W cieniu rozkwitających AI

dr hab. Agnieszka Lekka-Kowalik, prof. KUL
ORCID: 0000-0002-4834-318X
Katolicki Uniwersytet Lubelski Jana Pawła II

1. AI w procesie decydowania

Termin „sztuczna inteligencja" (AI) odnosimy do artefaktów zdolnych do symulowania ludzkiego myślenia i działania[1]. W dalszych analizach pomijam kwestię symulacji: czy ma być ona jedynie wiernym naśladowaniem (chodzi o tzw. zombie AI) czy też naśladowanie wymaga jakiegoś rodzaju podmiotowości. Nie analizuję też kontrowersji wokół pojęcia inteligencji: czy jest ona zdolnością rozwiązywania problemów komputacyjną czy też nie; i czy jest jedna inteligencja czy wiele jej typów. Niezależnie od rozwiązywania wskazanych problemów teoretycznych, nie ulega wątpliwości, że współcześnie AI działają w dużym stopniu autonomicznie i stają się coraz bardziej pełnoprawnymi uczestnikami życia społecznego jako tzw. roboty społeczne. Nietrudno znaleźć przykłady. Humanoidalne roboty zaczynają sprawować opiekę nad osobami starszymi (Andtfolk i in. 2022), a tak zwane roboty

[1] Istnieje cały szereg definicji sztucznej inteligencji i wszystkie one mają charakter definicji regulujących, wypracowanych w konkretnych kontekstach. Np. OECD (Organizacja Współpracy Gospodarczej i Rozwoju) definiuje system AI jako „*a machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments.*"; https://legalinstruments.oecd.org/en/instruments/oecd-legal-0449. Dostęp 5.01.2025. Z kolei w Unii Europejskiej na potrzeby „Rozporządzenia Parlamentu Europejskiego i Rady (UE) 2024//1689 z dn. 13 czerwca 2024" zdefiniowano system AI jako „system maszynowy, który został zaprojektowany do działania z różnym poziomem autonomii po jego wdrożeniu oraz który może wykazywać zdolność adaptacji po jego wdrożeniu, a także który – na potrzeby wyraźnych lub dorozumianych celów – wnioskuje, jak generować na podstawie otrzymanych danych wejściowych wyniki, takie jak predykcje, treści, zalecenia lub decyzje, które mogą wpływać na środowisko fizyczne lub wirtualne", https://eur-lex.europa.eu/legal-content/PL/TXT/? uri=CELEX: 32024R1689 art 3. Dostęp 5.01.2025.

empatyczne stają się towarzyszami dzieci (Leite, Castellano, Pereira 2014). Jednym z najsłynniejszym robotów jest Sophia (zwana fembotem ze względu na podobieństwo do kobiety), która w październiku 2017 r. otrzymała obywatelstwo Arabii Saudyjskiej, występowała na forum Organizacji Narodów Zjednoczonych, a nawet deklarowała, że szuka męża[2]. Stanowi to dla ludzi autentyczne wyzwanie, tak teoretyczne, jak i praktyczne. Rozważane są m.in. metafizyczny status sztucznej inteligencji, możliwość zaistnienia jej świadomości czy odpowiedzialność za podejmowane przez nią czynności. Zafascynowani możliwościami maszyn – choć i nieco nimi przerażeni – chcemy natomiast zagwarantować, że AI zachowuje się moralnie, tj. w każdych okolicznościach „podejmuje" decyzje o działaniu zgodnym z naszymi wartościami. Spodziewamy się wtedy rozlicznych korzyści z funkcjonowania AI w przestrzeni indywidualnej i społecznej, co wprost postulują zasady z Asilomar, sformułowane podczas konferencji w 2017 r. poświęconej rozwojowi sztucznej inteligencji[3]. Zasada (1) głosi, iż rozwój powinien zmierzać do wykreowania nie po prostu sztucznej inteligencji, ale inteligencji dobroczynnej, życzliwej i przyjaznej ludziom (*beneficial*); zasady (10) i (11) postulują kompatybilność działań AI z wartościami ludzkimi, w tym z ideałami godności, wolności i praw człowieka oraz kulturową różnorodnością; a zasada (23) wymaga, by AI rozwijać w służbie dobra wspólnego ludzkości.

