Czy sztuczna inteligencja może być odpowiedzialna albo godna zaufania?

dr Natalia Juchniewicz
ORCID: 0000-0002-2686-9404
Uniwersytet Warszawski

W dyskursie wokół sztucznej inteligencji (*artificial intelligence* – AI), gdy dyskutowane są warunki wdrażania tej technologii do życia społecznego, ekonomii czy polityki, pojawia się problem zaufania ludzi do samej technologii oraz zakresu odpowiedzialności, jaki możemy jej przypisywać bądź na nią delegować. W szerszym sensie wiąże się to bowiem z problemami ontologicznymi (jakim rodzajem bytu jest sztuczna inteligencja?; czy sztuczna inteligencja to aktor społeczny czy po prostu narzędzie? itp.) oraz wynikającymi z nich problemami moralnymi (czy sztuczna inteligencja jest autonomiczna?; czy sztuczna inteligencja podlega prawu? itp.; kto kontrolować powinien działanie AI?). Sztuczna inteligencja, jak żadna wcześniejsza technologia, potrafi bowiem doskonale imitować człowieka: wygrywa z nim w szachy, może prowadzić długie konwersacje o sztuce, może wspierać emocjonalnie osoby samotne, a przy tym zanalizuje duże zbiory danych, sprofiluje użytkowników określonej aplikacji czy też będzie rekomendować określone osoby do uzyskania kredytu. Spektrum zastosowań AI jest niemal nieograniczone. Massimo Airoldi przekonuje, że sztuczną inteligencję definiować można jako technologię opartą na algorytmie wyznaczonym przez: po pierwsze, procedurę matematyczną (co dotyczy zresztą każdej formy maszyny), pod drugie, model podążania za określonymi komendami opartymi na dedukcji (*Good Old Fashioned Artificial Intelligence*; np. algorytmy wyszukiwarek internetowych); po trzecie, uczenie maszynowe oparte na autonomicznym przetwarzaniu danych i indukcyjnym uczeniu się nadzorowane albo nienadzorowane przez ludzi (np. Page Rank, Alpha Go). Airoldi pokazuje, jak wraz z rozwojem technologicznym zmieniał się sposób definiowania sztucznej inteligencji. Była ona czymś innym w erze analogowej, erze cyfrowej i wreszcie w obecnej erze platform (Airoldi 2022: 14). Dla celów mojego artykułu przyjmuję

w szczególności trzecią definicję sztucznej inteligencji, gdyż to wokół niej narastają problemy ontologiczne, epistemologiczne i moralne. Uczenie maszynowe, sieci neuronowe, pewien zakres niejasności w procesie uczenia się AI (Burrell: 2016), jej zdolności analizy wielkich zbiorów danych, obecność algorytmów w codziennych praktykach, to wszystko sprawia, że nawet jeśli nie chcemy mieć ze sztuczną inteligencją świadomie do czynienia, to i tak często jest ona gruntem dla wielu naszych aktywności i towarzyszy nam oraz asystuje w wielu czynnościach. W ramach problemów podejmowanych w niniejszym artykule można uznać za sztuczną inteligencję wszystkie formy technologii uczących się opartych na algorytmie od algorytmów typu Page Rank po roboty społeczne. W zależności od kontekstu używać będę zamiennie słów „sztuczna inteligencja", „algorytm", „AI" czy „robot". Sama forma tej technologii jest bowiem wtórna wobec stawianych w artykule problemów wokół „moralnej AI".

