ISSN 2081-89-71
2024 nr 60, s. 271–286
DOI 10.21697/kmt.2023.60.15

Jarosław A. Sobkowiak
Uniwersytet Kardynała Stefana Wyszyńskiego w Warszawie
ORCID 0000-0003-3698-4366

„Actus humanus" w kontekście sztucznej inteligencji a odpowiedzialność osoby

Abstract

The aim of the presented article is to reflect on the issue of the 'actus humanus' in relation to artificial intelligence (AI) and human moral responsibility. In the first section, the characteristics of the human act are contrasted with the equivalent 'acts' of artificial intelligence, which are based on algorithms and machine learning processes. The next section outlines the technological and ethical framework for the use of artificial intelligence. The final section will discuss the proposal to limit human oversight of the machine learning process and entrust it to artificial intelligence referring exclusively to the created ethical framework of 'Constitutional AI' (CAI). The whole is rounded off with the latest proposal of 'Collective Constitutional AI' (CCAI), which aims to broaden accountability to include the social dimension of the creation of ethical frameworks and thus preserve indirect human oversight of artificial intelligence processes.

Keywords

Artificial Intelligence (AI), Constitutional AI (CAI), Collective Constitutional AI (CCAI), human moral responsibility, technological and ethical framework, 'actus humanus' – artificial intelligence

Abstrakt

Celem prezentowanego artykułu jest refleksja nad zagadnieniem „actus humanus" w odniesieniu do sztucznej inteligencji (AI) oraz odpowiedzialności moralnej człowieka. W pierwszej części dokonuje się zestawienia cech charakterystycznych dla aktu ludzkiego z odpowiednikiem „aktów" sztucznej inteligencji, które są oparte na algorytmach i procesach uczenia maszynowego. W dalszej części zostają nakreślone technologiczne i etyczne ramy wykorzystania sztucznej inteligencji. W ostatniej części zostanie omówiona propozycja ograniczenia nadzoru człowieka nad procesem uczenia maszynowego i powierzenie go sztucznej inteligencji odwołującej się wyłącznie do stworzonych ram etycznych (CAI). Całość zostaje dopełniona najnowszą propozycją „Collective Constitutional AI" (CCAI), która ma poszerzyć odpowiedzialność o społeczny wymiar tworzenia ram etycznych i w ten sposób zachować pośredni nadzór człowieka nad procesami sztucznej inteligencji.

Słowa kluczowe

Sztuczna inteligencja (AI), Constitutional AI (CAI), Collective Constitutional AI (CCAI), odpowiedzialność moralna człowieka, ramy technologiczne i etyczne, „actus humanus" sztucznej inteligencji

Wstęp

Prezentowany artykuł stanowi dopełnienie dwóch poprzednich tekstów poświęconych filozofii głupoty, a następnie przejściu od filozofii głupoty do sztucznej inteligencji, by w obecnie prezentowanym tekście postawić problem w najbardziej skrajnym kontekście – skoro coraz częściej pytamy o swoistą zastępowalność „człowiek – sztuczna inteligencja", czy zatem istnieje w jej działaniu odpowiednik „aktu ludzkiego", a jeśli nie, to czy decydujemy się na świadome „odczłowieczenie" szerokiego spektrum działań, jakich jeśli nie podmiotem, to z pewnością przedmiotem będzie człowiek?

W ocenie działania ludzkiego rozpatrujemy bowiem dwa typy działań: *actus hominis* i *actus humanus*. Oczywiście jeden i drugi akt wykonuje człowiek, z tą różnicą, że pierwszy jest dla człowieka czymś jakby zewnętrznym, nieangażującym tego, co dla niego najistotniejsze – obejmuje wszystkie działania mechaniczne, odniesione do podstawowych czynności życiowych – w tym zakresie może człowieka zastąpić sztuczna inteligencja. Drugi zaś typ aktu działania angażuje człowieka na sposób dla niego specyficzny, odwołując się do świadomości, wolności, motywacji, a w dalszej perspektywie naznaczony kreatywnością i empatią. Warto od razu rozwinąć ten wątek, który bardzo precyzyjnie wybrzmiewa w myśli św. Tomasza z Akwinu, czyli jak najbardziej klasycznym podejściu do aktu ludzkiego. Akwinata wskazuje, że akt ludzki musi być świadomy i wolny, gdyż tylko taki można poddać ocenie moralnej. W przeciwnym razie mamy do czynienia z drugim, wspomnianym już typem działania, jakim jest *actus hominis*, czyli akt niespecyficzny dla człowieka. Warto dodać, że powyższa myśl znajdowała swoich kontynuatorów[1].

Wolność aktu ludzkiego rozumiana jest jako synonim dobrowolności i niezdeterminowania. Co ważne, Tomasz z Akwinu odnosi wolność wyłącznie do bytu ludzkiego, a odnoszenie jej do czegokolwiek poza człowiekiem rozumie wyłącznie przez analogię. O wolności ludzkiej powie wprost – wolne jest to, co samo dla siebie jest przyczyną, a co dzisiaj określamy terminem podmiotowość[2].

W prezentowanym artykule podejmie się tę właśnie ścieżkę refleksji w odniesieniu do sztucznej inteligencji, najpierw zestawiając akt ludzkiego działania z aktem sztucznej inteligencji, stawiając przy tym pytanie czy i na ile działanie sztucznej inteligencji można porównać z działaniem ludzkim, nawet przez analogię. Postawi się pytanie, na ile sama sztuczna inteligencja może być „odpowiedzialna" za podejmowane działania i na ile człowiek może w tym kontekście uczynić siebie „nieodpowiedzialnym", czyli pozbawionym poczucia odpowiedzialności. Po tych teoretycznych rozważaniach podejmie się próbę nakreślenia ram technologicznych i etycznych dla zachowania odpowiedzialności (nadzoru) człowieka w odniesieniu do sztucznej inteligencji. W ostatniej części dokona się

[1] Godna polecenia w tym względzie, wychodząca w refleksji poza propozycje św. Tomasza jest książka Thomasa M. Osborne Jr, *Human Action in Thomas Aquinas, John Duns Scotus & William of Ockham*, Washington DC: The Catholic University of America Press 2014.

[2] A. Andrzejuk, *Wolność w doktrynie Tomasza z Akwinu*, w: *Wolność człowieka i jej granice. Antologia pojęcia w doktrynach polityczno-prawnych. Od Starożytności do Monteskiusza,* Uniwersytet Łódzki, Łódź 2019, s. 151-152.

omówienia bardzo istotnego tekstu Yuntao Bai i zespołu Anthropic dla zagadnienia uczenia maszynowego, poświęconego poziomowi nadzoru człowieka nad sztuczną inteligencją, ze wskazaniem, że przy zachowaniu pewnych ram konstytutywnych i etycznych działanie to można ograniczyć do minimum bez szkody dla człowieka.

„Actus humanus" – pomiędzy człowiekiem a sztuczną inteligencją

Brian Davies w rozdziale 17 książki „The Oxford Handbook of Aquinas" precyzyjnie omawia podstawowe różnice, jakie zachodzą pomiędzy „actus humanus" i „actus hominis". Pierwszy można rozumieć jako „czyn ludzki", drugi zaś jako „czyn człowieka". Czyn ludzki cechuje celowość i wynika on z działania rozumu praktycznego nastawionego na osiągnięcie określonego dobra[3]. Z kolei Thomas Williams analizuje akt ludzki w kontekście wolności. Wolność zaś jest swoistym dialogiem pomiędzy rozumem i wolą, co czyni z człowieka w takim działaniu „pana swoich działań" (*dominus suorum actuum*)[4]. Z kolei Christopher Alexander Franke i Joelma Marques de Carvalho poddają refleksji „actus humanus" pod kątem działania intelektu i woli, zaznaczając różnicę pomiędzy obszarem naturalnych zdarzeń (*genus naturae*) a sferą działania moralnego (genus moris), tym samym zauważając, że tylko akt ludzki odgrywa kluczową rolę w działaniu człowieka[5].