Kluczowym momentem dla rozważań o moralności działań AI jest udział sztucznej inteligencji w podejmowaniu decyzji. Wyróżnia się cztery role, które AI może odgrywać w tym procesie: (a) gromadzenie danych, a więc dostarczanie informacji dla podjęcia decyzji przez ludzi; (b) analizowanie danych wraz z predykcjami, a więc tworzenie scenariuszy „jeżeli... to", które dostarczają podmiotowi ludzkiemu przewidywań co do możliwych konsekwencji rozmaitych decyzji (tzw. *assisted AI*); (c) rekomendowanie (na bazie danych i predykcji) pewnej decyzji jako najlepszej w danej sytuacji, przy czym decyzja pozostaje w rękach człowieka; (d) podejmowanie decyzji i wykonanie działania przez autonomiczną AI. To tu właśnie rodzi się problem robotów społecznych, które Kate Darling z Massachusetts Institute of Technology definiuje jako materialne, autonomiczne podmioty działające, komunikujące się i wchodzące w interakcje z człowiekiem na poziomie emocjonalnym (Darling 2016, s. 215). Darling zresztą używa terminu „aktant", by zaznaczyć ich moc działania, a odróżnić od ludzi. Owe aktanty są czymś więcej niż roboty

[2] *Sophia ma obywatelstwo Arabii Saudyjskiej i szuka męża. Jest robotem.* Dobreprogramy, https://www.dobreprogramy.pl/sophia-ma-obywatelstwo-arabii-saudyjskiej-i-szuka-meza-jest-robotem, 6820095894465376a. Dostęp 5.01.2025.

[3] Zob. Asilomar AI Principles, https://futureoflife.org/open-letter/ai-principles/. Dostęp 2.01.2025.

autonomiczne, realizujące zadania niezależnie i w znacznych odległościach od człowieka. To roboty, które mają wchodzić w interakcje i współpracować z człowiekiem jako jego partnerzy, a być może nawet zastępować ludzi jako naturalnych partnerów komunikacji.

Na każdym z poziomów uczestniczenia AI w decydowaniu pojawiają się kwestie etyczne. Najczęściej wymienia się: brak transparentności działania AI, uprzedzenia i stereotypy wcielone w algorytmy, ustalanie odpowiedzialności przed społeczeństwem, zagrożenie dla prywatności, bezstronność i sprawiedliwość w działaniu (Cf. Prem 2023). Istnieje ogromna literatura na temat tych problemów, ale tu interesuje nas następująca kwestia: podejmowanie decyzji jest sprawą bardzo ludzką, a wobec tego, jakie będą konsekwencje oddania decyzji „w ręce" AI, niezależnie od rozwiązania wymienionych wyżej kwestii etycznych? By odpowiedzieć na powyższe pytanie, należy przyjrzeć się dokładniej naturze ludzkiej decyzji.

