

Cybersecurity and Law

2026 Nr 1(16) Numer specjalny

DOI: 10.35467/cal/226780



Analiza etyczności relacji człowiek – maszyna inteligentna

Analysis of the Ethics of the Human – Intelligent Machine Relationship

Teresa Maria GRABIŃSKA

Akademia Wojsk Lądowych

ORCID: 0000-0002-9131-2637

E-mail: gtate13@wp.pl

Streszczenie

Wartości etyczne przekładają się na normy moralne, obowiązujące w środowisku egzystencji i działania człowieka. Zmienia się ono szybko wraz z rozwojem zintegrowanych technologii i będzie się wypełniać maszynami inteligentnymi, czyli aktywnymi przedmiotami materialnymi i niematerialnymi, wyposażonymi w sztuczną inteligencję. Relacja człowiek-maszyna staje się zwrotna. Człowiek bowiem dotychczas był podmiotem działania i użytkownikiem narzędzi technicznych, natomiast obecnie maszyna inteligentna wyposażona w rodzaj wrażliwości (*sentience*), nabywa cech humanoidalnych. W artykule podjęto dyskusję z dwunastoma problemami etycznymi relacji człowiek-maszyna (B.P. Greena) oraz z trzema transhumanistycznymi zasadami oceny moralnej współdziałania człowiek-maszyna. Okazuje się, że normy moralnych będą zastępowane normami technicznymi, które będą optymalizować sprzężenie człowiek-maszyna. Sam zaś człowiek gatunku *Homo sapiens* będzie technologicznie ulepszany (*enhanced*) i w zastępowany przez transczłowieka (*transhuman*).

Słowa kluczowe

sztuczna inteligencja, normy moralne, automatyzacja etyki, prawa robotów, transhumanizm

Abstract

Ethical values translate into moral norms that apply in the environment of human existence and action. This environment is changing rapidly with the development of integrated technologies and it will be filled with intelligent machines (IM), i.e., active material and immaterial objects equipped with artificial intelligence. The human-IM relationship is becoming reflexive. Humans have previously been the subject of action and users of technical tools, but now, MI, equipped with a kind of sentience, is acquiring humanoid characteristics. There discussed B.P. Green's twelve ethical problems in human-IM relationships and three transhumanist principles for moral evaluation of human-IM interaction. It turns out that moral norms will be replaced by technical norms

that will optimize the human-IM interface. In turn, humans, the species *Homo sapiens*, will be technologically enhanced and replaced by transhumans.

Keywords

artificial intelligence, moral norms, ethical automation, robot rights, transhumanism

Wstęp

W artykule wszelkie zagadnienia związane z etyką będą się odnosić do maszyny inteligentnej (MI), rozumianej jak aktywny przedmiot materialny lub niematerialny, obdarzony sztuczną inteligencją (SI) albo reprezentujący SI – jak algorytm. Wybranie MI członem relacji w stosunku do człowieka jest uzasadnione z trzech powodów. Po pierwsze, SI nie ma jasnego statusu bytowego¹. Termin maszyna (*machine*) jest tu używany nie tylko w znaczeniu materialnego nośnika działania lecz także jak w języku angielskim, tj. “*the controlling and inner organization*”².

Po drugie, przeprowadzana analiza dotyczy relacji etyczności człowiek – MI w perspektywie podmiotowych aspektów obu jej członów. MI nie jest wyłącznie przedmiotem użytkowania, a człowiek nie jest wyłącznie użytkownikiem. Na współczesnym poziomie działania MI wyraźnie zaznacza się jej emancypacja, która powoduje interakcję z człowiekiem.

Po trzecie, interakcja człowiek – MI pociąga za sobą odpowiedzialność (moralną i prawną) za jej skutki: powstaje problem podmiotowości moralnej MI, czyli powinności moralnych obu czynników interakcji oraz związany z nim problem podmiotowości prawnej MI, który tu jest jedynie sygnalizowany.

Na obecnym poziomie rozwoju technologii zintegrowanych³ zagadnienia etyczne związane z SI można umieścić w 3 głównych nurtach: I – wartości etyczne tworzenia, doskonalenia i użytkowania MI oraz ich przestrzeganie; II – etyczność relacji MI – człowiek (*Homo sapiens*), zwany przez transhumanistów naturalnym⁴; III – wykorzystanie MI do korygowania lub kształtowania moralności oczekiwanej w danej społeczności.

¹ SI może najogólniej zostać określona jak symulowanie komputerowe ludzkiej inteligencji w oglądzie świata, rozpoznaniu mowy, posługiwaniu się językiem, wytwarzaniu przedmiotów materialnych i niematerialnych, analizie danych, formułowaniu informacji i posługiwaniu się nią, uczeniu się, budowaniu narracji itd.; zob. G. Hołub, *Artificial Intelligence: asking about its ontological status*, „Logos i Ethos” 2024, 30(2), s. 97-114.

² The New English Penguin Dictionary, red. R. Allen, London 2001. Termin maszyna w rozszerzonym znaczeniu angielskim został użyty np. w nazwie Turing machine (maszyna Turinga), czyli abstrakcyjne (modelowe) urządzenie, mające na celu symulowanie działania w wyniku wykonywania algorytmu.

³ MI współcześnie jest tworem nie tylko technologii informacyjnych lecz zintegrowanych z nimi technologii genetyki, nanofizyki i robotyki (GNR albo GRIN); odpowiednio zob. R. Kurzweil, *Progress and Relinquishment* [w:] M. More, N. Vita-More (red.), *The Transhumanist Reader*, Malden 2013, s. 451-453; T. Grabińska, *Człowiek i wartości w transhumanizmie*, «scriptum», Kraków 2024, s. 23.