Należy jednak zdawać sobie sprawę z faktu, że nie wszystkie nauki – jak filozofia czy teologia – tak właśnie podchodzą do problematyki woli czy wolności. Edward Nęcka proponuje zauważenie tej różnicy w trzech opcjach. Pierwsza opcja to właśnie wspomniane już dyscypliny, które odżegnując się od typowego pojęcia *science*, tym samym pokazują niewystarczalność badania problemu woli i wolności „szkiełkiem i okiem". Druga opcja to badanie zjawiska „na obrzeżach", czyli odżegnując się od przypisywania woli atrybutu wolności, skupiając się wyłącznie na „poczuciu" wolności. Oczywiście poczucie jest czymś subiektywnym, niemniej występuje na tyle powszechnie, że można je badać w sposób obiektywny (ankietą, wywiadem, eksperymentem). Trzecia opcja, to popularne we współczesnej psychologii stanowisko, pokazujące złudność ludzkiej wolności – jak opisuje Nęcka – np. robimy zakupy tylko dlatego, że inni też to robią[6]. Takie podejście do ludzkiej wolności zaczyna kreślić nieco szerszy obszar pod pytanie, czy sztuczna inteligencja może być zdolna, przynajmniej w jakimś wymiarze, do tego, co przypomina akt ludzki w ścisłym rozumieniu. W podobnym duchu wypowiada się wielu autorów, wykazując swoisty pat badawczy, który polega na tym, że zajmujemy się problematyką ludzkiej woli od wielu wieków i ciągle nie ma takiej definicji woli, na którą mogłaby się

[3] B. Davies, *Happiness,* w: *The Oxford Handbook of Aquinas*, rozdział 17, ed. B. Davies i E. Stump, Oxford University Press, 2012, s. 227n.

[4] Th. Williams, *Human Freedom and Agency,* w: *The Oxford Handbook of Aquinas*, rozdział 15, s. 199n.

[5] Ch.A. Franke, J. Marques de Carvalho, *Das Wesen der menschlichen Handlung bei Thomas von Aquin (The Essensce of Human Action in Thomas Aquinas),* „Revista Portuguesa de Filozofia", 2023, vol. 79 (1-2), s. 479n.

[6] E. Nęcka, *Wolna wola czy wolne weto: rola świadomości w czynnościach wolicjonalnych*, s. 11-13, https://www.publikacje.pan.pl/Content/117527/PDF/Necka.pdf, (dostęp 11.09.2024).

zgodzić większość nauk. W tym duchu wypowiadają się również E. Nęcka i J. Prusak[7], chociaż poza wskazaniem problemu, również niewiele wnoszą do bardziej precyzyjnego zdefiniowania woli. Przywołuję więc ten artykuł na potwierdzenie, że w dialogu filozofii i teologii z naukami eksperymentalnymi mamy dwa wyjścia: pomniejszać znaczenie refleksji i precyzji na temat woli w filozofii i teologii lub przyjąć, że dopóki nauki eksperymentalne nie wypracują precyzyjnej definicji woli (a skoro zauważa się jej przejawy, to rozum nakazuje ją zakładać), pozostaje przyjąć, że wypracowane filozoficzno-teologiczne stanowisko jest przynajmniej spójne i na poziomie teoretycznym pozwala na dalsze refleksje (J. Pastuszka[8], K. Jasiński[9], A. Karaś[10]).

Należałoby teraz prześledzić, jak przebiega działanie sztucznej inteligencji w kluczu „actus humanus". Jest to oczywiście opis przez analogię i nie można w jej działaniu w pełni odzwierciedlić ani odpowiednika władz w człowieku, ani też zależności, jakie w człowieku zachodzą między nimi. Mówiąc o odpowiedzialności sztucznej inteligencji, kładzie się nacisk nie tyle na jej autonomiczne działanie, co raczej na współdziałanie człowieka z nią. Tylko w tym wymiarze można mówić o jakiejkolwiek odpowiedzialności moralnej. Zatem ocenia się nie tyle działanie sztucznej inteligencji, co raczej interakcje, jakie zachodzą na styku działania: człowiek – sztuczna inteligencja[11].

Jak powinna przebiegać owa współpraca ze sztuczną inteligencją, by można mówić o jakiejkolwiek odpowiedzialności moralnej? Już wstępne prace nad dokumentem Parlamentu Europejskiego wskazywały jednoznacznie, że o odpowiedzialności AI można mówić wyłącznie w powiązaniu z człowiekiem. W przeciwnym razie wszelkie szanse jakie za sobą niesie mogą zostać zaprzepaszczone w sytuacjach nieprzewidywalnych, w których nie można by jednoznacznie wskazać podmiotu odpowiedzialności moralnej[12].

Powiązanie działania sztucznej inteligencji i człowieka w płaszczyźnie odpowiedzialności moralnej jest tak istotne, ponieważ sztuczna inteligencja jest zaprojektowana do osiągania konkretnych celów. Cele te są jednak definiowane przez twórców. Biorąc więc

7 E. Nęcka, J. Prusak, *Eksperymentalna psychologia woli : wolność i intencjonalność z perspektywy psychologii poznawczej i neuronauki*, 2016, s. 45-47.
https://www.academia.edu/66683464/Eksperymentalna_psychologia_woli_wolno%C5%9B%C4%87_i_intencjonalno%C5%9B%C4%87_z_perspektywy_psychologii_poznawczej_i_neuronauki, (dostęp 11.09.2024)

8 J. Pastuszka, *Filozoficzne i empiryczne pojęcie osoby ludzkiej*, „Roczniki Filozoficzne" vol. 11 (1963) nr 4, s. 45-60.

9 K. Jasiński, *Czyn doskonalący osobę. W kręgu perfekcjonizmu Karola Wojtyły*, „Studia Elbląskie" XIII(2012), s. 351-361.

10 A. Karaś, *Struktura aktu moralnego w ujęciu Mieczysława Gogacza*, „Studia Philosophiae Christianae" 47(2011) nr 2, s. 157-173.

11 H. Sheikh, C. Prins, E. Schrijvers, *Mission AI. The New System Technology*, Springer (Open Acces) 2023, część I, rozdział 2: *Artificial Inelligence: Definition and Background*, s. 20-22, https://link.springer.com/chapter/10.1007/978-3-031-21448-6_2 (dostęp 12.09.2024).

12 Parlament Europejski: *Sztuczna inteligencja: szanse i zagrożenia*, 20.06.2023, s. 4, https://www.europarl.europa.eu/pdfs/news/expert/2020/9/story/20200918STO87404/20200918STO87404_pl.pdf (dostęp 12.09.2024).

pod uwagę powstawanie zarówno skutków pozytywnych, jak i negatywnych, istotne jest rozważenie wszelkich implikacji etycznych, których podstawowym kierunkiem jest zabezpieczenie dobra całego społeczeństwa[13]. Dodatkowym argumentem wskazującym, że to ciągle odpowiedzialność osoby, jest założenie, że systemy AI są zorientowane na cel, którym jest realizacja określonych zadań lub rozwiązywanie konkretnych problemów. Istotnym kryterium etycznym musi pozostać zrównoważenie innowacji z odpowiedzialnością[14]. W tym kontekście można zaproponować kryteria, które w klasycznej refleksji etycznej były odnoszone do czynu o podwójnym skutku, a z takim właśnie mamy do czynienia w odniesieniu do zastosowania sztucznej inteligencji. Wspomniane kryteria brzmiały następująco: czyn sam w sobie musi być moralnie dobry lub obojętny; skutek bezpośredni czynu musi być dobry; skutek bezpośredni (dobry) jest celem działającego; istnieją proporcjonalnie ważne racje dla spełnienia tego czynu. Dopiero po przyłożeniu takich kryteriów do oceny konkretnego systemu AI można mówić o przeniesieniu refleksji na poziom moralny. Bez jasnego zdefiniowania oczekiwań etycznych sam imperatyw techniczny – możliwe techniczne staje się powinne moralnie – nie wystarczy.