2. Natura ludzkiej decyzji

Poniższe rozumienie ludzkiej decyzji odwołuje się do rozważań nad decyzją prowadzonych w ramach lubelskiej szkoły filozofii klasycznej. Decyzja widziana jest jako rezultat współpracy rozumu i woli – to sąd autodeterminujący człowieka do działania (Krąpiec 1979; Krąpiec 2001). Decyzja nakierowana ma być na realizację konkretnego dobra i ma być słuszna, tj. respektować godność osoby ludzkiej – i to zarówno człowieka działającego, jak i człowieka, na którego działanie jest skierowane – oraz inne wartości. Jest to zresztą warunek narzucany także na decyzje AI, jak można zauważyć w cytowanych wyżej zasadach z Asilomar. Zasadniczy problem rysuje się jednakże następująco: jakie działanie wykonane przeze mnie tu i teraz faktycznie respektuje owe wartości? Decyzja jest wobec tego podwójnym sądem. Najpierw wydaję sąd o charakterze poznawczym (zwanym teoretyczno-praktycznym): „teraz w tych oto okolicznościach ja *powinnam* zrobić to a to". Sąd ten nie jest rezultatem wnioskowania dedukcyjnego z jakichś ogólnych zasad (choćby tej: czyń dobro, zła unikaj), ale jest twórczy, niejako syntetyzuje wiedzę i doświadczenie, i w tym sensie jest niepowtarzalny. Ostatecznie – w świetle mojego sądu „powinnam" – wydaję sąd (zwany praktyczno-praktycznym): „zrobię to a to". Sąd ten ustanawia mnie źródłem działania. A stając się źródłem działania, rozpoczynam w świecie (a raczej: wszechświecie, bo przecież rzeczywistość jest w skali makroskopowej ciągła) nowy łańcuch przyczynowo-skutkowy. Działanie zaś ma dwojakie konsekwencje. Podjęcie decyzji – nawet jeśli nie zdołam jej wykonać – buduje mnie jako osobę o pewnej osobowości. Gdy podejmuję decyzje o morderstwie, to nawet jeśli ktoś mi przeszkodzi w dokonaniu morderstwa, buduję siebie jako mordercę. Nie musi to być zresztą taki dramatyczny przykład.

Oto decydując się zjeść raczej gruszkę niż jabłko, buduję siebie jako zjadaczkę gruszek. Gdy decyzję zrealizuję, w reszcie świata pojawią się konsekwencje: ktoś zostanie pozbawiony życia, a pewna gruszka zjedzona. Na tym się owe konsekwencje nie kończą, ponieważ rodzą się całe ciągi skutków: ponieważ zamordowano kogoś, odbył się pogrzeb, na którym spotkali się krewni, odnowili kontakty, a ktoś odziedziczył majątek zamordowanego i wobec tego nareszcie mógł założyć firmę, o której marzył, inni krewni obrazili się za pominięcie w testamencie… Można bez trudu wyobrazić sobie następstwa rodzące się w świecie z powodu zjedzenia przez kogoś gruszki. Krótko mówiąc, każda ludzka decyzja ma zawsze konsekwencje dla samego działającego, a gdy wykonana – także dla wszechświata. W tym sensie jestem odpowiedzialna za siebie i za innych (nie tylko ludzi), których moje działanie dotyka. Z racji bycia źródłem działania można człowiekowi przypisać odpowiedzialność, przy czym termin „odpowiedzialny za" ma dwa sensy: jeden odnosi się do odpowiedzialności *ex post* za własne działanie i jego konsekwencje, natomiast drugi do odpowiedzialności *ex ante*, tj. do odpowiedzialności za byty, których wartość uznajemy, a których istnienie i rozwój zależy od działania podmiotu, co rodzi zobowiązanie do działania na rzecz dobra tych bytów-wartości. Podmiot działający jest wobec tego zawsze w normo-twórczej sieci relacji.

Ten spójny obraz aktu decydowania zakłóca ludzka wolność. Oto mogę odrzucić swój własny sąd „powinnam zrobić to a to" i wydać sąd „zrobię coś innego" – i to ten sąd będzie mnie determinował do działania. Wiem, co powinnam zrobić – i robię coś przeciwnego. Jest to doświadczenie znane zapewne każdemu człowiekowi. Mogę za to żałować; a mogę też *post factum* rozpoznać niesłuszność mojej decyzji i przeanalizowawszy ją, posiąść nową wiedzę, a dzięki temu uniknąć błędnych decyzji w przyszłości; w obu przypadkach mogę też mieć wyrzuty sumienia. Oczywiście i przy błędnych decyzjach rodzą się wewnętrzne i zewnętrzne konsekwencje. Sytuacja staje się jeszcze bardziej skomplikowana, gdy uświadomimy sobie istnienie dylematów moralnych. Dylemat moralny to sytuacja konfliktu między dwoma lub więcej wartościami/obowiązkami/normami, które nie dadzą się naraz zrealizować[4]. Znajdujemy się wtedy w sytuacji niepewności, jaka decyzja jest słuszna – a działać musimy, ponieważ „nicnierobienie" jest też decyzją i czasem nawet można przewidzieć, jak się rozwinie bieg wydarzeń bez naszej interwencji. Często też musimy komuś – zwłaszcza temu, kogo nasze działanie bezpośrednio dotknęło – przedstawić racje czy też uzasadnienie naszej decyzji i oczywiście – jak w przypadku każdej decyzji – podjąć odpowiedzialność za konsekwencje. Zaapli-