⁴ Człowiek naturalny (natural human) jest terminem używanym przez transhumanistów na określenie człowieka biologicznie „czystego”, tzn. jeśli poddanego technicyzacji (umiarkowanej bionizacji), to w granicy gatunkowej normy fizjologicznej *Homo sapiens* i niescyborgizowanego, tj. niepoddanego technologicznemu wzmocnieniu (enhancement) zwłaszcza dyspozycji mentalnych, przewyższających normę biologiczną; np. zob. N. Bostrom, *Why I Want to be a Posthuman When I Grow Up* [w:] M. More, N. Vita-More (red.), *The Transhumanist Reader*, Malden 2013, s. 28-53, tu: 32; G. Hołub, *Stać się więcej niż człowiekiem?*, Lublin 2024. Transhumanizm zaś to współczesna ideologia (rozwijająca się od końca XX wieku) propagująca wizję technologicznego doskonalenia człowieka i świata, przekraczania biologicznych ograniczeń

Przedmiot niniejszych rozważań mieści się przede wszystkim w nurcie II.

Brian P. Green przedstawia⁵ przegląd 12 problemów etycznych, które powstają w realizacji celów (dóbr, wartości) człowieka naturalnego w relacji z MI. Są to: 1) zasady i bezpieczeństwo działania MI, 2) przejrzystość działania MI a jej specyficzna prywatność, 3) złe użycia MI, 4) dobre użycia MI, 5) stronniczość rozstrzygnięć MI, 6) bezrobocie wywołane stosowaniem MI, 7) wzrost nierówności społeczno-ekonomicznych w wyniku upowszechnienia MI, 8) automatyzacja etyki i pozbawienie człowieka kompetencji moralnych w relacjach z MI, 9) świadomość i prawa MI (robotów), 10) wzrost zależności od technologii ulepszających MI, 11) skutki społeczno-psychologiczne upowszechnienia MI, 12) skutki duchowe upowszechniania MI.

12 zagadnień etycznych wyróżnionych przez Briana P. Greena pozwala uszczegółowić tematykę nurtów I-III. Oto wstępne przyporządkowanie nurtom zagadnień. Jest ono symboliczne, gdyż przeprowadzona poniżej analiza pokaże, że wyróżnione zagadnienia należą jednocześnie do kilku nurtów:

I: (3), (4), (6), (7), (11), (12);

II: (1), (2), (5), (9);

III: (8), (10).

Wraz z dynamicznym rozwojem technologii GRIN zmiany środowiska życia i działania człowieka następują bardzo szybko, stąd reakcja na ich skutki może być wkrótce niewystarczająca. Tak ważne zatem jest podjęcie dyskusji w trybie odpowiedzialności za projektowanie nowych MI i stosowanie ich, rozwijającym np. etykę odpowiedzialności Hansa Jonasa⁶.

Trwa dyskusja o możliwości wytworzenia sztucznej świadomości i wzbogacenia nią MI⁷. Jednym ze stanowisk jest to, że sztuczna świadomość będzie konstruktem symulującym ludzką świadomość podobnie jak sztuczne wytwory materialne symulują przedmioty naturalne. Te wszak wzmacniają albo osłabiają ludzką aktywność i twórczość. Natomiast sztuczny wytwór obdarzony świadomością może ponadto generować zarówno rozpoznanie i decydowanie oczymś w inny sposób niż się to dzieje w wyborze celu i metod przez człowieka naturalnego. Może też wzmacniać lub zastępować ludzkie procesy świadomościowe. Zasadne jest zatem pytanie o to, z jakim skutkiem i kto (co) będzie za te transformacje odpowiedzialny (odpowiedzialne).

Zarysowany problem odpowiedzialności jest na tyle poważny, że wymaga oddzielnego opracowania. Tu problemy etyczności będą rozważane w sposób autorski lecz w ich porządku wyznaczonym przez B.P. Greena.

(1) Działanie i bezpieczeństwo działania MI: Czy to (tj. MI) działa i czy w taki sposób jak przewidywałoby się?

za pomocą technobiologii, produkowania tworów antropoidalnych i sprzężenia z otoczeniem ich egzystencji. Mają temu służyć zintegrowane technologie GNR (GRIN), zaangażowane w toczącą się już rewolucję przemysłową (Industry 4.0) i zapowiadaną Industry 5.0; K. Schwab, *The Fourth Industrial Revolution*, London 2017; T. Grabińska, op.cit., s. 23.

⁵ B.P. Green, *Ethical Reflections on Artificial Intelligence*, "Scientia et Fides" 2018, 6(2), s. 9-31.

⁶ H. Jonas, *Zasada odpowiedzialności. Etyka dla cywilizacji technologicznej*, tłum. M. Klimowicz, Platan, Kraków 1996.

⁷ O świadomości SI (artificial consciousness – AC, machine consciousness) zob. np. M. Farisco, K. Evers, J.P. Changeux, *Is artificial consciousness achievable? Lessons from the human brain*, "Neural Networks" 2024, no. 180, s. 106714.