Innym elementem, który musi być brany pod uwagę w ocenie moralnej, jest porównanie ludzkiej kreatywności z kreatywnością AI. Ludzka kreatywność jest zakorzeniona w emocjach i osobistych doświadczeniach, co nadaje jej unikalny charakter. Mówiąc o oryginalnym charakterze działalności człowieka, bierze się pod uwagę intuicję i spontaniczność. To one prowadzą do spontanicznych (czyli w zestawieniu z AI – nieprzewidywalnych), oryginalnych i twórczych rozwiązań. Ludzka kreatywność jest też bardzo mocno związana z kontekstem kulturowym i społecznym i również w tych kontekstach jest interpretowana. Z kolei sztuczna inteligencja opiera swoją kreatywność na algorytmach i dużych zbiorach danych, co pozwala na generowanie nowych, ale jednak nie oryginalnych treści, gdyż powstają one w oparciu o istniejące już wzorce. Niewątpliwie mocną stroną AI jest automatyzacja pewnych procesów, jednak pozbawienie treści odniesienia do emocji i doświadczeń sprawia, że jest ta twórczość mniej autentyczna i oryginalna. Szczególnie jest to widoczne w obszarze sztuki, gdzie brak takich typowo ludzkich doznań jak dotyk czy emocjonalna głębia[15]. Mamy wystarczająco dużo przykładów generowania treści i dzieł artystycznych przez sztuczną inteligencję (Generatywna sztuczna inteligencja – GenAI). Jednak jak zauważa A. Łukawski, „ważne jest rozróżnienie między

[13] H. Sheikh, C. Prins, E. Schrijvers, *Mission AI. The New System Technology*, dz. cyt., s. 30 i 32.

[14] M. Coeckelbergh, *Artificial intelligence, the common good, and the democratic deficit in AI governance*, Springer, Published online 22.05.2024, s. 4-6, link do artykułu https://link.springer.com/article/10.1007/s43681-024-00492-9 (dostęp 12.08.2024).

[15] L. Perska, *Zatarte granice między sztuczną inteligencją a ludzką kreatywnością*, Elblog, 16.06.2024, https://elblog.pl/pl/2024/06/16/zatarte-granice-miedzy-sztuczna-inteligencja-a-ludzka-kreatywnoscia/ (dostęp 13.09.2024).

kreatywnością wynikającą z prawdziwej innowacyjności a kreatywnością będącą rezultatem przetwarzania i rekonfiguracji istniejących danych”[16].

W czym zatem wyraża się specyfika aktu sztucznej inteligencji? Aktem AI jest nie tyle samo działanie (jak w przypadku człowieka), ale raczej „odniesienie”. To odniesienie, inaczej niż u człowieka, nie odnosi się do rozumu i woli, ale do algorytmów i procesów uczenia maszynowego. Najbardziej sztuczna inteligencja przypomina człowieka w tym, że się uczy, to znaczy jest zdolna do przeprogramowywania danych i poprawę ich wydajności bez podchodzenia na nowo do każdego kolejnego zadania. Interesujące w tym względzie jest stanowisko Polski po przyjęciu „Artificial Intelligence Act”[17] (13.03.2024). Polska wskazała 13 kluczowych kierunków rozwoju sztucznej inteligencji. Warto przywołać najważniejsze: ocena ryzyka, lista zakazanych praktyk, nadzór człowieka nad systemami wysokiego ryzyka, nałożenie obowiązku gwarantowania przejrzystości, uwzględnienie ochrony środowiska i wpływu na prawa podstawowe czy wprowadzenie kodeksu dobrych praktyk (code of practice)[18].

Ramy technologiczne i etyczne dla zachowania nadzoru człowieka nad „aktem” sztucznej inteligencji

Do zobrazowania zakresów nadzoru człowieka nad systemami sztucznej inteligencji niezbędne jest, przynajmniej w zalążku, nakreślenie typów i poziomów zaawansowania AI. Najbardziej popularny podział wyraża się we wskazaniu jej trzech typów: Artificial Narrow Intelligence (ANI), Artificial General Intelligence (AGI) oraz Artificial Super Intelligence (ASI). Pierwszy typ – to zresztą jedyny jaki istnieje obecnie – sprowadza się do systemów rozpoznawania mowy, autonomicznych pojazdów czy asystentów głosowych[19]. Drugi typ zakłada rozwój sztucznej inteligencji na poziomie maszyn świadomych kontekstu[20]. Trzeci typ odnosi się do przewyższenia zdolności człowieka na praktycznie wszystkich poziomach, nawet zakładając rozwój własnych emocji i potrzeb AI. Nie

[16] A. Łukawski, *Generatywna sztuczna inteligencja (GenAI) a kreatywność*, Zintegrowana Platforma Edukacyjna Ministerstwa Edukacji Narodowej, https://zpe.gov.pl/a/i-o-kreatywnosci-generatywnej-sztucznej-inteligencji-11-generatywna-sztuczna-inteligencja-genai-a-kreatywnosc/D16yodZxY, (dostęp 13.09.2024).

[17] Rozporządzenie Parlamentu Europejskiego i Rady (UE) 2024/1689 z dnia 13 czerwca 2024 r. w sprawie ustanowienia zharmonizowanych przepisów dotyczących sztucznej inteligencji oraz zmiany rozporządzeń (WE) nr 300/2008, (UE) nr 167/2013, (UE) nr 168/2013, (UE) 2018/858, (UE) 2018/1139 i (UE) 2019/2144 oraz dyrektyw 2014/90/UE, (UE) 2016/797 i (UE) 2020/1828 (akt w sprawie sztucznej inteligencji) (Tekst mający znaczenie dla EOG), European Union, EUR-Lex, https://eur-lex.europa.eu/legal-content/PL/TXT/?uri=CELEX:32024R1689, (dostęp 14.06.2024).

[18] Polskie stanowisko w sprawie AI Act, cyt. za https://www.gov.pl/web/ai/rewolucja-w-regulacji-wchodzi-w-zycie-akt-o-ai (dostęp 14.09.2024).

[19] S. Fourtané, *The Three Types of Artificial Intelligence: Understanding AI*, https://ir.westcliff.edu/wp-content/uploads/2020/01/The-Three-Types-of-Artificial-Intelligence-Understanding-AI.pdf, (dostęp 11.07.2024).

[20] Ch. Manning, *Artificial Intelligence Definitions*, Stanfort University, Human-Centered Artificial Intelligence, September 2020, https://hai.stanford.edu/sites/default/files/2020-09/AI-Definitions-HAI.pdf, (dostęp 12.04.2021).

trzeba dodawać, że ten typ sztucznej inteligencji budzi najwięcej obaw etycznych i zagrożeń[21]. Chociaż oczywiście nie można wykluczyć zagrożeń nawet na najniższym poziomie, o czym świadczą bieżące doniesienia prasowe. Jednym z ostatnich przypadków tego typu było nadpisywanie kodów wbrew czy obok świadomości i nadzoru człowieka, co w ocenie specjalistów jest rzeczą niedopuszczalną[22]

Na obecnym etapie rozwoju sztucznej inteligencji można stwierdzić, że systemy te nie mogą być moralnie odpowiedzialne, gdyż stworzył je człowiek. Odpowiedzialność ta obejmuje zakres od projektowania po implementację. Pojawia się jednak, wraz z rozwojem AI, problem coraz mniejszego kontrolowania systemu przez człowieka. Już na początku obecnego wieku przewidywano w tym względzie tzw. „luki odpowiedzialności" (responsability gap)[23]. Wszystko to pokazuje, jak wielka odpowiedzialność musi spoczywać na ludziach i jak po raz kolejny potwierdza się teza, że działanie AI można nazwać aktem wyłącznie w zestawieniu z nadzorującym człowiekiem. Pytaniem otwartym pozostaje, kiedy człowiek utraci (jeśli ciągle funkcjonujemy jeszcze w czasie przyszłym) realny nadzór nad pracą systemów?