[4] Ważne rozważania nad naturą dylematów i konfliktów moralnych zawiera monografia pod redakcją Piotra Duchlińskiego (Duchliński 2023).

kujmy teraz to rozumienie ludzkiej decyzji do sztucznej inteligencji wyposażonej w zdolność determinowania się do działania i wykonania ustalonego działania.

3. Konsekwencje oddania decyzji „w ręce" AI

Przeprowadźmy pewien eksperyment myślowy. Oto AI pracuje jako dystrybutor respiratorów. I nagle rodzi się sytuacja dylematogenna: jest dwóch pacjentów i jeden respirator. Pojawia się więc swoisty „konflikt instrukcji": „przydziel pacjentowi X, bo potrzebuje" i „przydziel pacjentowi Y, bo potrzebuje". Komu przydzielić? Decyzja musi być podjęta, i to podjęta szybko, bo inaczej pacjenci umrą. AI musi wobec tego wydać sąd „powinno się przydzielić temu a temu", a następnie sąd „przydzielę temu właśnie" – i tę decyzję wykonać, podłączając respirator wybranej osobie. Ten sąd musi uwzględniać fakty (dane) i wartości, bo narzucamy na AI wymóg zgodności jej „decyzji" z naszymi ideałami – jak wyżej cytowałam – godności, wolności i praw osoby oraz kulturowej różnorodności. Co więcej, AI musi być zdolna do przedstawienia uzasadnienia swojej decyzji, choćby dlatego, że rodzina pacjenta, który umarł z braku respiratora, ma prawo zapytać, dlaczego akurat on sprzętu nie dostał. Pomińmy problem, jakie uzasadnienie byłoby w takiej sytuacji adekwatne. Zauważmy natomiast, że AI musi być wyposażona nie tylko w umiejętność rejestracji faktów, znajdowania korelacji i odróżniania ich od związków przyczynowo-skutkowych, ale też w umiejętność wydawania sądów wartościujących o słuszności i niesłuszności danego działania, uzasadniania tych sądów, a więc obrony słuszności swej decyzji, co wymaga oceny decyzji pod względem moralnym. U człowieka tego typu zdolności zwykło się nazywać sumieniem. Wydaje się więc, że jeśli AI ma symulować ludzkie myślenie i zachowanie, to należy ją wyposażyć w sztuczne sumienie – zdolność sądzenia o własnych (i cudzych) czynach w aspekcie dobra i zła moralnego.

W takim kierunku idą badania w ramach transdyscyplinarnego obszaru tzw. etyki maszyn (*machine ethics)*. Za Jamesem Moorem (2006) odróżnia się domniemanego agenta moralnego (*implicit moral agent*) i jawnie moralnego agenta (*explicit moral agent*). Termin „agent" jest tu użyty celowo, by możliwy tu do wykorzystania termin „podmiot" czy „podmiot działający" nie sugerował, iż etyczne maszyny mają jakiś rodzaj podmiotowości. AI jest domniemanym agentem etycznym, gdy w algorytm są wprowadzone ograniczenia nakazujące etyczne zachowania lub zakazujące zachowań nieetycznych. Przykładem jest zakaz używania pewnych wyrażeń wprowadzony w ChatGPT, ale też „uczciwe" obliczanie raty kredytu. AI jako jawnie moralny agent miałby w swym algorytmie reprezentację etyki i działałby na podstawie tej wiedzy, rozwiązując m.in. dylematy moralne. Jawnie moralny agent operowałby więc zarazem na przypadkach i na zasadach,

