Sztuczna inteligencja w świecie etyki

Jakie praktyczne podejście do nauczania postępowania etycznego najlepiej sprawdzi się w domenie sztucznej inteligencji?

Żyjemy w czasach, kiedy sztuczna inteligencja, automatyzacja procesów oraz coraz powszechniejsze zastosowanie robotów, zaczynają przykuwać uwagę ogółu. Należy zwrócić uwagę na dynamiczny rozwój sztucznej inteligencji w dziedzinach takich jak ekonomia, gdzie ma ona za zadanie określanie popytu i podaży, przemysł, w którym sterowanie robotami przy taśmie produkcyjnej powierzone jest programowi komputerowemu czy w medycynie, w której lekarzy zmagających się z wyzwaniami diagnostycznymi wspiera sztuczna inteligencja.[1]

Początku sztucznej inteligencji należy upatrywać w rewolucji cyfrowej, która miała miejsce w Stanach Zjednoczonych tuż po II wojnie światowej. Sama nazwa pochodzi od amerykańskiego naukowca Johna McCarthy'ego, który ukuł go latem 1956 roku na kampusie Dartmouth College, dzięki czemu do dzisiaj uznawany jest za ojca tej dziedziny nauki. Konferencja naukowa o takiej właśnie nazwie, zorganizowana w tym samym roku nadała badaniom impet, który nie stracił na sile po dziś dzień.[2]

Według McCarthy'ego sztuczna inteligencja zajmuje się badaniami i tworzeniem inteligentnych podmiotów (ang: intelligent agents). Nazwa dziedziny może być także interpretowana jako umiejętność tworzonych w jej ramach przedmiotów do przejawiania inteligencji na wzór ludzki, co czyni z nich podmioty inteligentne.[3] Takie podejście nawiązuje do zaproponowanej przez Marvina Minsky'ego w 1960 r. definicji sztucznej inteligencji „jako nauki o maszynach realizujących zadania, które wymagają inteligencji, gdy są wykonywane przez człowieka."[4]

Do tej pory termin sztuczna inteligencje przybierał, raczej swodobnie, jedno z dwóch generycznych znaczeń. Zakres pojęciowy jest na tyle szeroki, że przed przystąpieniem do analizy implikacji zastosowania różnych metod nauczania zasad moralnych w dziedzinie sztucznej inteligencji należy przedstawić cztery różne podejścia i wynikające z nich implikacje dla wcześniej wspomnianego procesu w zestawianiu z egzemplarzami sztucznej inteligencji.

Robert Penrose w swojej książce „Shadows of the Mind" oznaczył czterema pierwszymi literami alfabetu łacińskiego różne stanowiska dotyczące istoty sztucznej inteligencji.[5] Pod literą A umieścił tzw. silną sztuczną inteligencję, której tłumaczenie z języka angielskiego na język polski nie oddaje jej istoty, zawierającej się w twierdzeniu, że umysł to program dla maszyny cyfrowej, który ma za zadanie, przy dostatecznym stopniu swojego zaawansowania,

---

[1] Bielecki A., Teoria i Zastosowanie Sztucznej Inteligencji, s.2-3
[2] Marciszewski W., Sztuczna Inteligencja, Znak, Warszawa, 1998, s.9
[3] https://www.sciencedaily.com/terms/artificial_intelligence.html
[4] Wawrzyński P, Podstawy Sztucznej Inteligencji, Oficyna Wydawnicza Politechniki Warszawskiej, 2015, s.1
[5] Penrose R., Shadows of the Mind, Oxford University Press, Oxford, 1994, s.12

wytwarzać świadomość.[6] Osoby przyjmujące podejście operacjonizmu oraz założenie, że Wszechświat jest gigantycznym komputerem mogą osiągnąć to stanowisko.[7]

Podejście B charakteryzuje się bardzo optymistycznymi ocenami dotyczącymi możliwości intelektualnych sztucznej inteligencji, mianowicie szansy na przewyższenie w tym zakresie człowieka. Należy jednak zaznaczyć, że stosując to podejście winno się wykluczyć możliwość wytworzenia przez egzemplarze sztucznej inteligencji świadomości, która, jak się okazuje, sama w sobie jest jednak niepotrzebna przy przejawianiu inteligentnych zachowań. Istotą podejścia B jest symulacja, dzięki której podmiot działania może być poddawany takim samym wyzwaniom jak w prawdziwej sytuacji (np. podczas symulacji lotu samolotem sprawdzane są te same umiejętności pilota). Nacisk na symulację w podejściu B, która czerpie z operacjonizmu, sprawia, że podejścia A i B wyrastają na tym samym gruncie traktowania myślenia jako procesu zachodzącego w układzie fizycznym oraz całkowicie algorytmicznego.[8] Pojawia się jednak znacząca różnica, do której nawiązuje Michał Heller, według niego przejawiająca się w interpretacji zdanego przez maszynę testu Turinga: podejście A – przypisanie maszynie świadomości, podejście B – mimo przejścia testu brak u maszyny świadomości.

Wyróżnikiem podejścia C jest zanegowanie podejścia algorytmicznego w zakresie symulacji świadomości. Co prawda świadomość jest wytworem ludzkiego mózgu, jednak procesy stojące za jej powstaniem nie mogą zostać wyjaśnione na bazie obliczeniowej. Wytłumaczeniem takiego stanowiska może być wiara w istnienie praw przyrody, których badaniem zajmują się do tej pory fragmentarycznie poznane działy fizyki (np. teoria chaosu), a które to prawa stoją za powstawaniem świadomości.[9]

Ostatnim nazwanym i wyeksplikowanym przez Penrose'a podejściem jest D, stojące w kontraście z podejściem A i jego algorytmami tworzącymi świadomość, a postulujące, że świadomość nie może zostać wyjaśniona ani w kategoriach fizykalnych ani obliczeniowych. Udział w stworzeniu tego stanowiska miał *post mortem* austriacki logik i matematyk Kurt Friedrich Gödel, który zasłynął ze sformułowania twierdzenia o niezupełności. Wnioskował on, że stanowisko D (wtedy jeszcze pod tą nazwą niewystępujące) jest słuszne, ponieważ, skoro mózg jest maszyną podlegającą prawom algorytmicznym oraz nie istnieje algorytm dowodzenia twierdzeń w sposób arytmetyczny (przy założeniu jego niesprzeczności)[10], świadomość musi brać się z jakiejś innej sfery niż fizykalna czy obliczeniowa.