W odniesieniu do powyższych refleksji wydaje się sprawą oczywistą, że człowiek – pod żadnym pozorem – nie może zwolnić się od odpowiedzialności. Zamiast analizować problem, na ile sztuczna inteligencja przejmie dotychczasowe zadania człowieka, może warto postawić pytanie, na ile w obliczu rozwoju AI to człowiek musi uświadomić sobie nowe obszary własnej odpowiedzialności? Można w tym kluczu odnotować dwa kierunki, które dotyczą wymiaru społecznego wpływu sztucznej inteligencji. Pierwszy wiąże się z rozwojem i kierunkiem przyszłych badań, które pozwolą na jeszcze lepsze wykorzystanie AI w wymiarze społecznym. Drugi wymiar bierze pod uwagę niebezpieczeństwa i przewidywanie skutków ubocznych jej zastosowania. W pierwszym obszarze można wyróżnić wpływ na szeroko pojętą służbę zdrowia, od profilaktyki poczynając, poprzez diagnostykę, aż po działania opiekuńcze nad chorymi. Ponadto AI może wspierać ochronę środowiska, monitorowanie przestrzeni miejskich oraz zarządzanie urbanizacją i przemysłem. Dotyczą także mobilności i bezpieczeństwa transportu. Drugi obszar zagrożeń musi być szczególnie analizowany przez różne organizacje międzynarodowe jak Organizacja Narodów Zjednoczonych, Unia Europejska czy inne o zasięgu ponadnarodowym[24]. W odniesieniu do wspomnianych organizacji warto powiązać działania AI z Agendą 2030 ONZ przyjętą rezolucją z dnia 25 września 2015 roku[25]. W tym kontekście (17 celów

[21] S. Fourtané, *The Three Types of Artificial Intelligence: Understanding AI*, art. cyt.

[22] M. Tomanek, *Zagrożenia sztucznej inteligencji. Model AI modyfikował swój kod*, Holistic News, 14.10.2024, https://holistic.news/model-ai-sam-chcial-zmienic-swoj-kod-naukowcy-sa-zaskoczeni/ (dostęp 14.10.2024).

[23] M. Constantinescu, C. Voinea, R. Uszkai, C. Vică, *Understanding responsibility in Responsible AI. Dianoetic virtues and the hard problem of context*, „Ethics and Information Technology" (2021), s. 807, https://link.springer.com/article/10.1007/s10676-021-09616-9, (dostęp 15.09.2024).

[24] M. Ghallab, *Responsible AI: requirements and challenges*, Springer Open, 3.09.2019 https://aiperspectives.springeropen.com/articles/10.1186/s42467-019-0003-z (dostęp 17.09.2024).

[25] *Agenda 2030 na rzecz zrównoważonego rozwoju (Transforming our world: the 2030 Agenda for Sustainable Development)*, tekst polski cyt. za https://www.gov.pl/web/rozwoj-technologia/agenda-2030 (dostęp 12.07.2021).

wskazanych w Agendzie) na związek wymiaru społecznego i sztucznej inteligencji wskazuje się w obszarze realizacji następujących celów: interpretacja i przetwarzanie zdjęć satelitarnych, szczególnie w aspekcie pozyskiwania żywności i rozwoju rolnictwa (SDG 2) oraz w programach ochrony środowiska (SDG 6, 13, 17); szybsze udostępnianie i interpretacja danych medycznych, szczególnie w aspekcie pandemii i innych zagrożeń epidemiologicznych (SDG 3), a także wspieranie procesu rozwoju tzw. inteligentnych miast (SDG 11)[26].

Mówiąc o odpowiedzialności za wykorzystanie sztucznej inteligencji, warto sięgnąć do modeli już wypracowanych, z całą świadomością faktu, że przygotowanie poszczególnych krajów, w tym krajów członkowskich Unii Europejskiej nie jest w tym względzie jednorodne. Niewątpliwie do liderów wdrażania ram etycznych i prawnych w zakresie AI należy Holandia, która w Global Index on Responsible AI[27] zajęła pierwsze miejsce. Do zasadniczych punktów rozwiązań holenderskich można zaliczyć: transparentność i odpowiedzialność w zarządzaniu AI – szczególnie jest to realizowane przez kampanie informacyjne oraz „open instruments", pozwalające na wgląd w to w jaki sposób powstają decyzje w oparciu o sztuczną inteligencję; wprowadzono ogólnokrajowy rejestr algorytmów, dzięki czemu obywatel ma pełną świadomość jakie algorytmy zostają wykorzystywane w decyzjach społecznych; znaczenie zaufania publicznego do rozwiązań AI – w tym momencie wraca klasyczne przekonanie o zaufaniu jako warunku komunikacji[28]; niski poziom nieufności wobec nowych technologii – co jest charakterystyczne dla społeczeństwa holenderskiego[29]. Oczywiście należy nieco ostudzić emocje związane z rozwiązaniami holenderskimi, gdyż można w nich znaleźć jeszcze wiele braków. Jednym z nich jest to, że wspomniany rejestr algorytmów jest opracowany na bardzo dużym poziomie ogólności. Bardziej szczegółowe wyszukiwanie nie przynosi oczekiwanych rezultatów. We wprowadzaniu algorytmów nie obyło się bez niespodzianek, szczególnie w sektorze podatkowym[30]. Inną bolączką rejestru jest to, że pomimo iż Holandię uważa się za lidera w tej kwestii, przez długi czas rejestr funkcjonował wyłącznie w języku niderlandzkim, co – biorąc pod uwagę znaczenie tematyki – wydaje się czymś dziwnym, zwłaszcza, że na tych rozwiązaniach mogłaby się wzorować, a przynajmniej uczyć na błędach całą Unia Europejska[31]. Są oczywiście pewne rozwiązania budzące nadzieję w tym względzie, cho-

[26] M. Ghallab, *Responsible AI: requirements and challenges*, art. cyt.

[27] D.M. Popa, *Frontrunner model for responsible AI governance in the public sector: the Dutch perspective*, 27.09.2024, https://link.springer.com/article/10.1007/s43681-024-00596-2 (dostęp 1.10.2024).

[28] Pisałem o tym szerzej – J.A. Sobkowiak, *Trust as a Condition of Getting to Know the Truth: The Anthropological Aspect* – „Studia Theologica Varsaviensia", 2016 nr 2, s. 195-208.

[29] D.M. Popa, *Frontrunner model for responsible AI governance in the public sector: the Dutch perspective*, art. cyt., s. 1-3.

[30] A. Stępień, E. Dęmska, *Postępowania przyszłości. Holenderski algorytm ostrzeżeniem dla Europy*, Crido: blog podatkowy, 1.02.2023 https://crido.pl/blog-taxes/postepowania-przyszlosci-holenderski-algorytm-ostrzezeniem-dla-europy/ (dostęp 2.10.2024).

[31] D.M. Popa, *Frontrunner model for responsible AI governance in the public sector: the Dutch perspective*, art. cyt., s. 9-10.

ciażby rejestr algorytmów w Amsterdamie, funkcjonujący już zarówno w języku niderlandzkim, jak również angielskim[32].

Powróćmy zatem do wątku wyjściowego niniejszego artykułu, do próby porównania aktu ludzkiego z aktem sztucznej inteligencji. W akcie ludzkim to, co istotne, to przedmiot, intencja i okoliczności. W literaturze przedmiotu dotyczącej klasycznych rozwiązań etycznych dwa pierwsze punkty wydają się dość oczywiste. Nawet jeśli przyjmiemy, że motywacja człowieka i motywacja sztucznej inteligencji różnią się zasadniczo, chociażby w odniesieniu do całej warstwy biologicznej, której nie zakłada sztuczna inteligencja, czy w dalszej perspektywie wizja postbiologicznej cywilizacji, wydaje się, że tym, co najbardziej wpływa na zbliżanie się do siebie „aktów" człowieka i maszyny, jest poziom okoliczności. To one sprawiają, że w odniesieniu do nich (czyli w reakcji na), nie zaś w działaniu podmiotowym (działaniu świadomym i wolnym) akt ludzki i akt AI stają się najbardziej podobne. I w tym wymiarze sztuczna inteligencja w pewien sposób trafniej, szybciej i skuteczniej reaguje na bodźce płynące ze świata zewnętrznego[33].