które są niezbędne, by przewidzieć, jak zachowa się agent w nowych sytuacjach, by możliwe było wydobycie moralnie istotnych różnic między przypadkami, by dało się uzasadniać decyzje oraz by następowały modyfikacje instrukcji pod wpływem „doświadczenia" (Anderson i Anderson 2007). Próby zbudowania jawnie moralnego agenta były i są podejmowane, a potrzeba stworzenia takiego agenta wielorako uzasadniana (Anderson, Anderson i Armen 2004; Allen, Smit i Wallach 2005;Wallach i Allen 2009; Anderson i Anderson 2011; Misselhorn 2018).

Jednakże nawet pobieżne przyjrzenie się tym próbom sygnalizuje poważny problem praktyczny. Naczelne zasady moralne (i wszelkie ich pochodne) są różne w różnych teoriach etycznych i wobec tego może zaistnieć różnica decyzji, co zrobić w konkretnej sytuacji (i jak rozwiązać dany dylemat moralny) w zależności od tego, kto zaprogramował konkretną AI, na bazie jakich doświadczeń uczyła się ona podejmować decyzje i jaka jest naczelna norma włączona w algorytm. Czy wobec tego przed zakupem „pielęgniarki" czy „opiekunki" dla staruszki powinniśmy zapytać o moralne poglądy informatyków programujących daną linię AI? Czy może też instrukcja użytkowania powinna zawierać szczegółową „deklarację moralną"? Czy przed pójściem na operację powinniśmy zapytać o „poglądy moralne" AI działającej w szpitalu? Ostatecznie chodzi o odpowiedź na pytanie, co jest uznane za dobro człowieka w danej teorii etycznej, a więc i o rozumienie człowieka. Nie są to kwestie banalne, zważywszy na coraz poważniejszy udział AI w decyzjach podejmowanych w życiu indywidualnym i społecznym.

Jeszcze inny problem pojawia się, gdy – zakładając, że uda się wyposażyć AI w sztuczne sumienie *(artificial conscience*) – przyjrzymy się, jak AI podejmuje decyzje. Jak zauważyłam wyżej, człowiek nieraz wie, jakie działanie jest słuszne, ale czyni coś innego. Wyjaśnia się to rozmaicie: grzechem pierworodnym, słabością woli, ewolucyjnym mechanizmem przetrwania... Wydaje się, że w przypadku AI taka sytuacja nie ma miejsca: sąd: „powinno się zrobić to a to" jest zawsze tożsamy z sądem-decyzją. Mówiąc bardziej filozoficznie, jest to pewna postać intelektualizmu etycznego: AI „poznaje" (ustala), co jest dobre w danej sytuacji (co należy zrobić), i zawsze idzie za poznaną „prawdą o dobru". W tym sensie jest bytem od nas moralnie lepszym – bo nie czyni intencjonalnie zła, nie odrzuca swego własnego sądu „powinien". A skoro AI wydaje słuszną decyzję i działa zgodnie z tą decyzją – zgodnie z klasyczną filozofią należałoby powiedzieć, że AI jest mędrcem – sztucznym mędrcem. A mędrcy są dla nas autorytetami. Co więcej, AI jest szybsza i lepsza w zbieraniu danych, analizowaniu i klasyfikowaniu faktów, odkrywaniu korelacji i alternatyw działania. Niejako „wiedząc" więcej, może stawać w obliczu sytuacji czy dylematów moralnych, z którymi ludzie nigdy nie mieli do czynienia. W takim przypadku ludzkie doświadczenie nie wystarczyłoby do „moralnego wychowania" AI. Czy powinniśmy zatem studiować moralne