Przedmiotem rozważań w punkcie (1) są efekty implementacji technologii SI w MI i ewentualne zagrożenia, jakie ze sobą niosą. W języku etyki chodzi o problem sprawczości MI. Działanie MI mogłoby być wtedy bezpieczne, gdy człowiek-użytkownik miał nad nią pełną kontrolę lub ona sama kontrolowałaby i korygowałaby skutki własnego działania. To drugie wymaga nie tylko elementu świadomościowego MI (rozpoznania otoczenia i samorozpoznania) lecz także woli. Człowiek, mimo swoich wielu barier biologicznych i poznawczych, posługuje się tzw. inteligencją ogólną, podczas gdy MI jest wyspecjalizowana w określonych czynnościach, w ramach tzw. rozpoznawania ograniczonych danych i symptomów oraz ograniczonej reaktywności na nie: cechują ją wąska sztuczna inteligencja (*artificial narrow intelligence* – ANI). W związku z tym jej działanie w okolicznościach wykraczających poza te przez nią rozpoznawane może ją unieczynnić, jak i wywołać nieprzewidywalne skutki. Dlatego współcześnie rozwijana ogólna sztuczna inteligencja – OSI (*artificial general intelligence* – AGI), która ma dorównać poziomowi ogólności ludzkiej inteligencji⁸.

Już obecnie kontrola człowieka po prostu stosującego MI (zewnętrzna audytowalność) staje się niewystarczająca. Coraz częściej niezbędne staje się komunikowanie człowieka z MI (w trybie interfejs). W konsekwencji, ze względu na technologiczne coraz większe zaawansowanie MI, aby relacja z maszyną miała charakter partnerski, człowiek będzie poddawany technicyzacji w postaci np. implantów wzmacniających funkcje poznawcze i komunikacyjne, a także reaktywność psychofizyczną. W ten sposób etyczność relacji dwóch podmiotów, tradycyjnie rozumiana jako przestrzeganie przez nich relacji norm postępowania w celu realizacji przyjętych wartości przesuwana się kierunku technicznej optymalizacji oddziaływania człowiek – MI.

(2) Przejrzystość działania MI a jej prywatność: Czy można to (działanie MI) zrozumieć w sensie oceny moralnej?

Sprawcą, jakim jest MI, trudno oceniać względem kryterium zasad moralności. Jej sprawczość jest bowiem nieprzejrzysta z brak dokładnego rozpoznania stanu jej dyspozycji do działania, utworzonych w procesie głębokiego uczenia się. Człowiek naturalny nabywa tych dyspozycji w wychowaniu, standardowym nauczaniu i trenowaniu, a w swoim działaniu ujawnia ich obecność lub ich brak. Więc jego czyn może być kwalifikowany moralnie, a owa kwalifikacja zwiększa lub zmniejsza zaufanie do niego. Natomiast margines zaufania do predykcji i przewidywań MI jest znacznie mniejszy, co wyzwalając powinno zwiększoną ostrożność w przyjmowaniu rozwiązań dyktowanych przez SI. Trudność jednak polega na znalezieniu odpowiedzialnego weryfikatora tych rozwiązań. Skoro jest niedostępny człowiekowi naturalnemu, to działanie MI nie jest przejrzyste (*transparent to inspection*). Można by je porównywać z jakąś prywatnością procesu *quasi-myślowego* MI⁹, faworyzującą lub dyskryminującą jakieś wartości lub grupy społeczne¹⁰.

⁸ Implementacja AGI wymaga algorytmów uwzględniających konteksty wyrażen językowych, uczenie się ze zrozumieniem, intuicyjną predykcję, samodzielne stawianie i rozwiązywanie problemów, elastyczność w dostosowaniu się do zmieniających się warunków działania itp.

⁹ B.P. Green, op.cit., s. 30.

¹⁰ N. Bostrom, E. Yudkowsky pokazują przykład, w którym mimo to że algorytm nie pracuje na danych o rasie, wyniki wyboru kredytobiorców przez MI wskazują na dyskryminację

Aby zwiększyć poziom przejrzystości swego działania MI powinna komunikować człowiekowi cele jej działania i sposoby ich realizacji. Wtedy jednak znów zjawia się potrzeba zaopatrzenia MI w jakiś rodzaj samoświadomości, jak i świadomości potrzeb człowieka-użytkownika. Ten proces komunikacji wymaga odpowiedniego wyposażenia użytkownika w dodatkowe technologiczne receptory (wprowadzane za pomocą częściowej cyborgizacji). Z tym, że realizacja przez MI potrzeb człowieka użytkownika, może być zanedo spersonalizowana zamiast spełniać kryterium obiektywności i prawdziwości przekazu informacji czy sugerowania właściwego rozstrzygnięcia i podejmowania decyzji. Ponownie więc trzeba zadać pytanie o weryfikatora treści komunikowanych przez MI i o przejrzystość przetwarzania przez nią danych.

Powstaje także problem ochrony świadomościowej struktury MI. Gdyby sprowadzić ją tylko do czegoś w rodzaju prawa patentowego (twórcy obiektu materialnego) czy autorskiego (twórcy obiektu niematerialnego), to zjawia się pytanie o podmiot odpowiedzialności moralnej (i prawnej) za skutki działania MI.

(3) Potencjalny ogrom zła wyzwalanego za pomocą MI: Jak przeciwdziałać mu, tj. złu wywoływanemu posługiwaniem się MI?

Człowiek swojej inteligencji używa także do działań destrukcyjnych, skutkujących złem moralnym. Umożliwienie mu w wyniku stosowania MI coraz większej i szybszej efektywności oraz precyzji wprowadzania w czyn złych zamiarów zagraża zagładą przynajmniej części ludzkości, o czym ostrzegał m.in. H. Jonas. Zaproponował zatem w swojej etyce odpowiedzialności imperatyw stwierdzający, że „wolno nam ryzykować własne życie, ale nie życie ludzkości”¹¹. Imperatyw ten skierowany jest zwłaszcza do tych, którzy eksperymentują z nowymi wytworami nauki i techniki oraz projektują ich wdrożenia, a w formie nakazu brzmi on następująco: „Postępuj tak, aby skutki twojego działania dały się pogodzić z ciągłością trwania autentycznego życia ludzkiego”, w wymiarze pokoleniowym. Owa autentyczność znaczy tu „integralność człowieka”¹². Zapewnia ona istnienie etyki.