W klasycznym rozumieniu okoliczności sprowadzały się do kilku fundamentalnych pytań: kto, co, gdzie, dlaczego, kiedy, jakimi środkami, w jaki sposób[34]. Nieco inaczej należy podchodzić do okoliczności, na które „reaguje" sztuczna inteligencja. Z punktu widzenia obszaru poszukiwań niniejszego artykułu ważne są przede wszystkim okoliczności technologiczne, a wśród nich dostęp do danych. Jak zauważa wielu autorów, dane nie są neutralne. Zawsze dotyczą czyjejś prywatności, ich zdobywanie może odbywać się na wiele sposobów. Chociaż więc sztuczna inteligencja korzysta z pozornie neutralnych informacji i danych, to sposób ich pozyskiwania i przetwarzania nie jest już neutralny[35]. Wśród najważniejszych kwestii etycznych wskazuje się na wykluczenie słabszych grup społecznych, gromadzenie danych i obawy dotyczące ich prywatności oraz szeroko pojęte zarządzanie algorytmiczne, które dotyczy bezpośrednio np. losu pracownika, a jest zupełnie pozbawione wymiaru ludzkiego, chociażby empatii. Algorytmy mogą również wykrywać prywatne dane zanim zostaną wprowadzone do modelu[36]. O pewnych systemowych niebezpieczeństwach AI wspominałem już, mówiąc o nadpisywaniu kodów przez system.

Warto również zwrócić uwagę na interesujący wątek powiązania teorii sprawiedliwości J. Rawlsa z obecnymi poszukiwaniami podstaw etycznych. Wśród zasadniczych

[32] City of Amsterdam Algorithm Register, https://algoritmeregister.amsterdam.nl/en/ai-register/ (dostęp 2.10.2024).

[33] J. Pilsner, *Circumstance*, w: *The Specification of Human Actions in St Thomas Aquinas* (ed. J. Pilsner), Oxford 2006, s. 172-298.

[34] T. Zadykowicz, Hasło: *Czyn ludzki*, w: *Jan Paweł II. Encyklopedia nauczania moralnego*, red. J. Nagórny, K. Jeżyna, Polskie Wydawnictwo Encyklopedyczne, Radom 2005, s. 131.

[35] *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence* by Kate Crawford Yale University Press, New Haven, CT, U.S.A 2021, polskie wydanie K. Crawford, *Atlas sztucznej inteligencji. Władza, pieniądze i środowisko naturalne*, wyd. Bo.wiem, Kraków 2024, szczególnie rozdział 3: *Dane*, s. 97.

[36] R. Townsent, P. Kostro, *Etyczne podejście do danych i AI ma biznesowy sens*, MITSloan, Management Review Polska, 8.09.2023, https://mitsmr.pl/b/etyczne-podejscie-do-danych-i-ai-ma-biznesowy-sens/PCr83KVmR (dostęp 4.10.2024).

kwestii zauważa się między innymi równość wolności podstawowych (Basic Liberties). Jest to nawiązanie do tradycyjnego nauczania Rawlsa, które polegało na tym, że wszelki postęp praw może dokonywać się tylko wtedy, jeśli w punkcie wyjście zapewni się wszystkim wolności podstawowe[37]. Innym wymiarem jest zasada równości szans (Equality of Opportunity)[38]. Rawls w swojej teorii stosując zasłonę niewiedzy, dążył do tego, aby wyeliminować uprzedzenia. W tym znaczeniu należy zadbać o równy dostęp do rekrutacji, oceny wyników pracy czy edukacji. Kolejnym czynnikiem jest dopuszczanie różnic (Difference Principle)[39]. W teorii Rawlsa przyjmuje się założenie, że dopuszczać nierówności można tylko wtedy, gdy są na korzyść najbardziej pokrzywdzonych. W tym sensie sztuczna inteligencja nie może pogłębiać nierówności. Podsumowując: związek podstaw etycznych dla zastosowań sztucznej inteligencji z teorię sprawiedliwości Rawlsa, wydaje się, że podstawowe zasady etyczne jawią się jako niezmienne, niezależnie od okoliczności. Należy bowiem pamiętać, że przywołaną teorię ogłosił Rawls już w 1971 roku.

W dotychczas prowadzonej refleksji porównano akt sztucznej inteligencji z aktem działania świadomego człowieka. Należy jednak zauważyć, przywołując chociażby koncepcję Paula Ricoeura, że w działaniu ludzkim istnieją działania dobrowolne i mimowolne, świadome i nieświadome[40]. Mimowolności nie da się wyraźnie oddzielić od dobrowolności, szczególnie w złożonym akcie ludzkim. Niemniej ciekawym związkiem myśli Ricoeura i refleksji nad sztuczną inteligencją jest kwestia tożsamości narracyjnej. Otóż dla Ricoeura była ona czymś podstawowym. Człowiek dzięki narracji, w poszczególnych jej etapach, odkrywał siebie coraz głębiej. Narracja leżała jednak po stronie podmiotowości osoby. Jeśli zaś dopuści się dane zewnętrzne, zwłaszcza takie, które w sposób jakby mimowolny będą miały wpływ na człowieka, wtedy zaczynamy dotykać podstawowej kwestii ludzkiej tożsamości jaką jest niepowtarzalność. W ten właśnie sposób – jakby mimowolnie – sztuczna inteligencja zaczyna coraz bardziej wiązać się z transhumanizmem[41] i przejściem od tego co specyficznie ludzkie w stronę postbiologii.

Ważne w podejściu do aktu sztucznej inteligencji jest uświadomienie sobie, że jednym z większych niebezpieczeństw „wymknięcia się" odpowiedzialności poza świadome spektrum człowieka jest dopuszczenie do dwóch typów błędów: „Omission Error" – zwanego błędem przeoczenia, gdy człowiek nie zauważa błędu oprogramowania oraz „Commission Error" – gdy człowiek nie potrafi zidentyfikować komunikatu o błędzie wynikającym

[37] S. Westerstrand, *Reconstructing AI Ethics Principles: Rawlsian Ethics of Artifcial Intelligence*, 5.08.2024, SpringerLink https://link.springer.com/article/10.1007/s11948-024-00507-y (dostęp 7.10.2024), s. 6-8 (cytuję za pobranym plikiem pdf).

[38] Tamże, s. 10-11.

[39] Tamże, s. 12-13.

[40] Warto przywołać chociażby publikacje z pierwszego okresu twórczości, poświęconego eidetyce woli czy drugiego, w którym przełomowe badania zostały zawarte w serii „Czas i opowiadanie", gdzie Ricoeur analizuje teorię czasu, działania i historii. Wpisuje się w to także rozwinięta przez francuskiego filozofa tożsamość narracyjna.

[41] M. Kumorek, *Sztuczna inteligencja a tożsamość narracyjna: perspektywa transhumanistyczna*, „Argument: Biannual Philosophical Journal", vol. 13 (2023) nr 1, s. 60-62.

z nieprawidłowo przebiegającego procesu automatyzacji, w którym to procesie system chce zastąpić informacje prawidłowe nieprawidłowymi[42]. Te mimowolne działania rodzą trudności na poziomie etycznym, chociażby w odniesieniu do pojazdów autonomicznych, gdzie system musi wybrać między dwoma grupami ofiar. Inną pobudką działań mimowolnych jest błąd algorytmu wynikający z niepełnych danych. Prowadzi to wtedy do niezamierzonych skutków[43]. Wraca po raz kolejny potrzeba odwołania się do bardziej ogólnych zasad etycznych, chociażby potrzeba wspomnianych już kryteriów działań o podwójnym skutku (drugi można tu uznać bądź jako nieprzewidywalny, bądź jako negatywny, zakładany jako możliwy), gdy jednak samo działanie jest podejmowane ze względu na akt oceniany pozytywnie w sensie etycznym.

