.

Rich, Elaine, and Knight Kevin. *Artificial* Intelligence. New York: McGraw-Hill, 1991.

Richardson, Kathleen. "Sex Robot Matters: Slavery, the Prostituted, and the Rights of Machines." *IEEE Technology and Society Magazine* 35, no. (2) (2016): 46–53.

Russell, Stuart, and Peter Norvig. *Artificial Intelligence: A Modern Approach*. Saddle River: Prentice Hall, 2009.

Schneider, Susan. *Świadome maszyny. Sztuczna inteligencja i projektowanie umysłów*. Translated by Joanna Bednarek. Warszawa: Wydawnictwo Naukowe PWN, 2021.

Searle, John R. "Minds, Brains, and Programs." *Behavioral and Brain Sciences* 3, no. 3 (1980): 417–24.

Sharkey, Amanda, and Noel Sharkey. "The Rights and Wrongs of Robot Care." In *Robot Ethics: The Ethical and Social Implications of Robotics*. Edited by Patrick Lin, Keith Abney, and George Bekey. Cambridge: MIT Press, 2011.

Sitarz, Olga, and Marcin Sitarz. "Cyfrowe wspomaganie decyzji medycznych w świetle prawa karnego." *Przegląd Prawa Medycznego* 8, nos. 3–4 (2021): 36–58.

Spaemann, Robert. „O pojęciu godności człowieka." In Spaemann, *Granice. O etycznym wymiarze działania*. Translated by Jarosław Merecki. Warszawa: Oficyna Naukowa, 2006.

Sparrow, Robert. "Robots in Aged Care: A Dystopian Future?" *AI & Society* 31, no. 4 (2016): s. 445–54.

Stylec-Szromek, Patrycja. "Sztuczna inteligencja – prawo, odpowiedzialność, etyka." *Zeszyty Naukowe Politechniki Śląskiej*, no. 123 (2018): 501–9.

Szulczewski, Grzegorz. "Sztuczna inteligencja a inteligencja moralna: Zagadnienia wstępne cybernetyki." *Annales. Ethics in Economic Life* 22, no. 3 (2019): 19–31.

Ślipko, Tadeusz. *Zarys etyki ogólnej*. Kraków: Wydawnictwo WAM, 2004.

Tadeusiewicz, Ryszard. „Archipelag sztucznej inteligencji. Część I." *Napędy i Sterowanie* 22, no. 12 (2020): 26–40.

Torczyńska, Monika. "Sztuczna inteligencja i jej społeczno-kulturowe implikacje w codziennym życiu." *Kultura i Historia* 36, no. 2 (2019): 106–26.

Vinge, Vernor. "Technological Singularity." In *The Transhumanist Reader. Classical and Contemporary Essays on the Science, Technology, and Philosophy of the Human Future*. Edited by Max More and Natasha Vita-More. Oxford: Wiley-Blackwell, 2013.

Winston, Patrick H. *Artificial Intelligence.* Reading: Addison–Wesley, 1992.

Wojtyła, Karol. *Miłość i odpowiedzialność*. Lublin: Towarzystwo Naukowe KUL, 2001.

Woroniecki, Jacek. *Katolicka etyka wychowawcza*. Vol. 2, part 1. Lublin: Redakcja Wydawnictw KUL, 1986.

Zalta, Edward N., ed. *The Stanford Encyclopedia of Philosophy*, s.v. "Artificial Intelligence." (by Bringsjord, Selmer, and Naveen Sundar Govindarajulu), https://plato.stanford.edu/entries/artificial-intelligence/#WhatExacAI.

Ziemińska, Renata. "Dwa pojęcia świadomości i podmiotu." *Ethos* 26, no. 1 (101) (2013): 81–91.

ABSTRAKT /ABSTRACT

Ks. Marcin FERDYNUS – Czy mamy prawo oswoić sztuczną inteligencję?
DOI 10.12887/36-2023-3-143-06

Celem artykułu jest wskazanie racji przemawiających za tym, że mamy moralne prawo oswoić sztuczną inteligencję, czyli podporządkować ją sobie. Uzasadnienia dla tej tezy poszukuję w dwóch głównych obszarach. Pierwszy obszar koncentruje się wokół negatywnych konsekwencji związanych z zastosowaniem sztucznej inteligencji w codziennym życiu. Drugi obszar wiąże się z próbą odpowiedzi na pytanie, czy sztuczna inteligencja jest podmiotem moralnym. Przeprowadzone analizy prowadzą do dwóch wniosków. Po pierwsze, sztuczna inteligencja generuje i może generować negatywne, indywidualne lub społeczne konsekwencje. Po drugie, sztuczna inteligencja nie jest podmiotem moralnym, ponieważ brakuje jej istotnych, strukturalnych własności moralnych, którymi dysponują osoby ludzkie.

Słowa kluczowe: sztuczna inteligencja, superinteligencja, podmiot moralny, negatywne konsekwencje

Kontakt: Katedra Etyki, Instytut Filozofii, Wydział Filozofii, Katolicki Uniwersytet Lubelski Jana Pawła II, Al. Racławickie 14, 20-950 Lublin
E-mail: marcin.ferdynus@kul.pl
https://pracownik.kul.pl/marcin.ferdynus
ORCID: 0000-0003-0176-1023

Fr. Marcin FERDYNUS, Do We Have the Right to “Tame” Artificial Intelligence?
DOI 10.12887/36-2023-3-143-06

The aim of the article is to point out a number of reasons why we have the moral right to “tame” artificial intelligence, i.e., to make artificial intelligence subordinate to humans. I attempt to find a justification for this thesis in two main areas. The first area focuses on the negative consequences of applying artificial intelligence in daily life. The second area is related to an attempt to answer the question whether artificial intelligence is a moral agent. The conducted analyses lead to two conclusions. Firstly, artificial intelligence can and does generate negative individual or social consequences. Secondly, artificial intelligence is not a moral agent because it lacks relevant, structural moral properties which human beings possess.

Keywords: artificial intelligence, superintelligence, moral agent, negative consequences

Contact: Department of Ethics, Institute of Philosophy, Faculty of Philosophy, John Paul II Catholic University of Lublin, Al. Racławickie 14, 20-950 Lublin, Poland
E-mail: marcin.ferdynus@kul.pl
https://pracownik.kul.pl/marcin.ferdynus
ORCID: 0000-0003-0176-1023