W warunkach poważnego zagrożenia ludzkiego zdrowia, życia i mienia jak na wojnie, gdy w interakcjach stron trudno jest egzekwować przyjęte normy moralne, tworzy się normy prawne, które określają dopuszczalne metody i skutki owych interakcji, jak również egzekwowanie odpowiedzialności stron za przekroczenie kodyfikowanych norm. Jak to jest trudne w praktyce świadczy mozolne formułowanie i problemy z egzekucją zapisów Międzynarodowego Prawa Humanitarnego Konfliktów Zbrojnych¹³. Czy w odniesieniu do interakcji człowiek – MI potrzebne będzie rozszerzenie pojęcia humanitarność ze względu na technologiczny charakter tych oddziaływań i – z drugiej strony – objęcie wytworów technologii obowiązkiem jej przestrzegania?

czarnoskórych kandydatów; N. Bostrom, E. Yudkowsky, *The Ethics of Artificial Intelligence* [w:] K. Frankish, W.M. Ramsey (red.), *Artificial Intelligence*, Cambridge 2024, s. 316-334.

¹¹ H. Jonas, op.cit., s. 39.

¹² Ibidem, s. 38.

¹³ Z. Falkowski, *Podstawowe zasady międzynarodowego prawa humanitarnego*, „Kwartalnik Policyjny” 2022, nr 4, s. 21-29.

(4) Potencjalny ogrom dobra wyzwalanego za pomocą MI: Jakie (dobro) zrealizować, wspomagając się MI?

Stosowanie MI może owocować dobrem, o jakim człowiek od dawna marzył dla siebie i innych. Oto wciąż poszerza się horyzont autentycznych ulepszeń we wszystkich sferach życia, jak m.in. w wyżywieniu (planowanie upraw w zakresie jakości i ilości oraz wyceny produktów, personalizowana żywność); relacjach interpersonalnych (media społecznościowe); obrotach finansowych (narzędzia płatnicze, instrumenty FinTech); działaniach wojennych – militarnych i niemilitarnych (automatyzacja pola walki); wzbogaceniu źródeł energii; opiece medycznej (diagnozowanie przyszłego stanu zdrowia, medycyna prewencyjna, inteligentna diagnostyka); wyręczeniu człowieka w ciężkiej niebezpiecznej pracy fizycznej; poszerzeniu możliwości samorealizacji w twórczości każdego rodzaju; edukacji (spersonalizowane programy nauczania).

Jednak już tylko wymienione dobrodziejstwa mają swoją negatywną stronę, o ile nie będzie się korzystać z udogodnień oferowanych przez MI w sposób ostrożny i odpowiedzialny, zgodnie z zaleceniem w punkcie (3).

(5) Dziedziczenie i zwielokrotnienie przez MI ludzkiej słabości w wprowadzaniu błędnych danych, jak i intencjonalnego konstruowania algorytmów: Czy to (takie wadliwe działanie MI) będzie sprawiedliwe?

Jeśli MI będzie pracować na błędnych lub fałszywych danych albo opierać się na algorytmach z wbudowaną intencją oceny albo skutku decyzyjnego lub praktycznego, to rozstrzygnięcie MI będzie skażone brakiem obiektywności, a więc pewną stronniczością. Szczególnie ważne z punktu widzenia zachowania życia ludzkiego, właściwego stanu środowiska naturalnego, skutków aktywności politycznej i społecznej, uniknięcia takiej lub innej katastrofy działanie MI musi być jak najmniej stronnicze, w tym sensie – sprawiedliwe, bo wtedy uwzględni cele optymalnie pozytywne dla całej populacji nimi objętej. A zatem podobnie jak w zaleceniu w punktach (1) i (2) pojawia się konieczność technicznej optymalizacji komunikowania się człowieka z MI.

W przykładzie przytoczonym w przypisie 10 MI może przypadkowo uwzględnić np. adresu zamieszkania wskazującego na dzielnicę zamieszkałą głównie przez czarnoskórych z rasą kandydata na kredytobiorcę. Wtedy jest pozornie stronnicza, lecz może być kierowana intencją nadaną z zewnątrz w wyniku manipulacji algorytmem np. celowego zawyżania wagi tej korelacji. To z kolei nie musi wskazywać na jej rasistowską intencję lecz na podwyższoną ochronę kredytodawcy, uwzględnioną w uczeniu maszynowym.

Stronniczość, powodująca niesprawiedliwe rozstrzygnięcia, może być nieintencjonalna, intencjonalna, ale może być wywołana swoistym „przekupstwem” MI przez użytkownika, który zaprogramuje MI na wyniki, w jakiś sposób jej nagradzane.

(6) Bezrobocie wywołane zastosowaniem MI: Czym wszyscy (ludzie) będą się trudnić?