W kontekście tak rozumianych ram podmiotowych – dla intelektualnej uczciwości – należy przywołać choć w zarysie inne spojrzenie na człowieka, który w moim przekonaniu stał się właśnie dlatego tak bardzo podatny nie tylko na myślenie innych, ale nawet na rozumienie sztucznej inteligencji jako „każdego innego". Jedną z myślicielek, która w sposób przekonujący opisuje konstytutywne przygotowanie człowieka do zajęcia takiego miejsca w świecie, jest nieżyjąca już od półwiecza Hannah Arendt. Jak zauważa Seyla Benhabib, Arendt analizując koncepcję sądu estetycznego I. Kanta, zaproponowała przeniesienie go na grunt publiczny. Proponuje ideę *sensus communis*, w której upatruje rozszerzenie indywidualnego osądu człowieka o widzenie jakie jest udziałem innych, dzięki temu człowiek może zestawić swoje widzenie świata z postrzeganiem go przez innych, by w ten sposób uniknąć iluzji subiektywności. Można tu dostrzec Kantowską ideę uniwersalizacji, która w tworzeniu własnego sądu każe niejako postawić się na miejscu każdego innego człowieka. Jak twierdzi Arendt, siła takiego sądu polega na porozumieniu i nawet wtedy, gdy człowiek jest sam w swojej refleksji nie zestawia siebie ze sobą i sobą jako innym, ale z innym w wymiarze przynajmniej wyobrażanego porozumienia. Człowiek potrzebuje więc innych, na których miejscu (in whose place) musi nauczyć się myśleć i postrzegać świat ich oczami[44].

Potrzeba konstytutywnych rozwiązań (CAI) i (CCAI)

Skoro człowiek pojmowany w ramach *sensus communis* nie obroni się sam przed sztuczną inteligencją jako swoistym wszechświatowym pojmowaniem rzeczywistości, należałoby go zabezpieczyć pewnymi jasnymi ramami czy kryteriami. Jednym z pierwszych podejść do rozwiązania tego problemu jest propozycja pewnych naczelnych zasad w kluczu tzw. Explainable AI. Najprościej można określić Explainable AI (XAI) jako wyjaśnialną sztuczną inteligencję, która odnosi się do zestawu technik i procesów, które pozwalają

[42] P. Henz, *Ethical and legal responsibility for Artifcial Intelligence*, SpringerLink 22.09.2021, https://link.springer.com/article/10.1007/s44163-021-00002-4 – cytuję pobrany PDF s. 3, punkt 3 artykułu: Behavioral science, (dostęp 7.10.2024).

[43] Tamże, s. 3-4.

[44] S. Benhabib, *Judgment and the Moral Foundations of Politics in Arendt's Thought*, „Political Theory" vol. 16 (1988) nr 1, s. 39-40.

ludziom zrozumieć i ufać decyzjom podejmowanym przez modele sztucznej inteligencji. XAI jest szczególnie istotna w przypadkach, gdy modele te są bardzo złożone. Celem XAI jest zapewnienie przejrzystości działania modeli AI, aby użytkownicy mogli zrozumieć, dlaczego model podjął określoną decyzję. Ma to kluczowe znaczenie w takich dziedzinach, jak medycyna, prawo, finanse, gdzie konsekwencje decyzji AI mogą mieć duży wpływ na życie i zdrowie ludzi. Dzięki XAI możemy: zrozumieć proces decyzyjny modelu; zidentyfikować i poprawić błędy w modelu; zapewnić odpowiedzialność za decyzje AI. Do głównych zasad można zaliczyć: przejrzystość – modele AI muszą być zrozumiałe, a ich działanie przejrzyste; interpretowalność – decyzje AI muszą być możliwe do zinterpretowania przez ludzi; uzasadnienie decyzji – AI powinna być w stanie wyjaśnić, dlaczego podjęła taką a nie inną decyzję, szczególnie w sytuacjach, gdzie jej działania mogą budzić wątpliwości etyczne lub prawne; odpowiedzialność – konieczność zapewnienia, że AI nie działa w sposób szkodliwy, a jej decyzje mogą być przypisane odpowiedzialnym twórcom systemów[45].

Powyższe nakreślone ramy wyjaśnialności i interpretowalności systemów sztucznej inteligencji odnoszą się do powszechnego prawa rozumienia, idei według której każdy człowiek ma prawo zrozumieć, dlaczego system podjął taką a nie inną decyzję. Jest to kluczowy element budowania zaufania do przyszłych działań sztucznej inteligencji, jak również odpowiedzialności za nią. Nie jest celem niniejszego artykułu analizowanie bogatej literatury w tym zakresie, lecz zwrócenie uwagi jak przejrzystość i interpretowalność mogą przyczynić się do podniesienia na wyższy poziom etyczny refleksji nad tym zagadnieniem. Jest też coraz więcej studiów przypadku, które potwierdzają w praktyce badawczej przyjęcie tego typu rozwiązań[46].

Na zakończenie proponowanych refleksji nad „aktem" sztucznej inteligencji pragnę odwołać się do bardzo istotnego – w moim przekonaniu – artykułu poświęconemu obronie sztucznej inteligencji przez nią samą. Badania te zostały przeprowadzone pod kierunkiem Yuntao Bai i zespołu, a opublikowane w grudniu 2022 roku[47]. To pierwsza na taką skalę próba tego, jak sztuczna inteligencja mogłaby nadzorować inną sztuczną

[45] *What is explainable AI?*, https://www.ibm.com/topics/explainable-ai (dostęp 11.07.2024). Szerzej na ten temat V. Božić, *Explainable Artificial Intelligence (XAI): Enhancing Transparency and Trust in AI Systems*, https://www.researchgate.net/publication/374478583_Explainable_Artificial_Intelligence_XAI_Enhancing_Transparency_and_Trust_in_AI_Systems (dostęp 14.06.2024), zob. także *Interpretability vs explainability: Understanding the Differences and Importance in the World of Artificial Intelligence*, https://www.xcally.com/news/interpretability-vs-explainability-understanding-the-importance-in-artificial-intelligence/ (dostęp 16.06.2024).

[46] N. Balasubramaniam, M. Kauppinen, A. Rannisto, K. Hiekkanen, S. Kujala, *Transparency and explainability of AI systems: From ethical guidelines to requirements*, „Information and Software Technology", July 2023, PDF https://www.sciencedirect.com/search?qs=Transparency%20and%20explainability%20of%20AI%20systems%3A%20From%20ethical%20guidelines%20to%20requirements&pub=Information%20and%20Software%20Technology&cid=271539 (dostęp 16.06.2024).

[47] Y. Bai et alii, *Constitutional AI: Harmlessness from AI Feedback*, https://arxiv.org/abs/2212.08073 – tekst artykułu – https://arxiv.org/pdf/2212.08073 (dostęp 18.10.2024).

inteligencję bez zbędnego udziału człowieka (without any human feedback labels for harms). Jest to próba zbudowania „nieszkodliwego asystenta AI", który po odpowiednim wytrenowaniu mógłby nadzorować inne systemy. Model ten autorzy nazywają „Constitutional AI" (CAI). Metoda ta nazywa się konstytucyjną, gdyż w zamyśle może trenować inne systemy wyłącznie poprzez krótkie listy zasad, czyli konstytucję[48].