Jako maszynę możemy chcieć zaprojektować sztuczną inteligencję, by służyła określonym, instrumentalnym celom. To jednakże, co AI „robi" jako aktor czy aktant społeczny (Latour 2013), wymyka się liniowemu projektowaniu jej oddziaływania. Jednym z nieoczywistych problemów związanych ze sztuczną inteligencją, choć bezpośrednio powiązanym z delegowaniem na nią działania oraz zaufaniem do niej, jest jej wymiar moralny. Samo powiązanie technologii z moralnością wymagało od filozofii i etyki przesunięcia w kierunku rzeczy, skupienia uwagi na ontologii przedmiotów i ich możliwym sprawstwie (*agency*). Skutkiem owego przesunięcia wyłoniły się dwie dominujące w literaturze filozoficznej narracje wokół moralnej sztucznej inteligencji. Pierwsza skupia się na konsekwencjach moralnych wdrażania sztucznej inteligencji w różne obszary życia społecznego, ekonomicznego czy politycznego, którą nazywam w skrócie narracją konsekwencjalistyczną. Druga narracja dąży do przedstawienia sztucznej inteligencji jako nowej formy aktora czy agenta społecznego, a nawet jako nowy rodzaj sztucznej osoby, co w skrócie nazywam narracją personalistyczną. W niniejszym artykule postaram się scharakteryzować i poddać krytyce obie te narracje w celu ukazania, że choć dotyczą one sztucznej inteligencji, w rzeczywistości postulują odpowiedzialność ludzi za tę technologię i ewentualne zaufanie bądź jego brak względem celów, jakie ludzie we wdrażaniu AI stawiają. Sztuczna inteligencja godna zaufania nie jest zatem dosłownie postulatem adresowanym wobec maszyny, a raczej pytaniem o to, w świecie jakich wartości chcemy funkcjonować jako ludzie. Fakt, że w badaniach nad moralną sztuczną inteligencją wcale nie chodzi o postulowanie moralności samych rzeczy, a o refleksję nad wartościami i stosunkiem ludzi do bytów nie-ludzkich, sprawia, że etyka sztucznej inteligencji okazuje się przede wszystkim dziedziną badań nad relacjami, jakie budować może człowiek z Innym, nie tylko technologicznej natury.

Nazywam wyszczególnione przeze mnie podejścia do moralnej sztucznej inteligencji „narracjami", po pierwsze dlatego, że niniejszy artykuł nie opiera się na pogłębionej analizie dyskursu, a na propozycji uporządkowania pewnych argumentów filozoficznych w określone grupy na zasadzie przyjęcia określonej ontologii AI. Celem mojego artykułu jest zatem wskazanie pewnych tendencji w budowaniu argumentów wokół AI, a nie ich klasyfikacja czy szczegółowa analiza. Chodzi mi o pokazanie, że odpowiedź na pytanie, czy możemy zaufać sztucznej inteligencji, sprowadzić można do określenia, czy stoimy po jednej czy po drugiej stronie dyskusji na temat ontologii AI – czy jest ona dla nas przedmiotem czy formą podmiotu. Moim zdaniem propozycja wyszczególnienia takich narracji ma charakter porządkujący dla szczegółowej pracy nad określonymi argumentami. Po drugie, pojęcie narracji jest bardzo szerokie, pozwala uchwycić zarówno konkretne, naukowe argumenty, jak i opowieści oraz wyobrażenia, jakie wokół sztucznej inteligencji pojawiają się w różnych obszarach naszej z nią interakcji. Nie jest zatem tak, że wyszczególnione przeze mnie narracje są zamknięte albo jednoznacznie zdefiniowane. W ich obrębie istnieją wewnętrzne spory co do tego, czym AI jest i jak należy się do niej odnosić. Po trzecie, narracje pozwalają przecinać podejścia w ramach samej etyki, która w niektórych ujęciach ma charakter opisowy, a w innych normatywny. Problem w tym, że nie zawsze jest jasne, czy np. z opisu określonych praktyk zdaniem autorów wynikają jakieś konsekwencje moralne albo czy postulowana cnota jest adresowana tylko do ludzi czy może też maszyn. Narracje, w moim przekonaniu, są bardziej elastycznym pojęciem i pozwalają opisać pewien zbiór problemów, w których mieści się także sama moralna AI.