Mimo, iż dalsza część pracy poświęcona jest zasadom etycznym egzemplarzy sztucznej inteligencji, warto przed podjęciem tego działania nawiązać do współczesnych definicji inteligencji ludzkiej w celu podkreślenia jej znaczenia w tworzeniu podstaw społeczeństwa. Celem pracy jest także klarowne zdefiniowanie już na samym wstępie pojęć, które wydają się ważne dla dalszego wywodu. Nie bez przyczyny nawiązać tu można do błędu popełnionego w prawie 1000-stronicowej pracy Guiforda „Natura Inteligencji Człowieka", w której badane zjawisko nie zostało ani razu *explicite* zdefiniowane.[11]

---

[6] Marciszewski W.: op.cit., s.16
[7] Penrose R.: op.cit., s.13
[8] Marciszewski W.: op. cit., s.17 -18
[9] Penrose R.: op. cit., s.14
[10] Gödel K.F., Über formal unentscheidbare Sätze der „Principia Mathematica" und verwandter Systeme I
[11] Guilford J.P., Natura Inteligencji Człowieka, PWN, Warszawa, 1978

Definicja łacińskiego słowa nakierowuje na współczesne rozumienie terminu inteligencja: „zdolność pojmowania, rozumienie, rozum, przenikliwość, bystrość."[12] Obecnie w literaturze można spotkać przynajmniej trzy znaczenia kryjące się pod słowem zdolność (jako składowa definicji inteligencji). Pierwsze z nich odnosi się do potencjalnych możliwości danej jednostki pod warunkiem spełnienia optymalnych warunków egzo- i endogenicznych. Drugie znaczenie kładzie nacisk na rzeczywiste i faktyczne uzdolnienie badanej jednostki, natomiast w trzecim należy odnieść się do rzeczywistego poziomu wykonania zadań, który został zmierzony przez pewien test.[13] Ciekawą i pasującą do tego wywodu definicją (aczkolwiek zbyt ogólną w innych przypadkach) jest ta zaproponowana przez Lindę Gottfredson: „Inteligencja jest bardzo ogólną zdolnością umysłową, która między innymi obejmuje umiejętność rozumowania, planowania, rozwiązywania problemów, myślenia abstrakcyjnego, rozumienia złożonych kwestii, szybkiego uczenia się oraz uczenia się na podstawie osobistego doświadczenia."[14] Popularna w niektórych kręgach definicja D.O. Hebba[15,16] (podział na podobną do genotypu inteligencję wrodzoną oraz podobne do fenotypu przejawianie inteligentnych zachowań) wydaje się wprowadzać zbyt dużą szczegółowość, niepotrzebną w dalszym wywodzie.

Zestawienie twierdzenia Gödla oraz powyższej, jakkolwiek ogólnej, definicji inteligencji prowadzi do wniosku, że traktując sztuczną inteligencję jako wcielenie maszyny Turinga, badacze mogą od razu porzucić marzenie o stworzeniu sztucznej inteligencji na wzór ludzki.[17] Mimo zastrzeżeń innych badaczy co do powyższego wnioskowania, wolno stwierdzić, że w przewidywalnej przyszłości problem nauki postępowania etycznego będzie dotyczył egzemplarzy sztucznej inteligencji przeznaczonej do pracy w konkretnej dziedzinie (w przeciwieństwie do ogólnej sztucznej inteligencji - Artificial General Intelligence).[18]

Badania i rozważania nad moralnością mogą być prowadzone na dwa sposoby. Stosując pierwszy z nich badacz oddziela się niejako od oceny przedmiotu badania słowami dobry i zły; celem badacza jest stworzenie opisu co dana społeczność uważa za słuszne moralnie. Drugie podejście zakłada zastosowanie oceny dobry/zły do napotkanych zachowań i stworzenie reguł postępowania dla ogółu.[19] W celu dokładniejszego opisu problemu poruszanego w tej pracy należy także odwołać się do słowa moralny i wybrać jedno z jego wielu znaczeń. W dalszej części pracy moralny będzie słowem neutralnym, czyli pozbawionym elementu pochwały czy nagany, nawiązującym do uczuć moralnych.[20]

Jak wynika z powyższego fragmentu, istotną rolę w formułowaniu ocen moralnych odgrywa poprawne użycie słów dobry i zły. Według literatury można się do nich odwoływać na trzech płaszczyznach: znaczenia neutralnego, opisowego i oceniającego.[21] Rozważając wspomnianą parę słów w ostatniej płaszczyźnie, milcząco implikuje się „podwójną relatywizację: 1. coś

---

[12] Pieńkos J., Słownik łacińsko-polski, Gdańsk 1996, s. 432
[13] Strelau J., Inteligencja Człowieka, Wydawnictwo "Żak", Warszawa 1997, s. 22
[14] Gottfredson L.S., Mainstream science on intelligence: An editorial with 52 signatories, w: Wall Street Journal 13.02.1994
[15] Strelau J.: op.cit., s.18
[16] Hebb D.O., The Organization of Behavior, John Wiley & Sons, 1949
[17] Lucas J. R., Minds, machines, and Gödel, Philosophy, z. 36, 1961
[18] Norvig P, S.J. Russell, Artificial Intelligence: A Modern Approach, Prentice Hall, 2016, s.1024
[19] Ossowska M., Podstawy Nauki o Moralności, t. I, De Agostini, Warszawa 2004, s. 59 - 60
[20] Ibid., s. 61
[21] Ibid., s.138-151

jest złe, dobre ze względu na ów wzór[22], z którym to coś zestawiamy; 2. jest ono równocześnie złe, dobre czy lepsze w porównaniu z innymi reprezentantami tej samej klasy."