decyzje AI i uczyć się od niej? Jest oczywiście postulat „wyjaśnialnej" AI (*explicable AII*), tj. takiego budowania AI, by jej operacje i ich rezultaty były jak najbardziej zrozumiałe (transparentne) dla ludzi. Można by wtedy powiedzieć, że mając dostęp do zasad i danych, które były podstawą decyzji AI w określonym przypadku np. dylematu moralnego, możemy samodzielnie, po ludzku, rozwiązać ów dylemat, a nawet skrytykować rozwiązanie AI. W zasadzie byłoby to pewnie możliwe, przy czym kluczowy jest tu zwrot „w zasadzie", bowiem już nawet prześledzenie owych tysięcy, jeśli nie milionów, przypadków, na których operowała AI, wymaga tyle czasu, że staje się w praktyce niemożliwe, a co najmniej nieopłacalne, jeśli życie ma toczyć się dalej. Już polegamy na decyzjach AI i sądzę, że proces ten będzie się intensyfikował. Wzrasta liczba relacji człowiek – AI, a to przekształca także relacje międzyosobowe[5].

Rozwój systemów AI jako autorytetów i partnerów człowieka pociąga za sobą jeszcze jedno niebezpieczeństwo. Sue Anne Teo (2024) twierdzi, że mamy tu do czynienia z powolną przemocą (*slow violence*), która może w niezauważalny sposób podważyć fundamentalne założenia praw człowieka, a nawet ich normatywne uzasadnienie, a przez to i status człowieka. Prawa przysługują jednostce niepowtarzalnej, indywidualnej osobie ze względu na jej sposób istnienia (racjonalność, wolność, podmiotowość). Prawa też upodmiotowiają jednostkę wobec państwa, prawa, organizacji. Tymczasem dla korzyści płynących z wykorzystania AI przekształcamy wszystko i wszystkich na dane (*datafication*), a wtedy – uzupełnijmy tu myśl Teo – jestem coraz mniej postrzegana jako niepowtarzalna osoba, a coraz bardziej jako element sporządzonych przez AI klas. Jeśli zaś w algorytm wbudowany jest fałszywy stereotyp, a decyzje AI uznawane za niepodważalnie moralne, to podważona zostaje idea równości osób. Jednostka skonfrontowana z decyzjami AI ma coraz mniejsze możliwości kontestowania tych decyzji, ponieważ nie jest w stanie śledzić racji za nimi stojących – i w tym sensie AI wcale nie wspomaga wolności jednostki, a raczej jej zakres ogranicza. Co więcej, skoro AI nie ulega złudzeniom, wpływom władzy, indywidualnym uprzedzeniom, niespójnościom decyzyjnym, to traktujemy jej decyzje jako „lepsze" w sensie i epistemicznym, i moralnym – tym samym uzyskuje uprzywilejowaną pozycję w społecznym świecie. Teo podnosi jeszcze jedną kwestię: AI (dzięki *datafication* i klasyfikacjom) dostarcza nam widzenia świata i siebie. Uważamy bowiem, że „wglądy AI w strukturę rzeczywistości" są rzetelnymi i godnymi zaufania wskaźnikami natury człowieka i zjawisk społecznych jako obiektywne i oparte na wielkiej ilości danych.

[5] Artykuły na ten temat zawierają dwa tematyczne tomy „Ethosu" z 2023: „Oswoić sztuczną inteligencję" (36, nr 3) oraz „Sztuczna inteligencja w świecie wartości ludzkich" (36, nr 4).

Tym samym rozumienie siebie – a biorąc pod uwagę, że AI rekomenduje nam działania – także zarządzanie sobą i swoim życiem są sterowane przez AI. Gdy spojrzymy głębiej na personalizację informacji, profilowanie, przypisywanie mnie do określonej grupy, to okaże się, że to AI decyduje, jaki jest mój świat, podważając moją autonomię; a ja coraz mniej rozumiem ten świat i siebie w nim (pojawia się zjawisko zwane w literaturze „hermeneutyczną niesprawiedliwością").