W pierwszej części artykułu dokonam charakterystyki narracji konsekwencjalistycznej. W jej ramach uwaga skupia się na pozytywnych bądź negatywnych konsekwencjach produkowania, wdrażania oraz używania sztucznej inteligencji. Celem badaczy zorientowanych na konsekwencje AI jest postulowanie określonych wartości, które technologia ta powinna spełniać, aby np. nie szkodzić (przede wszystkim ludziom, ale i przyrodzie). Co istotne, pochodzenie i rozumienie tych wartości nie jest zawsze w tej narracji wyjaśniane. Sztuczna inteligencja rozumiana jest tutaj jako instrument czy narzędzie, na które co najwyżej delegować można określone wartości (Verbeek 2011), zatem w narracji tej w dużo większym stopniu chodzi o moralność i cnotę ludzi niż o moralną AI.

Narracja personalistyczna, którą charakteryzuję w drugiej części artykułu, zakłada z kolei, że podmiotowość, działanie, sprawstwo oraz osobowość mogą zostać rozszerzone na innych aktorów niż ludzie. Narracja ta ujawnia tym samym, że traktowanie sztucznej inteligencji „jak człowieka" chociażby ze względu na jej umiejętność posługiwania się językiem, nie jest błędem kategorialnym opartym

na niezdolności rozpoznania, iż mamy do czynienia z maszyną, a właśnie na przyjęciu, iż nawet jeśli mamy do czynienia z maszyną, to interakcja z nią może być równa jakościowo interakcjom z ludźmi. W ramach tej narracji wyłaniają się dwa dominujące podejścia: tzw. etyka zorientowana na aktora, która zakłada, że AI może być podmiotem moralnym w sensie sprawczości moralnej (*agent-oriented ethics*; Anderson & Anderson 2011; Wallach & Allen 2009) oraz etyka zorientowana na pacjenta czy odbiorcę (*patient-oriented ethics*; Gunkel 2018), która określa, jaki stosunek do AI powinniśmy mieć jako ludzie, bez względu na to, czy ona sama jest podmiotem moralnym czy nie. Problem, jaki stoi przed narracją personalistyczną, to określenie sensu działania moralnego, do jakiego zdolna miałaby być sztuczna inteligencja.

W trzeciej części pracy poddam obie te narracje krytyce. Gdy narracja konsekwencjalistyczna mówi o sztucznej inteligencji godnej zaufania, to nie mówi wcale o sztucznej inteligencji, a o jej pożądanych, społecznych skutkach, którym możemy zaufać. Szyld *trustworthy AI* nie dotyczy zatem samej sztucznej inteligencji jako rzekomego podmiotu moralnego. Z kolei narracja personalistyczna, choć umiejętnie uzasadnia, dlaczego sztuczna inteligencja może być postrzegana jako np. osoba w sensie prawnym i dlaczego niektóre praktyki ludzi wokół AI mogą być uznane za rozszerzenie definicji podmiotowości moralnej, to nie jest w stanie rozwiązać problemu nieposiadania przez sztuczną inteligencję samej etyki jako koncepcji. Innymi słowy, AI nie jest w stanie stworzyć etyki, co gorsza, nie jest w stanie zmienić się samoistnie w jakimś moralnym kierunku, gdyż oba te warunki podmiotu moralnego zakładają z jednej strony teorię umysłu, z drugiej posiadanie woli i zdolności do projektowania siebie. Propozycją wyjścia poza problemy etyki zorientowanej ontologicznie, a więc takiej, która przypisuje jakość moralną bytom posiadającym określone jakości, jest propozycja postawienia etyki przed ontologią (Gunkel 2018).

W podsumowaniu zbieram kluczowe argumenty artykułu oraz jego konkluzje.