Etyce jako nauce nad moralnością przypisuje się ogromne znaczenie w zakresie badań nad możliwością przejawiania przez sztuczną inteligencję zachowań sprowadzających się do oceny wyrażonej za pomocą słów „złe" i „dobre". Wielką grupę zagadnień omawianych w ramach etyki można skategoryzować jako postępowanie ludzkie, jego motywy oraz dyspozycje, których skutkiem są pewnie zachowania[23]. Obiektywne stwierdzenie, że dane zachowanie jest dobre lub złe nie sprawia ludziom zbyt dużych trudności. Jednak, o czym będzie mowa w dalszej części pracy, klasyfikowanie tych ocen według pewnych reguł nastręcza ogromnych trudności osobom, które intuicyjnie mogą ocenić moralność (tu wyjątkowo: słuszność) danego, pojedynczego zachowania, ale nie są w stanie wykazać racji stojących za ich decyzją.[24] W tym miejscu zarysowuje się tytułowy problem referatu, a mianowicie jak nauczyć okazy sztucznej inteligencji klasyfikowania zachowań w kategoriach moralności tychże czynów.

Przydatne w świetle przedstawionych do tej pory trudności z definiowaniem dobrych i złych zachowań jest wprowadzenie pojęcia relatywizmu moralnego, którego istota polega na występowaniu znaczącej różnicy w normach moralnych różnych epok historycznych (relatywizm historyczny) i/lub różnych kultur, społeczności (relatywizm kulturowy)[25] i/lub przedstawicieli różnych zawodów (relatywizm socjologiczny).[26] Relatywizm moralny pod postacią relatywizmu metaetycznego może także nawiązywać do przekonania, że nie ma systemu nadrzędnego względem innych.[27] Ogólnie rzecz ujmując, relatywizm stoi w opozycji do stanowiska nakazującego uznanie jednego systemu norm, a co za tym idzie bezkompromisowość w stosunku do zasad postępowania moralnego.[28] To bezkompromisowe stanowisko można określić jako absolutyzm (lub rygoryzm) moralny. Z jednej strony wydaje się stanowić odpowiednie i wygodne podejście do procesu nauczania sztucznej inteligencji zasad etyki, ponieważ z góry wiadomo jakich reguł egzemplarze mają przestrzegać. Wadą tego podejścia jest fakt, że pogląd, iż coś, co jest dobre nie według jakiegoś standardu, lecz z samej racji posiadania właściwości bycia dobrym sprawia wrażenie absurdu.[29] Z punktu widzenia omawianego w referacie zagadnienia istotnym dla rozwiązania problemu nauki okazów sztucznej inteligencji etyki wydaje się być relatywizm kulturowy oraz relatywizm socjologiczny. Oba z nich mogą odegrać znaczącą rolę w projektowaniu systemów etycznych przeznaczonych na potrzeby sztucznej inteligencji.[30, 31, 32, 33, 34]

---

[22] Przykładowo w przypadku dobrego lekarza, stosując słowo „dobry", podciągamy go pod wzór lekarza.
[23] Ossowska M., Podstawy Nauki o Moralności, t. II, De Agostini, Warszawa 2004, s. 674
[24] Helm L., Muehlhauser L., The Singularity and Machine Ethics, w: Singularity Hypotheses, Springer 2012 , s. 11
[25] Rachel J., The Elements of Moral Philosophy, McGraw-Hill, New York 2002, s.18-19
[26] Brzeziński T., Etyka lekarska, PZWL, Warszawa 2012, s.2-3
[27] Harman G., What is moral relativism w: Philosophical Studies Series in Philosophy, z. 13 , s. 143
[28] Pospiszyl I., Patologie Społeczne, PWN Warszawa 2008, s. 19
[29] The Oxford Handbook of Ethical Theory, red. Copp D., Oxford University Press, Oxford 2006, s. 242
[30] Yampolskiy R.V., Artificial Intelligence Safety Engineering: Why Machine Ethics Is a Wrong Approach, w: Philosophy and Theory of Artificial Intelligence, Springer-Verlag, Berlin Heidelberg 2013, s.389
[31] Machine Ethics, red. Anderson M., Anderson S.L., Cambridge University Press 2011, s.9
[32] Machine Ethics, red. Anderson M., Anderson S.L., Cambridge University Press 2011, s.83

Opis (nie)moralnych zachowań egzemplarzy sztucznej inteligencji podlega dziedzinie etyki robotów (ang. machine ethics)[35], która ma na celu zapewnienie, że podmioty myślące w sposób sztuczny i działające w pewnej społeczności będą zachowywać się zgodnie z normami moralnymi obowiązującymi w tej społeczności.[36] Postępowanie etyczne (w neutralnym znaczeniu tego słowa) musi zostać przedstawione w postaci obliczeniowej, aby podmioty nieludzkie były w stanie dorównać lub przewyższyć ludzi w posługiwaniu się etyką.[37] Etyka robotów pomaga także udzielić odpowiedzi na pytania: „Jak stworzyć komputer (maszynę cyfrową), który będzie w stanie rozróżnić dobro od zła, promować to pierwsze i powstrzymywać się od czynienia tego drugiego?"[38] oraz w przypadku odpowiedzi przeczącej na to pytanie, „Czy maszyna cyfrowa mogłaby przynajmniej pozorować zachowanie wskazujące na jej etyczne deliberacje?"[39] Omawiając temat etyki robotów warto także zwrócić uwagę na przestrogę, którą wprost nazywa R.W. Picard w swojej książce „Affective Computing": Im więcej wolności przyznamy maszynom [cyfrowym], tym bardziej potrzebować będą standardów etycznych."[40]

Jednym z warunków skutecznego nauczania zasad postępowania etycznego egzemplarzy sztucznej inteligencji jest zaznajomienie się z pojęciami moralnej sprawczości oraz *moral patiency*[41]. Aby lepiej zrozumieć istotę tych pojęć należy rozbić je na części składowe: moralny, sprawczość, *patiency*. Ze względu na dokonane już omówienie zagadnienia moralności oraz nawiązania do konotacji słowa moralny, przejdę do dwóch kolejnych składowych. Każdy, kto jest zdolny do odczuwania stanów intencjonalnych, może być zaliczony do kategorii podmiotów sprawczych.[42] Okazuje się, że po dziś dzień sprawczością charakteryzują się jedynie ludzie i zwierzęta. Doprecyzowując sprawczość o element moralności i otrzymując w ten sposób moralną sprawczość, należy stwierdzić, że jest ona tym, co jest niezbędne do stawiania ocen moralnych.[43] Koniecznym jest także wyróżnienie *moral patients*[44], czyli tych obiektów i istot, wobec których podmioty sprawcze posiadają pewne obowiązki.[45]