MI w postaci coraz to bardziej wyspecjalizowanych i efektywniejszych robotów lub algorytmów zastępuje pracę ludzką zarówno z powodu coraz to większej efektywności MI, jak i minimalizacji kosztów zatrudnienia. Człowiek naturalny staje się bezrobotny, w związku z czym traci tradycyjny sens własnej egzystencji i misji, jaką miałby do spełnienia w życiu. Staje się niepotrzebny. Taki zaś stan usposabia do depresji albo agresji w stosunku do otoczenia (ludzkiego, przyrodniczego, kulturowego), które przestaje mieć jakąkolwiek wartość. Innym wyjściem jest popadnięcie w nałogi, co już teraz jest widoczne w postaci siecioholizmu, która uzależnia egzystencję w sferze mentalno-psychicznej od stałej obecności w przestrzeni wirtualnej. Docelowo transhumaniści¹⁴ wieszczą w ogóle wirtualizację człowieka.

Człowiek pozbawiony pracy może odnajdować sens życia w interakcjach rodzinnych i przyjacielskich, oddawać się różnym hobby, uprawiać sport, egzystować w postaci *homo ludens*, o ile będzie miał zabezpieczone podstawowe warunki życia. Jest to o tyle niejasne, że transhumaniści przewidują zhierarchizowanie społeczne wśród populacji w różnym stopniu zbionizowanej i scyborgizowanej, a więc złożonej z osobników w różnym stopniu współdziałających z technologizowanym otoczeniem, a stąd w nierówny sposób korzystający z dobrodziejstw w nim oferowanych.

Człowiek naturalny pozbawiony technologicznych ulepszeń jest ustawiony w tej hierarchii nisko, tuż za zwierzętami i w sposób traktowania ich przez człowieka będzie stosowany wobec niego przez bardziej technologicznie wzmocnionych (*enhanced*). Dlatego postępujące technologizowanie środowiska ludzkiego życia wymusi poddawać się owemu wzmocnieniu. Problem jednak w tym, że nowa hierarchizacja społeczna będzie prawdopodobnie miała odpowiednik w zróżnicowaniu stanu posiadania. A zatem naturalnych ludzi nie będzie stać na owo ulepszenie; powstanie też gradacja ludzi ulepszonych.

(7) Wzrost nierówności społeczno-ekonomicznych w wyniku upowszechnienia MI: Co komu będzie dane?

W wyniku wprowadzania technologizacji maleją koszty produkcji, co podnosi zyski właścicieli środków produkcji i pozwala im je przeznaczać na rozwój nowych zintegrowanych technologii, inwestować w nowe generacje MI i upowszechnianie ich zastosowania. Poziom dywersyfikacji bogactwa będzie zapewne szybko rosnąć, przy czym nadzieja na jego redystrybucję jest płonna. Pogłębiać się będą nierówności społeczno-ekonomiczne i to w skali globalnej, włącznie z upadłością biedniejszych krajów, których nie stać będzie na wprowadzenie ujemnego podatku dochodowego (*negative income tax*)¹⁵.

Pewną propozycją jest zaangażowanie ludzi pozostających bez zatrudnienia w usługach, które miałyby na celu podnoszenie estetyczności otoczenia, opieki nad potrzebującymi. MI, jak na razie, nie dominuje w tym zakresie nad człowiekiem naturalnym, ale może się to wkrótce zmienić. Ludzka

¹⁴ Zob. np. D. Cecchetto, *Humanesis. Sound and Technological Posthumanism*, Minneapolis – London 2013.

¹⁵ Negative income tax (NIT) to koncepcja dofinansowania przez państwo biedniejszych obywateli do progu podatku dochodowego.

kreatywność, wrażliwość emocjonalność także może zostanie symulowana działaniem MI, która stanie się tworem *sentience*.

(8) Automatyzacja etyki i de-skilling moralny w relacjach z MI: Brak praktyki czyni niedoskonałym.

Tradycyjnie doskonalenie moralne odbywa się wraz z doskonaleniem umiejętności wykonywania zadań zarówno w określaniu użytecznych i moralnie pozytywnych celów, jak i perfekcyjności ich osiągania w zgodzie z systemem wartości i zasadami moralnymi. Rozwój MI powoduje brak zapotrzebowania na ludzkie kompetencje zarządcze i wykonawcze. Człowiekowi naturalnemu, pozbawionemu umiejętności, jak i temu technologicznie wzbogaconemu i w jakimś kierunku wyspecjalizowanemu, przyjdzie, być może, jedynie oswajać otoczenie w wyniku wykonywania funkcji poznawczych i reaktywnych na zasadzie *quasi*-instynktownej, przy czym proces ten będzie się odbywał w ramach bardziej lub mniej zaawansowanej optymalizacji technologicznej tej symbiozy.

Można się spodziewać powstania dyscypliny nazywanej etyką maszyn (*machine ethics, machine morality*)¹⁶, której celem jest wyposażenie MI w zdolność podejmowania decyzji moralnych. Tak zbudowana MI byłaby sztucznym sprawcą czynów moralnych (*artificial moral agent*)¹⁷, co dotychczas było przedmiotem narracji *science fiction*. Jednak pojawiające się już i przewidywane sytuacje rozbieżności decyzji podejmowanych przez MI z ludzkimi intencjami warunkowanymi moralnie mogą doprowadzić do konfliktów o katastrofalnych skutkach. Stąd wysiłki tworzenia maszynowej etyki są coraz bardziej intensywne.

Podobnie jak przejmowanie umiejętności praktycznych przez MI zubaża wykonawczo człowieka naturalnego, tak i – jak można przypuszczać – automatyzacja etyki zuboży go moralnie. Wystąpienie kolapsu w kierunku sugerowanego już zachowania *quasi*-instynktownego i dostosowawczego jest realne, a także w kierunku infantylizacji¹⁸, której może od człowieka wymagać MI z powodu wcześniej wymuszonego na nim braku doświadczenia w podejmowaniu decyzji moralnych.