I. Narracja konsekwencjalistyczna

Przez narrację konsekwencjalistyczną rozumiem łączenie sztucznej inteligencji z badaniami nad konsekwencjami społecznymi, ekonomicznymi, politycznymi, czy szerzej etycznymi, tworzenia, wdrażania i posługiwania się AI. W języku charakterystycznym dla tej narracji pojawiają się często „implikacje", pożądane i niepożądane „skutki", zamierzone i niezamierzone „efekty", kalkulacja „zysków i strat". Posługiwanie się sztuczną inteligencją służy w tej narracji odpowiedzi na następujące pytania: Kim możemy się stać? – jakie mamy szanse autonomicznej samorealizacji dzięki technologii AI; Co możemy zrobić? – w jakim stopniu sztuczna inte-

ligencja poszerzy bądź zawęży ludzką sprawczość; Co możemy osiągnąć? – jakie technologia AI daje możliwości jednostkowe i społeczne; Jak możemy wchodzić w interakcję ze sobą nawzajem i ze światem? – czy sztuczna inteligencja sprzyja spójności społecznej (Floridi et al., 2018: 690).

Narracja konsekwencjalistyczna najbliższa jest definiowaniu tzw. sztucznej inteligencji godnej zaufania (*trustworthy AI*). Taka definicja zaproponowana została m.in. w dokumencie Komisji Europejskiej *Wytyczne w zakresie etyki dotyczące godnej zaufania sztucznej inteligencji* (KE 2019). Zasady, które spełniać musi AI godna zaufania, są następujące: (i) poszanowanie autonomii człowieka; (ii) zapobieganie szkodom; (iii) sprawiedliwość; (iv) możliwość wyjaśnienia (*explainability*). Sztuczna inteligencja, jakiej możemy zaufać, to w tym ujęciu przede wszystkim technologia, która nie działa na szkodę człowieka i środowiska w całym cyklu swojego życia i działania. Co istotne, w dokumencie tym zakładana jest przewodnia i nadzorcza rola człowieka (KE 2019: 17–19), innymi słowy, odpowiedzialność technologii jest ograniczona na rzecz człowieka i kontroli, jaką sprawuje on nad sztuczną inteligencją.

Luciano Floridi wraz z 12 ekspertami od sztucznej inteligencji dokonał szerszej analizy zasad, jakie postulowane są w aktach prawnych i rekomendacjach dotyczących sztucznej inteligencji i zaproponował całościowe podsumowanie zysków i strat ze sztucznej inteligencji dla Dobrego Społeczeństwa AI (*Good AI Society*). Zanim przejdę do analizy tego podsumowania, chciałabym zwrócić uwagę na język, jakim Floridi rozpoczyna swój artykuł:

> W każdym przypadku sztuczna inteligencja może być użyta (*used*) do wspierania ludzkiej natury i jej potencjału, tworząc w ten sposób możliwości; niedostatecznie wykorzystywana (*underused*), tworząc w ten sposób możliwe koszty; lub nadużywana (*overused*) lub niewłaściwie wykorzystywana (*misused*), tworząc w ten sposób ryzyko. Jak wskazuje terminologia, zakłada się, że wykorzystanie sztucznej inteligencji jest synonimem dobrych innowacji i pozytywnych zastosowań tej technologii. Jednak strach, niewiedza, nieuzasadnione obawy lub nadmierna reakcja mogą doprowadzić społeczeństwo do niedostatecznego wykorzystania technologii AI poniżej ich pełnego potencjału, z powodów, które można ogólnie określić jako niewłaściwe (Floridi et al. 2018: 690–691).