---

[33] Machine Ethics, red. Anderson M., Anderson S.L., Cambridge University Press 2011, s.165
[34] Bostrom N., The Ethics of Artificial Intelligence , s. 13
[35] Należy zwrócić uwagę, że polskie tłumaczenie nie w pełni oddaje znaczenie angielskiego terminu, który odwołuje się do całościowej dziedziny sztucznej inteligencji, a nie tylko pewnych jego elementów (robotów) jak jego polskie tłumaczenie.
[36] Anderson M., Anderson S.L., Machine Ethics: Creating an Ethical Intelligent Agent, w: AI Magazine z. 28(4), 2007, s.15
[37] Ibid., s.16
[38] Powers T.M., Deontological Machine Ethics, w: Association for the Advancement of Artificial Intelligence Fall Symposium Technical Report, 2005, s.1
[39] Powers T.M., Prospects for Kantian Machine, w: IEEE Intelligent Systems z. 21(4), lipiec-sierpień 2006, s. 46
[40] Picard R.W., Affective Computing, The MIT Press, Cambridge 1997, s.134: „The greater the freedom of a machine, the more it will need moral standards."
[41] Zdecydowałem się na pozostawienie tego terminu w języku angielskim, ponieważ w języku polskim nie są dostępne żadne szeroko uznane tłumaczenia. W mojej opinii, autorskie „moralne zobowiązanie" nie oddaje istoty całego zagadnienia kryjącego się pod angielskim odpowiednikiem.
[42] Gunkel D.J., The Machine Question – Critical Perspective on AI, Robots and Ethics, The MIT Press, Cambridge 2012, s. 20-21
[43] The Oxford Handbook…: op.cit., s. 18
[44] Autorskie tłumaczenia „moralny obligatariusz" oraz „moralny wierzyciel" w pełni nie oddają znaczenia angielskiego terminu.
[45] Gunkel D.J.: op. cit., s. 93

Przydatnym w dalszych rozważaniach będzie dokonanie dalszego podziału moralnych podmiotów sprawczych. Zachowanie tych podmiotów sprawczych, które działają *explicite* (ang. *explicit moral agents*) w kategoriach związanych z etyką można przyrównać do zachowania Deep Blue oraz Deeper Blue[46] w dziedzinie szachów: *explicit moral agents* stosują zasady etyczne takie imperatyw kategoryczny Kanta w celu ustalenia swoich dalszych czynności (analogia do ruchów w grze w szachy). Należy podkreślić, że *explicit moral agents* niezależnie podejmują decyzję w zgodzie z predefiniowanymi zasadami oraz na podstawie tych zasad mogą podać uzasadnienie swojego działania.[47] Twierdząca odpowiedź na pytanie dotyczące istnienia takich podmiotów wskazuje na to, że etyka może istnieć w maszynach cyfrowych i być w nich w pewien sposób reprezentowana.[48]

Do mniej zaawansowanej pod względem związku z etyką kategorii inteligentnych maszyn cyfrowych należą *implicit moral agents*, które są domyślnie ograniczone do unikania nieetycznych zachowań, nie wykazując przy tym umiejętności do wnioskowania na podstawie wcześniej zdefiniowanych pryncypiów etycznych.[49] Badacze wątpią w możliwości *implicit moral agents* w zakresie przeprowadzania rozumowania etycznego w kompletnie dla nich nowych sytuacjach i dziedzinach.[50] Przykładem *implicit moral agents* są bankomaty oraz autopiloci.[51]

Kategorią stojącą w ostrym kontraście z implicit moral agent jest podmiot o pełnej sprawczości moralnej (ang. *full moral agent*), który charakteryzuje się świadomością, sprawczością oraz wolną wolą[52] i z tego powodu może być pociągany do odpowiedzialności za swoje czyny. Jedynym do tej pory znanymi przykładem *full moral agents* są ludzie, którzy nie zostali ubezwłasnowolnieni na drodze prawnej. Badacze spierają się czy stworzenie egzemplarza sztucznej inteligencji z taką właśnie charakterystyką pełnej sprawczości moralnej jest słuszne z moralnego i/lub racjonalnego punktu widzenia.[53]

W celu uniknięcia nieporozumień dotyczących niezależnych decyzji oraz autonomii podmiotów je podejmujących należy zwrócić uwagę na rozróżnienie funkcjonalnej i moralnej niezależności (autonomii), którą tłumaczy etyk wojskowy George Lucas, Jr. następującymi słowami: odkurzacz Roomba oraz rakieta Patriot są niezależne w sensie elastyczności podejmowanych decyzji w warunkach, których zmiana nie wymaga ludzkiej interwencji w zmodyfikowany z kolei proces decyzyjny; nie są one jednak niezależne w sensie moralnym, ponieważ nie mogą przerwać swojego działania w przypadku posiadania moralnych obiekcji.[54]

---

[46] http://theconversation.com/twenty-years-on-from-deep-blue-vs-kasparov-how-a-chess-match-started-the-big-data-revolution-76882
[47] Van Rysewyk S.P., Pontier M., Machine Medical Ethics, Springer 2014, s.11
[48] Moor J.H., Is Ethics Computable, w: Metaphilosophy, z. 26(1-2), 1995, s.1-21
[49] Van Rysewyk S.P., Pontier M.: op.cit., s. 11
[50] Yueh-Hsuan W., Chien-Hsun C., Chuen-Tsai S., Toward The Human-Robot Coexistence Society: On safety Intelligence for Next Generation Robots, s. 5-6
[51] Moor J.H., The Nature, Importance, and Difficulty of Machine Ethics, w: IEEE Intelligent Systems z. 21(4), lipiec-sierpień 2006, s. 19
[52] Van Rysewyk S.P., Pontier M.: op.cit., s. 11
[53] Van Rysewyk S.P., Pontier M.: op.cit., s.11
[54] Etzioni A., Etzioni O., Incorporating Ethics into Artificial Intelligence, w: The Journal of Ethics, 2017, s.7