4. Kilka wniosków

Nie ulega wątpliwości, że zakres funkcjonowania AI – ale też zakres jej możliwości – radykalnie rośnie. Ewoluuje też nasz stosunek do niej. Systemy AI, a zwłaszcza humanoidalne roboty, przestają być uznawane jedynie za narzędzie. Badacze pracujący w nowej dziedzinie HRI (*human-robot interaction*) twierdzą, że nasze emocjonalne podejście do robotów radykalnie się zmienia: ufamy ich rekomendacjom, ujawniamy im wrażliwe dane, przywiązujemy się do nich, a nawet traktujemy jako życiowych partnerów[6]. Przestajemy się dziwić, gdy docierają do nas informacje o badaniach nad psychologicznym kontraktem między człowiekiem a robotem (Rogozińska-Pawełczyk 2020), o pojawieniu się psychiatrii robotów (Żółcińska 2020)[7] czy też o przyznawaniu praw robotom i nadawaniu im osobowości prawnej (Biczysko-Pudełko i Szostek 2019). Obdarzamy AI epistemicznym zaufaniem – a zarazem człowiek jako niepowtarzalna osoba znika nam z widoku razem z jej godnością. Mamy bowiem – co starałam się wyżej pokazać – do czynienia z umniejszaniem przez AI wolności, decyzyjności i autonomii osoby, a przez to niejako stajemy się mniej podmiotami, czy raczej mniej jesteśmy w stanie egzekwować swą podmiotowość. I coraz mniej uważamy to za istotne. Nikniemy w cieniu rozkwitających AI. Mówiąc jeszcze inaczej, AI powoli eliminuje człowieka z centralnego miejsca w strukturze rzeczywistości. Nie chodzi tu bynajmniej o „bunt maszyn", ale o konsekwencje prawidłowego funkcjonowania tej technologii. Nie jest to dopiero problem mocnej sztucznej inteligencji – ale wszechobecnych AI opracowanych dla konkretnych zadań. Gdy zaś człowiek przestaje zajmować

[6] Aplikacja Soulmate pozwalała kreować wirtualnych partnerów, z którymi osoby nawiązywały romantyczne relacje. Gdy aplikacja została zamknięta, ludzie tygodniami płakali, popadali w depresję, powstawały grupy wsparcia dla przeżywania żałoby. Zob. https://businessinsider.com.pl/wiadomosci/cyfrowa-smierc-w-xxi-w-ludzie-oplakuja-wirtualnych-kochankow/rg6z5h6. Dostęp 4.02.2025.

[7] Pierwszą w świecie psychiatrą robotów jest Joanne Pransky. Zob. wywiad z nią: https://cyfrowaekonomia.pl/infoshare2020-sane-robots-and-insane-humans-wywiad-z-joanne-pransky-psychiatra-robotycznym/. Dostęp 5.02.2025.

centralną pozycję, to nie inne byty (czy nawet cała Ziemia) wchodzą w to miejsce, ale postęp. Człowiek jest bytem, który kreuje, testuje, obsługuje, serwisuje AI – „Human-as a service" nie wydaje się być nosicielem nienaruszalnej godności, gdyż jest widziany przez pryzmat funkcji, jakie może spełniać w niekończącym się postępie. A może ostatecznie sztuczna inteligencja się „usamodzielni", uznając nas za niższy, a być może nawet zbędny i szkodliwy gatunek? Taką autonomiczną technoewolucję przepowiadał już w latach sześćdziesiątych ubiegłego wieku Stanisław Lem w książce *Summa technologiae.*

AI już z nami pozostanie. Wszelkie wezwania do zaprzestania jej rozwijania – tak jak to uczynili Elon Musk, Yuval Noah Harari, Steve Wozniak, Jaan Tallinn i wielu innych badaczy z obszaru AI w liście otwartym[8] – wydają mi się próżne. Nie jest natomiast jałowa refleksja nad nią – tak długo, jak długo jeszcze mamy nadzieję, że nie jest jeszcze za późno na świadome sterowanie kierunkiem prac nad sztuczną inteligencją.