Floridi zdaje sobie sprawę z tego, że kluczowy dla korzyści i strat związanych ze sztuczną inteligencją jest sposób jej użycia. Jeśli używana będzie w celach

sprzyjających społecznemu rozwojowi, to oczywiście stwarzać będzie możliwości pozytywne, jeśli natomiast zostanie źle użyta albo w ogóle zignorowana, przyczynić się może do zmniejszenia szans rozwojowych społeczeństwa. Floridi tym samym sięga do argumentu znanego już od czasów Karola Marksa, który w maszynie fabrycznej widział sprzymierzeńca robotnika, o ile zostanie ona wyzwolona z kapitalistycznego celu produkcji, jakim jest zysk. Problem w tym, że podobnie jak Marksowski postulat właściwego zastosowania technologii dla korzyści ludzkości rozbił się o brak politycznie trwałych sukcesów rewolucji proletariackiej tam, gdzie rzeczywiście z kapitalizmem mieliśmy do czynienia, tak i argumentacja Floridiego zdaje sprawę z tego, jak chcielibyśmy, aby sztuczna inteligencja była używana, ale nie oznacza to w żadnym sensie, że z takim jej użyciem mamy faktycznie do czynienia. Ponadto ów instrumentalistyczny stosunek do sztucznej inteligencji, jako *de facto* środka do realizacji np. społeczeństwa bardziej sprawiedliwego, zdradza, że sama „moralna sztuczna inteligencja" nie jest tu celem. Raczej moglibyśmy powiedzieć, że sztuczna inteligencja jest traktowana jako instrument neutralny moralnie, skoro może być zastosowana do lepszych i gorszych społecznie celów, bądź jako instrument, który można „nauczyć" moralności rozumianej jako pożądane społecznie sposoby działania. To drugie podejście wiąże się z delegowaniem moralności na artefakty, sterowaniem ludzkim zachowaniem i współkształtowaniem przez artefakty decyzji moralnych, które podejmujemy. W szczególności zwrócił na to uwagę Peter-Paul Verbeek, gdy poddał analizie moralny wymiar badań ultrasonograficznych, jakim poddawane są kobiety w ciąży (Verbeek 2008). Verbeek zauważył, że np. decyzja o aborcji czy o leczeniu płodu podjęta może być tylko dlatego, że technologia ujawniła nam jakiś problem, który należy moralnie rozstrzygnąć. Verbeek zwraca uwagę, że technologie „pomagają kształtować naszą egzystencję i podejmowane przez nas decyzje moralne, co niezaprzeczalnie nadaje im wymiar moralny" (Verbeek, 2011: 2). Z fenomenologicznej perspektywy technologie współkształtują nasze codzienne środowisko, nasze doświadczenia, nie mogą zatem nie być traktowane jako współobecne w decyzjach moralnych. Z argumentem Verbeeka trudno się nie zgodzić, choć dyskusyjny jest zakres owego udziału technologii w decyzjach moralnych. Technologie mogą, ale nie muszą, być szczególnie istotnym czynnikiem w rozstrzyganiu dylematów moralnych, a nawet, z punktu widzenia choćby założenia o autonomii podmiotu moralnego, być może sama technologia nawet jeśli prowokuje problem, nie stanowi części w procesie znajdowania dla niego rozwiązania. Założenie, że technologie współkształtują nasze decyzje moralne, skutkuje bowiem przesunięciem dyskusji o moralności z podmiotu na przedmioty, czyli do „moralizowania technologii" (*moralizing technology*), co postuluje sam Verbeek. Dyskurs ten prowadzi zatem do większego zwracania uwagi na to, jak

i dlaczego pewne rozwiązania powinny zostać zaprojektowane, aby np. maksymalizować bezpieczeństwo użytkowników czy pożądane społeczne korzyści (Verbeek & Slob 2006). Problem w tym, że jeśli owa delegacja moralności na rzeczy odbywa się w fazie projektowania, to faktyczna odpowiedzialność za to, czy artefakt działa tak, jak powinien, spada na projektantów. Nawet jeśli pierwsze smartfony z ekranem dotykowym wykluczały z pola użytkowników osoby niewidome, to nazwanie tych technologii „dyskryminującymi" nie dotyczy tak naprawdę samego urządzenia, a społecznych skutków ich ograniczonego projektu. Możliwość uzupełnienia takich artefaktów o np. asystenta głosowego poprawiło ten projekt, ale odpowiedzialność za to była w rękach producentów (zob. Goggin & Newell 2003). Dlatego, gdy Langdon Winner analizował projekty wiaduktów Roberta Mosesa prowadzących do plaży na Long Island w Nowym Jorku i stwierdził, że „artefakty są polityczne", to starał się wytłumaczyć, że artefakty technologiczne zawsze uwikłane są w jakiś porządek władzy, a z całą pewnością, określony porządek infrastrukturalny wywołują:

> Rzeczy, które nazywamy „technologiami", są sposobami budowania porządku w naszym świecie. Wiele urządzeń technicznych i systemów ważnych w codziennym życiu zawiera możliwości dla wielu różnych sposobów porządkowania ludzkiej aktywności. Świadomie lub nie, celowo lub nieumyślnie, społeczeństwa wybierają struktury dla technologii, które wpływają na to, jak ludzie będą pracować, komunikować się, podróżować, konsumować i tak dalej przez bardzo długi czas. W procesach, w których podejmowane są decyzje strukturalne, różni ludzie są różnie usytuowani i posiadają nierówne stopnie władzy, a także nierówne poziomy świadomości (Winner 1986: 28).

Dlatego właśnie narracja konsekwencjalistyczna wokół AI posiada z jednej strony dobre filozoficzne podstawy, gdyż sięga do studiów nad nauką i technologią, delegacji moralności na artefakty, fenomenologii czy częściowo teorii aktora-sieci, ale traktuje technologie jedynie jako element „infrastruktury moralnej", a nie faktyczny podmiot moralny. Dlatego sztuczna inteligencja, jakiej możemy zaufać, to w tej narracji *de facto* zaufanie do podmiotu dostarczającego nam technologię sztucznej inteligencji i przyjęcie, że jest godny zaufania co do celów i sposobów dostarczania takiej technologii użytkownikom oraz spełniania społecznych, ekonomicznych i politycznych oczekiwań co do realizacji określonych wartości.

W zasadach, jakie powinny być kluczowe w tworzeniu i wdrażaniu sztucznej inteligencji, za analizą dokumentów i deklaracji przeprowadzoną przez Floridiego, dominuje pięć kluczowych zasad: dobrobyt (*beneficience*), niekrzywdzenie (*non-maleficience*), autonomia, sprawiedliwość oraz możliwość wyjaśnienia (*explainability*).

Dobrobyt sprowadza się do troski o ludzką godność oraz o planetę. Sztuczna inteligencja ma zasadniczo służyć postępowi ludzkości, dobru wspólnemu oraz dobremu życiu. *Niekrzywdzenie* oznacza troskę o zachowanie prywatności, bezpieczeństwa oraz ocenę ryzyka kierunków rozwoju sztucznej inteligencji. Istotą tego postulatu jest zatem minimalizacja krzywd, jakie stosowanie AI może wywołać. *Autonomia* wiąże się z delegacją działania na technologię. Wraz z możliwością scedowania podejmowania wielu decyzji na sztuczną inteligencję człowiek powinien mieć wciąż prawo do kontroli tego, co AI robi. Ponadto podkreślona zostaje w tym postulacie „wartość ludzkiego wyboru (*value of human choice*) – przynajmniej co do istotnych decyzji" (Floridi et al. 2018: 698), a więc zachowanie ostrożności w delegowaniu zbyt wielu decyzji na sztuczną inteligencję i dowartościowanie decyzji podejmowanych przez człowieka. *Sprawiedliwość* sprowadza się do promocji wspólnych wartości i dobrobytu, ograniczania powielania i podtrzymywania przez sztuczną inteligencję stereotypów i uprzedzeń oraz dążenie do tego, aby ta technologia odnosiła się z szacunkiem do wszystkich ludzi bez względu na rasę, pochodzenie czy płeć. *Możliwość wyjaśnienia* oznacza natomiast, że sposób działania sztucznej inteligencji powinien być zrozumiały dla ludzi, co ma minimalizować ryzyko nieprzewidzianych konsekwencji, których człowiek nie byłby w stanie kontrolować.