W tym miejscu warto nawiązać do koniecznych charakterystyk jakimi powinna legitymować się sztuczna inteligencja. W badaniach[55] zwraca się uwagę transparentność, przewidywalność wyników działania egzemplarzy oraz odporność na manipulację. Odpowiednie rozwiązanie postawionego na wstępie tej pracy problemu nie może odbiegać od wymienionych wyżej charakterystyk bezpiecznego okazu sztucznej inteligencji. Ponadto każde z proponowanych poniżej rozwiązań, aby zostać przynajmniej wstępnie zaakceptowanym, musi spełniać warunek spójności, polegającej na tym, że algorytm, który otrzymał dwa lub więcej razy ten sam zestaw danych wejściowych za każdym razem musi zwrócić ten sam wynik.[56] Ponadto kompletność algorytmu, czyli zwracanie poprawnego rozwiązania dla wszystkich poprawnych zestawów danych wejściowych, odgrywa znaczącą rolę w selekcji rozwiązania tytułowego problemu.[57]

Po przeanalizowaniu i wyeksplikowaniu wszystkich składowych niezbędnych do zrozumienia etyki robotów można przejść do klasyfikacji i opisu poszczególnych metod implementacji kodeksów etycznych. Istnieją dwa nadrzędne podejścia do tytułowego problemu: oddolne (bottom-up) oraz odgórne (top-down) . Pierwsze z nich zasadza się na założeniu, że podmioty nie reprezentują *explicite* swoich zasad moralnych, lecz postępują według nich tak jakby były one *implicite* zawarte w każdym ich czynie.[58]

Nietrywialną charakterystyką podejścia oddolnego jest niewątpliwie niezachwiana zdolność do bieżącego uwzględniania materiału wejściowego pochodzącego z różnych środowisk i sytuacji.[59] Wadą tego podejścia jest natomiast nieznajomość lub niepewność co do celów za pomocą, których należy określić skuteczność danego systemu oddolnego. Należy, bowiem w tym miejscu zaznaczyć, że podczas projektowania systemów oddolnych, działanie każdego z nich podlega ocenie wystawianej na podstawie spełnienia jednego lub większej liczby celów. W przypadku zbyt dużej liczby celów, których spełnienie stawia się systemowi, może dojść do rozbieżności skutków, jakie poszczególne rozwiązania (każde optymalne w stosunku do innego celu) wywierają na środowisko zewnętrzne.[60]

Sztandarowym ucieleśnieniem podejścia oddolnego w praktyce jest wykorzystanie uczenia maszynowego[61] oraz zastosowanie zasad kazuistyki, polegającej na wnioskowaniu przez analogię do problemów moralnych z uznanymi rozwiązaniami w celu rozwiązaniu nowych przypadków (kazusów).[62] Przykładem zastosowania bardziej zaawansowanych rozwiązań jest podejście wykorzystujące sieci neuronowe[63], które w jednym z badań, mających na celu rozstrzygnięcie sporu dotyczącego zasadności partykularyzmu (moralnego)[64], zostały użyte

---

[55] Bostrom N., The Ethics of Artificial Intelligence, s. 2-3
[56] Anderson M., Anderson S.L., Armen C., Towards Machine Ethics, s. 2
[57] Anderson M., Anderson S.L., Armen C.: op. cit., s. 3
[58] Wallach W., Allen C., Smit I., Machine Morality: Bottom-up and Top-down Approaches for Modeling Human Moral Faculties, s. 3
[59] Wallach W., Allen C., Smit I.: op. cit., s. 4
[60] Helm L., Muehlhauser L.: op.cit., s. 4
[61] Samodoskonalenie się systemów sztucznej inteligencji przy pomocy zgromadzonych doświadczeń (danych) i nabywanie na tej podstawie wiedzy.
[62] The Oxford Handbook…: op.cit., s. 627
[63] Struktura matematyczno-informatyczna wykonująca obliczenia lub przetwarzająca sygnały poprzez rzędy elementów zwanych sztucznymi neuronami.
[64] The Oxford Handbook…: op.cit., s. 627

do wykazania, że mimo przydatności partykularnego toku rozumowania pewien stopień generalizacji (tu: użycia zasad etycznych odgórnie) jest niezbędne.[65]

Cechą charakterystyczną drugiego podejścia jest korzystanie z uprzednio wyspecyfikowanych teorii etycznych (zarówno tych będących dziełem filozofów jak i tych zawartych w świętych księgach), z których otrzymuje się nakazy, zakazy oraz powinności w konkretnej sytuacji.[66] Do chwili obecnej podejście to było bardzo mało popularne wśród praktyków zajmujących się etyką robotów.[67] Odgórne podejście do problemu nauczania zasad moralnych ma jedną niewątpliwą zaletę – pozwala uniknąć czasochłonnych ludzkich rozważań na wzór danego systemu etycznego (np. imperatywu kategorycznego Kanta) i zastąpić je ogromną mocą obliczeniową maszyny cyfrowej.[68] Refleksja nad wadami natomiast wprowadza istotę historii etyki, czyli ciągłego konfliktu pomiędzy poszczególnymi teoriami, polegającego na dialektycznym wytykaniu niedociągnięć każdego z nich.[69]