Pewnym optymizmem mają napawać projekty edukacji moralnej ludzi, prowadzone przez MI wyposażonej w odpowiednie algorytmy maszynowej etyki. Można jednak wskazać na trudności z tym związane nie tylko z powodu różnic kulturowych mających wpływ na zasady moralne, lecz także w związku ze znaczeniem wolnej woli w decyzjach moralnych, a w związku z tym odpowiedzialności za nie.

(9) Świadomość robotów i prawa robotów: Czy bez sprawiedliwości dla robotów nie zapanuje pokój?

W związku z projektem wyposażenia MI w świadomość, wrażliwość i moralność rozwija się dyskusja na temat praw robotów (zaawansowanych MI), podobnych do praw człowieka (naturalnego). Prawom jednak towarzyszą

¹⁶ L.M. Pereira, A.B. Lopes, *Machine Ethics. From the Machine Morals to the Machinery of Morality*, Cham 2020.

¹⁷ B.P. Green, *op.cit.*, s. 9-31.

¹⁸ *Ibidem*, s. 9-31.

obowiązki, a egzekwowanie praw i obowiązków określa stan sprawiedliwości społecznej. O ile rozpisanie listy praw robotów jest do pomyślenia, to powstają poważne problemy z określeniem ich powinności i wdrożenia zasad sprawiedliwości. Określenie praw i obowiązków zawsze występuje w relacji jednego podmiotu do drugiego, a te się komplikują. Tym razem to nie tylko relacja człowiek (naturalny) – człowiek (naturalny), ale człowiek (naturalny) – transczłowiek (o różnym stopniu wzmocnienia technologicznego), robot – człowiek taki lub inny, robot – hybryda taka lub inna, człowiek taki lub inny – hybryda taka lub inna itd., nie rozpatrując już stosunku do tak lub inaczej tworzonych instytucji. Podobnie pojawiają się wielorakie relacje między w różnym stopniu zaawansowanymi MI.

Urzeczywistnienie transhumanistycznej różnorodności biotechnologicznej¹⁹ prowadzi do powstania różnych hybryd ludzko-technicznych, zwierzęco-ludzkich, zwierzęco-technicznych. Pojawi się więc problem tzw. osobowości prawnej członów tych wielorakich relacji. Stworzenia sprawiedliwego prawa w tak złożonej sytuacji wydaje się być utopią. Realnym acz niebezpiecznym rozwiązaniem jest totalne ujenolicenie praw i obowiązków, a zatem wykluczenie sprawiedliwości, przynajmniej rozumianej po arystotelesowsku. Stworzenie idealnego sędziego w postaci wyspecjalizowanej MI, mogącej te wszystkie skomplikowane relacje analizować, klasyfikować i nadawać im status prawny wcale nie musi zdemokratyzować stosunków nowego społeczeństwa. Poza tym pozostałoby jeszcze do rozwiązania to, jak ma wyglądać egzekucja praw i obowiązków.

W tradycyjnych społecznościach rządzonych niesprawiedliwie narasta zwykle niezadowolenie całego społeczeństwa lub poszczególnych grup, które często przeradza się w bunt²⁰. Możliwy jest zatem bunt grup nie do końca wzbogaconych technologicznie czy w różnym stopniu przekształconych technicznie. Bunt, jeśli nie jest totalny, to rozkłada się na różnie zachodzące na siebie konflikty między grupami. Prawa nadane robotom niczego wobec tego nie rozwiązują. Zamiast bowiem pokojowego istnienia nowego społeczeństwa²¹ raczej będzie ono się charakteryzować ciągłym wewnętrznym konfliktem, o ile nie zapanuje władza totalitarna.

(10) Zależność od technologii ulepszających MI: *Nie ma odwrotu od ściślejszego sprzężenia transosobników z technologią.*

Technicyzacja człowieka i technologizacja środowiska jego bytowania coraz bardziej uzależnia nie tylko ludzi naturalnych lecz również wszelkich transosobników²² od technologii. Umożliwia ona, po pierwsze, ich działanie, ustalanie relacji między nimi oraz między obiektami MI w różnych kombinacjach członów relacji, wymienionych w punkcie (9). Można wprost wskazać na symptomy tworzenia się technokultury, jak i technocentrycznego humanizmu.

Nazwa technocentryczny humanizm²³ może ponadto stać się pusta ze względu na unieważnienie czynnika ludzkiego. Aby do tego nie doszło należałoby

¹⁹ N. Bostrom, Transhumanist Values, "Journal of Philosophical Research" 2005, no. 30, s. 3-14.

²⁰ B.P. Green, op.cit., s. s. 9-31.

²¹ Pokojowe współistnienie technicznie przekształconej społeczności jest jedną z wartości transhumanistycznej Bostroma; N. Bostrom, op.cit., s. 3-14.

²² Transosobnik jest nazwą ogólną przekształconego technologicznie człowieka (czyli transczłowieka), zwierzęcia, jak i hybrydy biologiczno-technologicznej.

²³ T. Grabińska, op.cit., s. 23.

zadbać o podtrzymanie ludzkiej aktywności w świecie uzależniającym wszelkie działanie od MI. Ważnym bowiem problemem może się stać poszukiwanie równowagi między wydajnością zautomatyzowanych i autonomicznych systemów współdziałających ze sobą, a więc wykazujących pewne scentralizowanie decyzji i sposobów wykonania, a jakością i sprawnością wykonania zadań zarządczych i produkcyjnych²⁴. Prostota i umiar cechujące człowieka naturalnego może istotnie przeciwdziałać zacieraniu się działania takich systemów. Musi być jednak spełniony warunek włączenia go do swoistej zewnętrznej kontroli MI, co prawdopodobnie wymagać będzie jednak wzbogacenia technologicznego jego funkcji poznawczych i decyzyjnych.