Pierwszym krokiem jest tzw. skalowanie nadzoru. Zakłada się, że nadzór AI może być bardziej efektywny niż nadzór człowieka. Podejmuje się także próbę łączenia nadzoru człowieka i nadzoru AI w celu polepszenia efektywności nadzoru w stosunku do nadzoru jaki człowiek i maszyna mogłyby wykonywać osobno. Kolejnym motywem było zbudowanie takiego systemu nadzoru AI, który ze świadomością, że już dzisiaj zdolności AI przekraczają ludzkie możliwości kontroli, mógłby właśnie dzięki odpowiednio wytrenowanej AI kontrolować inne systemy (w artykule ukazuje to rysunek 2, który na wykresie pokazuje „wyższość" nadzoru sztucznej inteligencji nad nadzorem człowieka. Novum tego badania było również to, że przyjęto zasadę, iż asystent AI, który na trudne pytania odpowiada „nie wiem" nie jest w niczym pomocny. Zatem spróbowano ograniczyć tzw. nieszkodliwość na rzecz pomocniczości – „A Harmless but Non-Evasive (Still Helpful) Assistant" – (s. 3).

Uczenie asystenta AI odbywało się na dwóch etapach: etap nadzorowany (krytyka – poprawa – uczenie nadzorowane). Polega to ogólnie na tym, że kiedy pada niewystarczająca odpowiedź, prosi się system o poprawienie zgodne z wytycznymi (CAI), dokonuje się kolejnej próby aż do momentu, gdy asystent osiąga pożądany poziom. Drugi etap to etap porównania, ale zamiast tradycyjnego oceniania odpowiedzi przez człowieka, tym razem ocenia AI w oparciu o zasady konstytucyjne. W pierwszych próbach badania nadzoru AI okazało się, że nadzorowanie przez sztuczną inteligencję dorównuje w prostych modelach nadzorowi człowieka (s. 6n). W trzecim punkcie artykułu opisano dokładnie przebieg badania. Przynajmniej w założeniu autorów przyjęto, że metoda kontrolowania przez AI bez nadzoru człowieka zwiększa ilość odpowiedzi nieszkodliwych oraz zdecydowanie poprawia zależność „nieszkodliwość-pomocniczość" (s. 7-9). Wynikałoby z tego, że modele nadzoru pracujące autonomicznie będą docelowo lepiej kontrolowały inne modele sztucznej inteligencji.

Powyższe badania cieszą się na ogół uznaniem wśród znawców problematyki. Zwraca się jednak uwagę na dwie zasadnicze kwestie: czy w pełni zautomatyzowany proces nadzoru jest w stanie wychwycić wszystkie niuanse ludzkich wartości i zasad etycznych oraz druga kwestia – czy w tworzeniu zasad konstytutywnych AI nie powinno być etapu, w którym te zasady tworzy się w oparciu o szerszy kontekst społeczny[49]. Etap ten nazywa się „Collective Constitutional AI" (CCAI) w odróżnieniu od omówionego wcześniej (CAI).

[48] Tamże, s. 1. Dalsze odniesienia do stron będę podawał bezpośrednio w tekście – JS.

[49] Są już pierwsze próby takich rozwiązań – S. Huang, D. Siddarth et alii, *Collective Constitutional AI: Aligning a Language Model with Public Input*, https://ar5iv.labs.arxiv.org/html/2406.07814v1, (dostęp 18.10.2024).

Oczywiście są to jakieś próby „uetycznienia" procesów, które człowiek coraz bardziej oddaje maszynie. Z jednej strony rozwój sztucznej inteligencji jest tak galopujący, że przekracza granice możliwości nadzoru człowieka – można co najwyżej ograniczyć tempo procesu. Można też przyjąć drugą ścieżkę rozwoju i przyjąć, że skoro nie można zatrzymać postępu prac nad rozwojem AI, jak też samego uczenia maszynowego z jego wielorakimi odmianami, to przynajmniej niech nadzoruje je maszyna w oparciu o „ludzkie" ramy etyczne.

W jakim kierunku pójdą dalsze prace nad rozwojem sztucznej inteligencji i która z dróg okaże się właściwa, pokaże przyszłość. Można się tylko zastanawiać, czy warto powstrzymać ludzki potencjał rozwoju, by mogły w sposób coraz mniej kontrolowany rozwijać się systemy sztucznej inteligencji i czy ostatecznie nie zgubi nas bezkrytyczne rozumienie idei postępu i zwykłe ludzkie lenistwo?

Bibliografia

Agenda 2030 na rzecz zrównoważonego rozwoju (Transforming our world: the 2030 Agenda for Sustainable Development), tekst polski cyt. za https://www.gov.pl/web/rozwoj-technologia/agenda-2030 (dostęp 12.07.2024).

Andrzejuk A, *Wolność w doktrynie Tomasza z Akwinu*, w: *Wolność człowieka i jej granice. Antologia pojęcia w doktrynach polityczno-prawnych. Od Starożytności do Monteskiusza,* red. O. Górecki, Wydawnictwo Uniwersytetu Łódzkiego, Łódź 2019, s. 149-170.

Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence by Kate Crawford Yale University Press, New Haven, CT, U.S.A 2021, polskie wydanie K. Crawford, *Atlas sztucznej inteligencji. Władza, pieniądze i środowisko naturalne*, wyd. Bo.wiem, Kraków 2024, szczególnie rozdział 3: *Dane*, s. 97-128.

Bai Y. et alii, Constitutional AI: *Harmlessness from AI Feedback*, s. 1-34, https://arxiv.org/pdf/2212.08073 (dostęp 18.10.2024)

Balasubramaniam N., Kauppinen M., Rannisto A., Hiekkanen K., Kujala S., *Transparency and explainability of AI systems: From ethical guidelines to requirements*, „Information and Software Technology", July 2023, PDF – https://www.sciencedirect.com/search?qs=Transparency%20and%20explainability%20of%20AI%20systems%3A%20From%20ethical%20guidelines%20to%20requirements&pub=Information%20and%20Software%20Technology&cid=271539 (dostęp 16.06.2024).

Benhabib S., *Judgment and the Moral Foundations of Politics in Arendt's Thought*, „Political Theory" vol. 16 (1988) nr 1, s. 29-51.

Božić V., *Explainable Artificial Intelligence (XAI): Enhancing Transparency and Trust in AI Systems*, https://www.researchgate.net/publication/374478583_Explainable_Artificial_Intelligence_XAI_Enhancing_Transparency_and_Trust_in_AI_Systems (dostęp 14.06.2024) – artykuł PDF (preprint), s. 1-21.

City of Amsterdam Algorithm Register, https://algoritmeregister.amsterdam.nl/en/ai-register/ (dostęp 2.10.2024).

Coeckelbergh M., *Artificial intelligence, the common good, and the democratic deficit in AI governance*, Springer, Published online 22.05.2024, s. 4-6, link do artykułu https://link.springer.com/article/10.1007/s43681-024-00492-9 (dostęp 12.08.2024).

Constantinescu M., Voinea C., Uszkai R., Vică C., *Understanding responsibility in Responsible AI. Dianoetic virtues and the hard problem of context*, „Ethics and Information Technology" (2021), s. 803-814, https://link.springer.com/article/10.1007/s10676-021-09616-9, (dostęp 15.09.2024).

Davies B., *Happiness,* w: *The Oxford Handbook of Aquinas*, rozdział 17, ed. B. Davies i E. Stump, Oxford University Press, 2012, s. 227-237.

Fourtané S., *The Three Types of Artificial Intelligence: Understanding AI*, https://ir.westcliff.edu/wp-content/uploads/2020/01/The-Three-Types-of-Artificial-Intelligence-Understanding-AI.pdf, (dostęp. 11.07.2024).

Franke Ch.A, Marques de Carvalho J., *Das Wesen der menschlichen Handlung bei Thomas von Aquin (The Essensce of Human Action in Thomas Aquinas)*, „Revista Portuguesa de Filozofia", 2023, vol. 79 (1-2), s. 479-506.

Ghallab M., *Responsible AI: requirements and challenges*, Springer Open, 3.09.2019 https://aiperspectives.springeropen.com/articles/10.1186/s42467-019-0003-z (dostęp 17.09.2024).