Pierwszy z analizowanych systemów etycznych, które mogłoby stanowić rozwiązanie tytułowego problemu jest etyka kantowska. Jej wykładnię stanowi kilka prac pruskiego filozofa, jednak najlepiej jest ona wyartykułowana w „Krytyce Czystego Rozumu". Filozof zwraca uwagę na to, że powinność moralna nie wywodzi się ani od Boga, ani ludzkich ustanowień, władz i społeczności, ani preferencji i pragnień, lecz z rozumu[70], co stanowi o szansie przełożenia tej postawy etycznej na grunt obliczeniowy za pomocą symulacji racjonalnego podejmowania decyzji.[71] Kant stawia warunek, że tezy poznawcze, do których obrony przystępujemy, muszą dotyczyć rzeczywistości dostępnej ludzkiemu doświadczeniu. Ważnym elementem tego kantowskiego podejścia etycznego jest odwoływanie się do pytania „Co powinienem czynić?", co prowadzi do określenia maksym, czyli zasad działania, które powinny zostać przyjęte.[72] Typową cechą maksym w rozumieniu Kanta jest to, że są to zasady, które mogłaby przyjąć pewna zbiorowość bez zakładania czegokolwiek na temat „pragnień podmiotów działających i stosunków społecznych.[73] Jednym z najbardziej znanych konstruktów etycznych związanych z pruskim filozofem jest imperatyw kategoryczny (prawo moralne), które głosi: „Postępuj tylko według takiej maksymy, dzięki której możesz zarazem chcieć, aby stała się prawem moralnym."[74] Kluczowym dla demoralizacji przeciwników rozwoju sztucznej inteligencji elementem etycznej filozofii Kanta jest „formuła celu samego w sobie" zgodnie, z którą „człowieczeństwo zarówno we własnej osobie, jak i w osobie każdego innego zawsze [uznawane] zarazem jako cel, nigdy jako środek."

Jak każda postawa etyczna, tak i ta pociąga za sobą falę krytyki, która, w tym przypadku, kładzie nacisk na zbytni formalizm, rygoryzm i abstrakcyjność.[75] Podmiot stosujący zalecenia etyki kantowskiej *in verbatim* jest pozbawiony możliwości uwzględniania różnić pomiędzy poszczególnymi przypadkami. Kolejny zarzut dotyczy oparcia teorii Kanta na analizie

---

[65] Guarini M., Particularism and the Classification and Reclassification of Moral Cases, w: IEEE Intelligent Systems, z. 21(4), lipiec-sierpień 2006
[66] Wallach W., Allen C., Smit I.: op. cit., s. 2
[67] Wallach W., Allen C., Smit I.: op. cit., s. 4-5
[68] Wallach W., Allen C., Smit I.: op. cit., s. 5
[69] Etzioni A., Etzioni O.: op. cit., s. 4
[70] Przewodnik po Etyce, red. Peter Singer, Wydawnictwo "Książka i Wiedza", Warszawa 2000, s.214
[71] Powers T.M., Prospects for Kantian Machine…
[72] The Oxford Handbook…: op.cit., s. 482
[73] Przewodnik po Etyce: op. cit., s. 216
[74] The Oxford Handbook…: op.cit., s. 489
[75] Przewodnik po Etyce: op. cit., s. 222

przypadków granicznych, która w żaden sposób nie zapewnia użytkownikom algorytmu moralnego postępowania, pozostawiając ich niejako samym sobie w dokonywaniu wyborów nienaruszających żadnych zakazów i spełniających wszystkie powinności.

Drugim systemem etycznym, na podstawie którego próbuje się stworzyć system etyczny dla sztucznej inteligencji jest deontologia. Jej główne założenie można wyłożyć w postaci pewnych rodzajów działań, które będąc samymi w sobie złe, nie mogą stanowić etapu przejściowego w dążeniu do jakichkolwiek celów (nawet tych godnych pochwały czy moralnie nakazanych).[76,77] Teorię deontologiczną można sformułować jaką tą, „która albo nie określa dobra niezależnie od słuszności, albo nie interpretuje słuszności jako maksymalizacji dobra…"[78] Dobrym przykładem postawy deontologicznej jest negatywny stosunek do kłamstwa w ogóle, ale nie dlatego, że może mieć ono złe skutki, lecz dlatego, że kłamanie jest po prostu niesłuszne. Omawiana postawa etyczna posługuje się rygorami, które przeważnie mają charakter negatywny i nie mogą być równoważone przez sformułowania pozytywne.[79] Inną charakterystyką norm (rygorów) deontologicznych jest istnienie ich granic, poza którymi nic nie jest zakazane.[80] Warto także zwrócić uwagę na to fakt, że wąskie ukierunkowanie rygorów prowadzi do racji deontologicznych, które „z pełną mocą zwracają się przeciwko czynieniu czegoś, a nie przeciwko zdarzeniu się tego."[81]

Problematyczny element tradycji deontologicznej stanowi zagadnienie niesłuszności pewnych czynności, które w myśl zasad omawianego systemu nie mogą być pod żadnym pozorem podejmowane. Nasuwa się pytanie według jakich kryteriów można zaliczyć czynność jako (nie)słuszną, na które można odpowiedź w różny sposób: z tradycji religijnej, z ludzkiej intuicji, z jednej z formalnych zasad (np. imperatyw kategoryczny)[82]. Nawiązując do wcześniejszych fragmentów dotyczących relatywizmu kulturowego i niezdolności ludzi do skodyfikowania ultymatywnych zasad moralnych[83] dwie pierwsze odpowiedzi mogą zostać odrzucone jako nieprzydatne, natomiast imperatyw kategoryczny sprawia pewne, już nazwane, problemy.

Dwudziestowieczny brytyjski filozof W.D. Ross sformułował na podstawie swoich obserwacji dotyczących niepoprawności niektórych form monizmu etycznego teorię obowiązków *prima facie,* która ani nie zakłada, że któreś obowiązki muszą podlegać priorytetyzacji ani, że nie sugeruje, jakoby proponowane zasady miały być ze sobą spójne.[84] W sytuacjach, w których dochodzi do konfliktu pomiędzy obowiązkami ciążącymi na podmiocie, musi on na podstawie swojego osądu (i najwyraźniej bez pomocy żadnej teorii) znaleźć punkt równowagi. Sytuacja w jakiej stoi ten podmiot nie wydaje się łatwa – jak sam W.D. Ross postulował: „(…)ważniejsze jest, żeby nasza teoria zgadzała się z faktami, niż żeby była prosta."[85] Pocieszającym jest fakt, że sam konflikt pomiędzy dwoma obowiązkami wskazuje na ich istotność. Według Ross'a nie istnieje jedna wersja hierarchii wartości