(11) Skutki społeczno-psychologiczne upowszechnienia MI: Za mało i za dużo.

Część negatywnych skutków społeczno-psychologicznych, powstałych w wyniku upowszechniania MI została ukazana w punktach (6) i (7). Problem alienacji i wykluczenia społecznego będzie narastał wraz postępowaniem technologizacji środowiska życia i działania człowieka naturalnego, ale i innych projektowanych transosobników wrażliwych (*sentience*). Przewiduje się wprost kryzys samotności, niezależnej od wieku. Już teraz w Wielkiej Brytanii powołano ministra ds. samotności, gdyż okazuje się, że już kilka lat temu 200 tys. seniorów w tym kraju nie rozmawiało z kimś bliskim od ponad miesiąca²⁵. I znów panaceum wydaje się tworzenie towarzystwa w postaci MI, wyspecjalizowanych w konwersacji i empatii. Jednak inne trudności upowszechniania MI ukazane w punktach (7), (9), (10) mogą przeszkodzić w realizacji tego zamierzenia. Optymiści są jednak zdania, że MI może pomóc w pielęgnowaniu i rozwijaniu relacji międzyludzkich, ale wtedy trzeba by się zdać na automatyzację etyki i nadanie uprawnień robotom, a to, jak zostało zaznaczone w punkcie (8) i (9), rodzi nowe problemy.

(12) Wpływ MI na ludzką duchowość: Co będzie znaczyć człowieczeństwo?

To, co wyróżnia człowieka (*Homo sapiens*) wśród wszystkich istniejących na Ziemi organizmów to duchowość rozumiana jak szczególny spłot świadomości, inteligencji i rozumności oraz wolitywności. Pytanie o to, czy wzmacnianie technologiczne tych wyznaczników człowieczeństwa zwiększy jego ludzki potencjał, jest ważne. Odpowiedź na nie nie jest jednoznaczna. Brzmi „tak” – gdy zwiększy jego sprawność, ale w ścisłej zależności od technologii (a więc z umniejszeniem np. wolitywności i znaczenia samoświadomości). Odpowiedź brzmi „nie” – jeśli stanie się on trybikiem w systemie złożonym z połączonych przedmiotów MI – zarządczych lub wytwórczych. Techniczne instrumentarium, zainicjowane przez człowieka może zatem przejąć owe wyjątkowe ludzkie dyspozycje, jak to częściowo zostało ukazane w punkcie (6).

W rozważaniach przeprowadzonych gdzie indziej został sformułowany postulat, że zachowanie osoby ludzkiej we wzbogacanym technicznie człowieku – tak, jak jest ona rozumiana w personalizmie, powinno być warunkiem *sine qua*

²⁴ B.P. Green, op.cit., s. 9-31.

²⁵ Ibidem, s. 9.

non, ograniczającym technicyzację człowieka i technologizację jego otoczenia²⁶. Osoba ludzka jest bytem psychofizycznym, obdarzonym rozumem i wolą, dynamicznym, kreatywnym i odpowiedzialnym za podejmowane i wykonane czyny. Rozważania zatem zwłaszcza o odpowiedzialności powinny towarzyszyć tworzeniu nowego technologiczowanego społeczeństwa i technokultury.

Podsumowanie

Przeprowadzone rozważania prowadzą do pytania o status moralny MI. Jeśli komunikacja i współdziałanie z MI stanie się powszechne, to powstanie problem norm postępowania zaangażowanych w nie stron (częściowo wymienionych w (6)). Choć niektórzy autorzy w uczestniczą w projektowaniu katalogu norm moralnych, to bardziej przyszłościowo realna jest postawa przyjęcia regulacji jednak technologicznej (samooptymalizującej się) niż zasad moralności, podobnych do tradycyjnych, tj. opartych na hierarchii wartości etycznych. Jeśli już jakieś wartości miałyby w tym przypadku znaczenie, to raczej użyteczne. Chodziłoby przy tym o uwzględnienie nie tyle korzyści poszczególnych zaangażowanych osobników humanoidalnych i zhybrydizowanych, ile o utrzymanie homeostazy tak połączonych systemów. Trzy zasady oceny moralnej interakcji wymienione przez N. Bostroma i E. Yudkowsky'ego²⁷ między różnymi jej stronami (jak w 6) są zbyt prymitywne, nawet jeśli uwzględnić wyłącznie użyteczność interakcji.

Zasada „niedyskryminacji podłoża” (*Principle of Substrate Non-Discrimination* – PSND), wymagająca równego statusu moralnego dla obu stron o tej samej funkcjonalności, różniących się jedynie podłożem implementacji SI jest możliwa do spełnienia wyłącznie w relacjach człowiek – MI, jeśli by funkcjonalność człowieka i MI dała się określić jako identyczna. A to przecież jest niemożliwe bez posługiwania się jakimś technicznym analizatorem też o określonej SI, i to raczej w rodzaju OSI. W związku z różnorodnością stron interakcji należałoby przyjąć różne statusy moralne, jeśli by tę zasadę traktować poważnie.

Uzupełniającą zasadą miałyby być zasada „niedyskryminacji ontogenezy” (*Principle of Ontogeny Non-Discrimination* – POND), zachowującą PSND, niezależnie od historii „rodowej” powstania każdej ze stron. Koresponduje ona ze współczesnymi nurtami antyspecyzmu²⁸. Nic nowego zatem nie wnosi i nie rozwiązuje problemu kompatybilności funkcjonalności.