Henz P., *Ethical and legal responsibility for Artifcial Intelligence*, SpringerLink 22.09.2021, https://link.springer.com/article/10.1007/s44163-021-00002-4 (dostęp 7.10.2024), artykuł w PDF, s. 1-5.

Huang S, Siddarth D. et alii *Collective Constitutional AI: Aligning a Language Model with Public Input*, https://ar5iv.labs.arxiv.org/html/2406.07814v1, (dostęp 18.10.2024).

Interpretability vs explainability: Understanding the Differences and Importance in the World of Artificial Intelligence, https://www.xcally.com/news/interpretability-vs-explainability-understanding-the-importance-in-artificial-intelligence/ (dostęp 16.06.2024).

Jasiński K., *Czyn doskonalący osobę. W kręgu perfekcjonizmu Karola Wojtyły*, „Studia Elbląskie" XIII(2012), s. 351-361.

Karaś A., *Struktura aktu moralnego w ujęciu Mieczysława Gogacza*, „Studia Philosophiae Christianae" 47(2011) nr 2, s. 157-173.

Kumorek, *Sztuczna inteligencja a tożsamość narracyjna: perspektywa transhumanistyczna*, „Argument: Biannual Philosophical Journal", vol. 13 (2023) nr 1, s. 59-74.

Łukawski A., *Generatywna sztuczna inteligencja (GenAI) a kreatywność*, Zintegrowana Platforma Edukacyjna Ministerstwa Edukacji Narodowej, https://zpe.gov.pl/a/i-o-kreatywnosci-generatywnej-sztucznej-inteligencji-11-generatywna-sztuczna-inteligencja-genai-a-kreatywnosc/D16yodZxY, (dostęp 13.09.2024).

Manning Ch., *Artificial Intelligence Definitions*, Stanfort University, Human-Centered Artificial Intelligence, September 2020, https://hai.stanford.edu/sites/default/files/2020-09/AI-Definitions-HAI.pdf, (dostęp 12.04.2021).

Nęcka E., *Wolna wola czy wolne weto: rola świadomości w czynnościach wolicjonalnych*, s. 11-31, https://www.publikacje.pan.pl/Content/117527/PDF/Necka.pdf, (dostęp 11.09.2024).

Nęcka E., Prusak J., *Eksperymentalna psychologia woli : wolność i intencjonalność z perspektywy psychologii poznawczej i neuronauki*, 2016, s. 41-60.

https://www.academia.edu/66683464/Eksperymentalna_psychologia_woli_wolno%C5%9B%C4%87_i_intencjonalno%C5%9B%C4%87_z_perspektywy_psychologii_poznawczej_i_neuronauki, (dostęp 11.09.2024).

Osborne T.M. Jr, *Human Action in Thomas Aquinas, John Duns Scotus & William of Ockham*, Washington DC: The Catholic University of America Press 2014.

Parlament Europejski: *Sztuczna inteligencja: szanse i zagrożenia*, 20.06.2023, s. 1-6, https://www.europarl.europa.eu/pdfs/news/expert/2020/9/story/20200918STO87404/20200918STO87404_pl.pdf (dostęp 12.09.2024).

Pastuszka J., *Filozoficzne i empiryczne pojęcie osoby ludzkiej*, „Roczniki Filozoficzne" vol. 11 (1963) nr 4, s. 45-60.

Perska L., *Zatarte granice między sztuczną inteligencją a ludzką kreatywnością*, Elblog, 16.06.2024, https://elblog.pl/pl/2024/06/16/zatarte-granice-miedzy-sztuczna-inteligencja-a-ludzka-kreatywnoscia/ (dostęp 13.09.2024).

Pilsner, J. *Circumstance*, w: *The Specification of Human Actions in St Thomas Aquinas* (ed. J. Pilsner), Oxford 2006, s. 172-298.

Polskie stanowisko w sprawie AI Act, cyt. za https://www.gov.pl/web/ai/rewolucja-w-regulacji-wchodzi-w-zycie-akt-o-ai (dostęp 14.09.2024).

Popa D.M., *Frontrunner model for responsible AI governance in the public sector: the Dutch perspective*, 27.09.2024, https://link.springer.com/article/10.1007/s43681-024-00596-2 (dostęp 1.10.2024).

Rozporządzenie Parlamentu Europejskiego i Rady (UE) 2024/1689 z dnia 13 czerwca 2024 r. w sprawie ustanowienia zharmonizowanych przepisów dotyczących sztucznej inteligencji oraz zmiany rozporządzeń (WE) nr 300/2008, (UE) nr 167/2013, (UE) nr 168/2013, (UE) 2018/858, (UE) 2018/1139 i (UE) 2019/2144

oraz dyrektyw 2014/90/UE, (UE) 2016/797 i (UE) 2020/1828 (akt w sprawie sztucznej inteligencji) (Tekst mający znaczenie dla EOG), European Union, EUR-Lex, https://eur-lex.europa.eu/legal-content/PL/TXT/?uri=CELEX:32024R1689, (dostęp 14.06.2024).

Sheikh H., Prins C., Schrijvers E., *Mission AI. The New System Technology*, Springer (Open Acces) 2023, część I, rozdział 2: *Artificial Inelligence: Definition and Background*, s. 15-41, https://link.springer.com/chapter/10.1007/978-3-031-21448-6_2 (dostęp 12.09.2024).

Sobkowiak J.A., *Trust as a Condition of Getting to Know the Truth: The Anthropological Aspect* – „Studia Theologica Varsaviensia", 2016 nr 2, s. 195-208.

Stępień A., Dęmska E., *Postępowania przyszłości. Holenderski algorytm ostrzeżeniem dla Europy*, Crido: blog podatkowy, 1.02.2023 https://crido.pl/blog-taxes/postepowania-przyszlosci-holenderski-algorytm-ostrzezeniem-dla-europy/ (dostęp 2.10.2024).

Tomanek M., *Zagrożenia sztucznej inteligencji. Model AI modyfikował swój kod*, Holistic News, 14.10.2024, https://holistic.news/model-ai-sam-chcial-zmienic-swoj-kod-naukowcy-sa-zaskoczeni/ (dostęp 14.10.2024).

Townsent R., Kostro P., *Etyczne podejście do danych i AI ma biznesowy sens*, MITSloan, Management Review Polska, 8.09.2023, https://mitsmr.pl/b/etyczne-podejscie-do-danych-i-ai-ma-biznesowy-sens/PCr-83KVmR (dostęp 4.10.2024).

Westerstrand, S. *Reconstructing AI Ethics Principles: Rawlsian Ethics of Artifcial Intelligence*, 5.08.2024, SpringerLink https://link.springer.com/article/10.1007/s11948-024-00507-y (dostęp 7.10.2024).

What is explainable AI?, https://www.ibm.com/topics/explainable-ai (dostęp 11.07.2024).

Williams Th., *Human Freedom and Agency,* w: *The Oxford Handbook of Aquinas*, rozdział 15, ed. B. Davies i E. Stump, Oxford University Press, 2012, s. 199-208.

Zadykowicz T., Hasło: *Czyn ludzki*, w: *Jan Paweł II. Encyklopedia nauczania moralnego*, red. J. Nagórny, K. Jeżyna, Polskie Wydawnictwo Encyklopedyczne, Radom 2005, s. 129-133.

Biogram

Jarosław Andrzej Sobkowiak – doktor nauk teologicznych w zakresie teologii moralnej, studia specjalistyczne z etyki – Centre Sévres, Paryż, adiunkt i kierownik Zakładu Dydaktyki Mediów w Instytucie Edukacji Medialnej i Dziennikarstwa, Dyrektor Centrum Komunikacji Społecznej i Cyfrowej UKSW, prodziekan ds. studenckich i kształcenia ustawicznego, redaktor naczelny półrocznika „Studia Theologica Varsaviensia", członek zarządu (sekretarz) Stowarzyszenia Teologów Moralistów, członek Polskiego Towarzystwa Komunikacji Społecznej.