---

[76] Przewodnik po Etyce: op. cit., s. 247
[77] The Oxford Handbook…: op.cit., s. 427
[78] Rawls J., Teoria Sprawiedliwości, PWN, Warszawa 1994, s. 46
[79] Przewodnik po Etyce: op. cit., s. 250
[80] Fried C., Right and Wrong, , Harvard University Press Cambridge, 1978, s.9-10
[81] Nagel. T., The View from Nowhere, Oxford University Press, New York, 1986, s. 177
[82] Przewodnik po Etyce: op. cit., s. 254
[83] Road Vehicle Automation, red. Meyer G., Beiker S., Springer International Publishing 2014, s.7
[84] Przewodnik po Etyce: op. cit., s. 261
[85] Ross W.D., The Right and the Good, Clarendon Press, Oxford, 1930, s. 19

moralnych, jednak analiza jego pracy wskazuje na siedem obowiązków *prima facie*: wierność, zadośćuczynienie, wdzięczność, sprawiedliwość, dobroczynność, godziwość, samodoskonalenie.[86] Okazuje się, że główną przeszkodą na drodze uniwersalnego zastosowania strategii Ross'a stanowi mnogość układów równań, które opisywałaby możliwą do podjęcia czynność.[87, 88]

Ostatnim omawianym stanowiskiem etycznym jest konsekwencjalizm, który zaleca jako najlepszą reakcję na wyznawane wartości ich propagowanie.[89] Kluczowymi pojęciami dla teorii teleologicznych są opcja oraz prognoza, które oznaczają, odpowiednio, dającą się urzeczywistnić możliwość oraz porządek świata związany z daną opcją. W każdym wyborze podmiot powinien kierować się tym, która opcja daje największe szanse realizacji związanej z nią prognozy.[90] Należy zwrócić uwagę, że podejście to wydaje się niepraktyczne, ponieważ przewidzenie wszystkich konsekwencji danego czynu może być albo bardzo czasochłonne (potrzebny czas przewyższa czas dostępny dla decydenta) albo niemożliwe. Głównym zarzutem kierowanym w stronę analizowanego systemu jest zawarta w nim bierna zachęta do rozważenia wyboru opcji z okropnymi, (w innych systemach) niemoralnymi czynami, tylko dlatego, że opcje te mogą charakteryzować się najwyższą opłacalnością dla podmiotu i ogółu.[91]

Szczególny przypadek konsekwencjalizmu stanowi utylitaryzm, opierający się na zasadzie, że jeśli jakaś rzecz ma być dobra, to musi być dobra dla kogoś, czy to dla jednostki czy ogółu. Jak sama nazwa wskazuje w tym systemie podmioty stojące przed wyborem pomiędzy pewnymi czynami odwołują się do pojęcia użyteczności, która okazuje się bardzo szerokim i niejasnym podejściem. W krytykowanym podejściu hedonistycznym Benthama użyteczność to to samo, co przydatność w zakresie promowania przyjemności i unikania cierpienia.[92] Można także mówić o utylitaryzmie dobrobytu, sprowadzającym się do zaspokajania interesów i stanowiącym pewien rodzaj zabezpieczenia przed przekładaniem krótkoterminowego i krótkowzrocznego zaspokajania preferencji na rzecz ochrony długofalowych interesów podmiotów.

Abstrahując od typu użyteczności, palącym problemem podejścia utylitarystycznego jest kwestia jej maksymalizacji. Uznając za działanie słuszne te z największą liczbą jednostek użyteczności podmiot decyzyjny natrafia na problem sumowania poszczególnych składników końcowego wyniku. Aby osiągnąć jednoznaczną decyzję musi założyć porównywalność dóbr oraz ludzi, co zwłaszcza w ostatnim przypadku jest kwestią podlegającą burzliwej dyskusji.[93]

Z perspektywy *exemplum* sztucznej inteligencji stosującego algorytmy i przekształcenia liczbowe podejście utylitaryzmu wydaje się *prima facie* najskuteczniejszym. O jego atrakcyjności stanowi fakt, że maszyny cyfrowe, w przeciwieństwie do ludzi[94], rzeczywiście

---

[86] Ross W.D.: op. cit., s. 16-47
[87] Anderson M., Anderson S.L., Armen C., s. 3
[88] Naprzeciw wychodzi moc obliczeniowa dzisiejszych komputerów, która stanowi podstawę projektu W.D. Advises kierowanego przez M.Anderson, S.L.Anderson ora C.Armen i skutecznie doradzającego przy prostych wyborach moralnych.
[89] Przewodnik po Etyce: op. cit., s. 273
[90] The Oxford Handbook…: op. cit., s. 383
[91] Przewodnik po Etyce: op. cit., s. 276
[92] Przewodnik po Etyce: op. cit., s. 284
[93] Przewodnik po Etyce: op. cit., s. 287
[94] Anderson M., Anderson S.L., Machine Ethics: Creating an Ethical Intelligent Agent: op. cit., s. 18

przeprowadzają obliczenia. Warto w tym miejscu zwrócić uwagę na zachwianą (dla niektórych osób) sprawiedliwość. W myśl utylitaryzmu o byciu „dobrym" i „złym" decyduje przyszła użyteczność, natomiast obecny stan, który jest oceniany w tym właśnie paradygmacie powstał na bazie przeszłych zdarzeń. Jak wykazano w jednym z badań, kierowcy samochodów preferują utylitaryzm w stosunku do innych systemów etycznych jedynie w sytuacjach, gdy sami nie są uczestnikami zdarzenia – w przeciwnym razie preferują system priorytetyzujący ich własne dobro.[95]

Analiza obecnego stanu badań nad implementacją systemów etycznych w sztucznej inteligencji wskazuje na możliwe korzyści z wynikające z połączenia podejścia oddolnego i odgórnego. Jedną ze wskazówek dotyczących tytułowego problemu jest sposób w jaki ludzie mogą przyswajać umiejętności w zakresie etyki: najpierw otrzymują pewien zestaw predefiniowanych zasad, których muszą się trzymać jako dzieci (element podejścia odgórnego), a następnie stając się coraz bardziej samodzielnymi zaczynają je modyfikować na podstawie własnych doświadczeń i wynikających z nich przekonań (element podejścia oddolnego).[96] Biorąc pod uwagę dwie oddzielne kategorie pojęć najbardziej obiecującym wydaje się podejście oparte na obowiązkach *prima facie* W.D. Ross'a. Warto także zastanowić się nad rolą sztucznej inteligencji we współczesnym społeczeństwie. Czy pełna autonomia opisana we wstępie pod postacią postaw A, B, C, D jest konieczna? Czy sztuczna inteligencja powinna być traktowana raczej jako pomocnik człowieka? W drugim przypadku pełna odpowiedzialność za czyny cyfrowego partnera spadłaby na właściciela, użytkownika lub producenta eliminując w pełni problem etyki robotów.[97]