Trzecia zasada „subiektywnego tempa czasu” (*Principle of Subjective Rate of Time* – PSRT) uwzględnia (ale tylko częściowo) podnoszony wyżej problem różnorodności stron interakcji i lokuje go na poziomie egzotycznych umysłów – różniących się od ludzkiego i od siebie rodzajem świadomości i rozumności, czyli tzw. osobowością. Jednym z przejawów egzotyczności umysłów byłoby subiektywne odczuwanie czasu, co miałoby wpływ na tempo oraz efekt porozumiewania się i działania. PSRT wymaga, aby w przypadkach, gdy czas trwania doświadczenia wewnętrznego MI jak np. w procesach *quasi-*

²⁶ T. Grabińska, op.cit., s. 23.

²⁷ N. Bostrom, E. Yudkowsky, op.cit., 316-334.

²⁸ Antyspecyzm (antygatunkowość) to współczesny nurt kulturowy promujący hasła równości gatunkowej organizmów, przeciwstawiający się zwłaszcza dominacji gatunku *Homo sapiens*. Koresponduje z przekazem radykalnych ruchów proekologicznymi (tzw. terroryzmu ekologicznego). Zob. np. T. Morton, *Humankind, Solidarity with Non-human people*, New York 2017.

biologicznych ma podstawowe znaczenie normatywne, tj. wpływa na sposób i czas reakcji, uwzględniać to w ocenie moralnej. Ten wymóg jest kompletnie nie do zrealizowania bez specjalnej technicznej samooptymalizacji relacji między zróżnicowanymi i podlegającymi ciągłym zmianom stronami interakcji, uwzględniającej m.in. tzw. współodczuwanie.

Jeśli status moralny traktować jako przymiot każdej ze stron odznaczających się pewnym poziomem świadomości i wrażliwości, to w ich wzajemnych interakcjach, gdy poziom i jakość rozumności i świadomości wykazują znaczne różnice, to niemożliwe jest wyrównanie statusu moralnego stron, czyli udzielenie im równych praw i obowiązków względem siebie i otoczenia. Tylko optymalizacja technologiczna lub zarządzanie autorytatywne (czynniki decyzyjne) są w stanie zaprowadzić konwencjonalny i wciąż ewoluujący porządek. Jednak wiązanie go z moralnością i sprawiedliwością jest antropocentrycznym mirażem, widocznym już mechanistycznym modelem człowieka i społeczności w siedemnastowiecznej propozycji Thomasa Hobbesa²⁹.

Tylko w relacji człowiek (naturalny) – MI (pod warunkiem zachowania dwóch pierwszych zasad Bostroma-Yudkowsky'ego) możliwe byłoby określenie etyczności tej relacji, a więc statusu moralnego stron interakcji. Takie stanowisko wpisuje się w połączenie antropocentryzmu i technocentryzmu, w którym strona humanoidalna nie zdominowałaby strony pochodzącej od człowieka naturalnego. Te poszukiwania powinny systematycznie wspierać stanowienie prawa w nowym środowisku działania człowieka i tworów humanoidalnych.

Bibliografia

- Allen R. (red.), *The New English Penguin Dictionary*, London 2001.
- Bostrom N., *Transhumanist Values*, "Journal of Philosophical Research" 2005, no. 30.
- Bostrom N., *Why I Want to be a Posthuman When I Grow Up* [w:] More M., Vita-More N. (red.), *The Transhumanist Reader* Malden 2013.
- Bostrom N., Yudkowsky E., *The Ethics of Artificial Intelligence* [w:] K. Frankish, Ramsey W.M. (red.), *Artificial Intelligence*, Cambridge 2024.
- Cecchetto D., *Humanesis. Sound and Technological Posthumanism*, Minneapolis – London 2013.
- Elangovan U., *Industry 5.0. The Future of the Industrial Economy*, CRC Press – Boca Raton – New York 2017.
- Falkowski Z., *Podstawowe zasady międzynarodowego prawa humanitarnego (MPH)*, „Kwartalnik Policyjny” 2022, nr 4.
- Farisco M., Evers K., Changeux J.P., *Is artificial consciousness achievable? Lessons from the human brain*, "Neural Networks" 2024, no. 180.
- Grabińska T., *Człowiek i wartości w transhumanizmie*, «scriptum», Kraków 2024.
- Green B.P., *Ethical Reflections on Artificial Intelligence*, "Scientia et Fides" 2018, no. 6(2).
- Hobbes T., *Lewiatan czyli materia, forma i władza państwa kościelnego i świeckiego*, tłum. C. Znamierowski, Warszawa 2005.
- Hołub G., *Artificial Intelligence: asking about its ontological status*, „Logos i Ethos” 2024, no. 30(2).
- Hołub G., *Stać się więcej niż człowiekiem?*, Lublin 2024.

²⁹ T. Hobbes, *Lewiatan czyli materia, forma i władza państwa kościelnego i świeckiego*, tłum. C. Znamierowski, Warszawa 2005; T. Grabińska, op.cit., s. 23.

Jonas H., Zasada odpowiedzialności. Etyka dla cywilizacji technologicznej, tłum.
M. Klimowicz, Kraków 1996.
Kurzweil R., Progress and Relinquishment [w:] More M., Vita-More N. (red.), The
Transhumanist Reader Malden 2013.
Pereira L.M., Lopes A.B., Machine Ethics. From the Machine Morals to the
Machinery of Morality, Cham 2020.
Morton T., Humankind, Solidarity with Non-human people, New York 2017.
Schwab K., The Fourth Industrial Revolution, London 2017.