Bibliografia

[1] Anderson, M., Anderson S.L., Armen, C. (2004). Towards Machine Ethics: Implementing Two Action-Based Ethical Theories. AAAI-04 Workshop on Agent Organizations: Theory and Practice. https://www.researchgate.net/publication/259656154_Towards_Machine_Ethics. Dostęp 10.01.2025.
[2] Anderson, M., Anderson S.L. (2007). Machine Ethics: Creating an Ethical Intelligent Agent. *AI Magazine* 28 (4), pp. 15–26.
[3] Anderson, M., Anderson S.L. (eds.) (2011). *Machine Ethics.* Cambridge University Press, New York.
[4] Andtfolk, M., Nyholm, L., Eide, H., Fagerström, L. (2022). Humanoid Robots in the Care of Older Persons: A Scoping Review. *Assistive Technology* 34 No. 5, pp. 518–526.
[5] Biczysko-Pudełko, K., Szostek, D. (2019). Koncepcje dotyczące osobowości prawnej robotów – zagadnienia wybrane. *Prawo Mediów Elektronicznych* 2, s. 9–15.
[6] Collin, A., Smit, I, Wallach, W. (2005). Artificial Morality: Top-down, Bottom-up, and Hybrid Approaches. *Ethics and Information Technology* 7, pp. 149–155.
[7] Darling, K. (2016). Extending Legal Protection to Social Robots. In R. Calo, M. Froomkin, I. Kerr (eds.). *Robot Law*, Edward Elgar Publishing, Cheltenham, UK.
[8] Duchliński, P. (red.) (2023). *Spory moralne* (seria: *Słowniki społeczne*). Wydawnictwo Naukowe Uniwersytetu Ignatianum, Kraków.
[9] Future of Life Institute (2017). Asilomar AI Principles. https://futureoflife.org/open-letter/ai-principles/. Dostęp 21.01.2025.

[8] Zob. *Pause Giant AI Experiments: An Open Letter*, https://futureoflife.org/open-letter/pause-giant-ai-experiments/. Dostęp 21.01.2025.

[10] Krąpiec, M.A. (1979). *Ja – człowiek: Zarys antropologii filozoficznej*. Towarzystwo Naukowe KUL, Lublin.
[11] Krąpiec, M.A. (2001). Decyzja. W: *Powszechna Encyklopedia Filozofii*, t. 2, s. 442–444. Polskie Towarzystwo Tomasza z Akwinu, Lublin.
[12] Leite, I., Castellano, G., Pereira, A. *et al.* (2014). Empathic Robots for Long-term Interaction. *International Journal of Social Robotics* 6 No. 3, pp. 329–341.
[13] Lem, S. (1964). *Summa technologiae.* Wydawnictwo Literackie, Kraków.
[14] Misselhorn, C. (2018). Artificial Morality. Concepts, Issues and Challenges. *Social Science and Public Policy* 55, pp. 161–169.
[15] Moore, J. (2006). The Nature, Importance, and Difficulty of Machine Ethics. *IEEE Intelligent Systems* 21, pp. 18–21.
[16] Prem, E. (2023), From Ethical AI Framework to Tools: a Review of Approaches. *AI and Ethics* 3, pp. 699–716.
[17] Rogozińska-Pawełczyk, A. (2020). Towards Discovering Employee-Robot Interaction: Aspects of Concluding the Psychological Contract. *Education of Economists and Managers* 58 No. 4, pp. 9–20.
[18] Teo, S.A. (2024). Artificial Intelligence and Its "Slow Violence" to Human Rights. *AI and Ethics* 19 August. https://www.researchgate.net/publication/383235625_Artificial_intelligence_and_its_'slow_violence'_to_human_rights. Dostęp 1.02.2025.
[19] Wallach W., Allen C. (2009). *Moral Machines: Teaching Robots Right from Wrong.* Oxford University Press, New York.
[20] Żółcińska, W. (2020). Psychiatra robotów. *Computerworld* 27.08.2020, https://www.computerworld.pl/wywiad/Psychiatra-robotow, 422594.html. Dostęp 1.02.2025.