[95] Bonnefon J., Shariff A., Rahwan I., The Social Dilemma of Autonomus Vehicle, w: Science 352(6293)
[96] Etzioni A., Etzioni O.: op. cit., s. 4
[97] Etzioni A., Etzioni O.: op. cit., s. 10

Bibliografia

Anderson M., Anderson S.L., Armen C., *Towards Machine Ethics*

Anderson M., Anderson S.L., *Machine Ethics: Creating an Ethical Intelligent Agent*, w: AI Magazine z. 28(4), 2007

Bielecki A., *Teoria i Zastosowanie Sztucznej Inteligencji*

Bonnefon J., Shariff A., Rahwan I., *The Social Dilemma of Autonomus Vehicle*, w: Science 352(6293)

Bostrom N., *The Ethics of Artificial Intelligence*

Brzeziński T., *Etyka lekarska*, PZWL, Warszawa 2012

Etzioni A., Etzioni O., *Incorporating Ethics into Artificial Intelligence*, w: The Journal of Ethics, marzec 2017

Fried C., *Right and Wrong*, Harvard University Press Cambridge, 1978

Gödel K.F., *Über formal unentscheidbare Sätze der „Principia Mathematica" und verwandter Systeme I*

Gottfredson L.S., *Mainstream science on intelligence: An editorial with 52 signatories,* w: Wall Street Journal 13.02.1994

Guarini M., *Particularism and the Classification and Reclassification of Moral Cases,* w: IEEE Intelligent Systems, z. 21(4), lipiec-sierpień 2006

Guilford J.P., *Natura Inteligencji Człowieka*, PWN, Warszawa, 1978

Gunkel D.J., *The Machine Question – Critical Perspective on AI, Robots and Ethics*, The MIT Press, Cambridge 2012

Harman G., *What is moral relativism* w: Philosophical Studies Series in Philosophy, z. 13

Hebb D.O., *The Organization of Behavior*, John Wiley & Sons, 1949

Heller M., *Filozofia Sztucznej Inteligencji*, w: „Znak" 9(484)/1995

Helm L., Muehlhauser L., *The Singularity and Machine Ethics*, w: Singularity Hypotheses, Springer 2012

http://theconversation.com/twenty-years-on-from-deep-blue-vs-kasparov-how-a-chess-match-started-the-big-data-revolution-76882 [dostęp: 25.06.2017, 21:20]

https://www.sciencedaily.com/terms/artificial_intelligence.htm [dostęp: 21.06.2017, 17:50]

Lucas J. R., *Minds, machines, and Gödel*, Philosophy, z. 36, 1961

Machine Ethics, red. Anderson M., Anderson S.L., Cambridge University Press 2011

Marciszewski W., *Sztuczna Inteligencja*, Znak, Warszawa, 1998

Moor J.H., *Is Ethics Computable*, w: Metaphilosophy, z. 26(1-2), 1995

Moor J.H., *The Nature, Importance, and Difficulty of Machine Ethics*, w: IEEE Intelligent Systems z. 21(4), lipiec-sierpień 2006

Nagel. T., *The View from Nowhere*, Oxford University Press, New York, 1986

Norvig P, S.J. Russell, *Artificial Intelligence: A Modern Approach*, Prentice Hall, 2016

Ossowska M., *Podstawy Nauki o Moralności*, t. I, De Agostini, Warszawa 2004

Ossowska M., *Podstawy Nauki o Moralności*, t. II, De Agostini, Warszawa 2004

Penrose R., *Shadows of the Mind*, Oxford University Press, Oxford, 1994

Picard R.W., *Affective Computing*, The MIT Press, Cambridge 1997

Pieńkos J., *Słownik łacińsko-polski*, Gdańsk 1996

Pospiszyl I., *Patologie Społeczne*, PWN Warszawa 2008

Powers T.M., *Deontological Machine Ethics*, w: Association for the Advancement of Artificial Intelligence Fall Symposium Technical Report, 2005

Powers T.M., *Prospects for Kantian Machine*, w: IEEE Intelligent Systems z. 21(4), lipiec-sierpień 2006

Przewodnik po Etyce, red. Peter Singer, Wydawnictwo "Książka i Wiedza", Warszawa 2000

Rachel J., *The Elements of Moral Philosophy*, McGraw-Hill, New York 2002

Rawls J., *Teoria Sprawiedliwości*, PWN, Warszawa 1994

Road Vehicle Automation, red. Meyer G., Beiker S., Springer International Publishing 2014

Ross W.D., *The Right and the Good*, Clarendon Press, Oxford, 1930

Strelau J., *Inteligencja Człowieka*, Wydawnictwo "Żak", Warszawa 1997

The Oxford Handbook of Ethical Theory, red. Copp D., Oxford University Press, Oxford 2006

Van Rysewyk S.P., Pontier M., *Machine Medical Ethics,* Springer 2014]

Wallach W., Allen C., Smit I., *Machine Morality: Bottom-up and Top-down Approaches for Modeling Human Moral Faculties*

Wawrzyński P, *Podstawy Sztucznej Inteligencji*, Oficyna Wydawnicza Politechniki Warszawskiej, 2015

Yampolskiy R.V., *Artificial Intelligence Safety Engineering: Why Machine Ethics Is a Wrong Approach,* w: Philosophy and Theory of Artificial Intelligence, Springer-Verlag, Berlin Heidelberg 2013

Yueh-Hsuan W., Chien-Hsun C., Chuen-Tsai S., *Toward The Human-Robot Coexistence Society: On safety Intelligence for Next Generation Robots*