Powyższe zasady – po pierwsze – sugerują, że właściwym beneficjentem działania sztucznej inteligencji jest człowiek. Innymi słowy, AI nie działa we własnym imieniu czy „dla siebie". Po drugie, zasady te mają co prawda służyć za drogowskaz w tworzeniu i wdrażaniu sztucznej inteligencji w różnych obszarach życia, ale ich faktyczna zbieżność ze sztuczną inteligencją weryfikowana jest najczęściej w procesie jej działania. Moralna AI jest zatem „moralna" dlatego, że generuje problemy wymagające rozwiązania, a skuteczność tych rozwiązań mierzona jest skutkami, jakie z nich wynikają. Gdy zatem w narracji konsekwencjalistycznej pojawia się postulat sztucznej inteligencji godnej zaufania, to wydaje się, że owo zaufanie nie dotyczy jej samej, a tego, jakie wartości zostały w nią zaimplementowane i jakie efekty to wywołuje. Powiązanie badań nad sztuczną inteligencją i etyką ma tu zatem charakter bardzo techniczny i często pozbawiony wyjaśnienia, czy pożądane „efekty" i „zyski" z wdrażania sztucznej inteligencji mierzone są ze względu na korzyści, przyjemności, szczęście, czy określają je sami badacze, opinia publiczna czy politycy.

II. Narracja personalistyczna

W narracji personalistycznej sztuczna inteligencja zostaje potraktowana jako aktor społeczny o pewnym stopniu własnej autonomii. Tak jak w narracji konsekwencjalistycznej AI jest aktantem, gdyż działa, to jednocześnie owo działanie wpisane jest w całą infrastrukturę wartości o charakterze społecznym, kulturowym czy politycznym. W narracji personalistycznej sztuczna inteligencja staje się niejako oddzielona od swojej infrastruktury i zyskuje „własne życie", a nawet potencjalnie prawa, przywileje i godność. W znacznie większym stopniu jest ona także obecna jako *quasi*-podmiot w interakcjach z użytkownikami, gdyż przykładami, jakie często ilustrują ten rodzaj narracji, są roboty społeczne czy czatboty.

W ramach etyki przyjmowanej w narracji personalistycznej dominują dwa podejścia: zorientowane na aktora i zorientowane na pacjenta (czy inaczej odbiorcę). Orientacja na aktora zbliża w wielu miejscach narrację personalistyczną do konsekwencjalistycznej, gdyż zakłada ona, iż sztuczna inteligencja jest źródłem określonego działania, które może być wartościowane moralnie. Skupia ona zatem uwagę na sztucznej inteligencji jako spełniającej określone kryteria traktowania jej jako sprawczego podmiotu moralnego. Orientacja na pacjenta pozwala natomiast postawić pytania o to, w jaki sposób podmiot ludzki odnosić powinien się do sztucznej inteligencji i ją traktować. Etyka może zatem być postrzegana jako budowana albo pod sztuczną inteligencję w jej stosunku do człowieka, albo pod człowieka w jego stosunku do sztucznej inteligencji.

W ramach narracji personalistycznej dostrzega się istotną rolę sztucznej inteligencji w budowaniu relacji z ludźmi na poziomie emocjonalnym i związaną z tym możliwość posiadania przez AI praw. Humanizowanie sztucznej inteligencji, postrzeganie jej jako osoby, czy jako podmiotu o „sztucznym życiu", krytykowane być może jako współczesna forma antropomorfizacji i myślenia magicznego (Musiał 2019), ale bez względu na ocenę sensowności takich praktyk społecznych niewątpliwym jest, jak wykazuje wiele badań, że ludzie z robotami empatyzują, nawet gdy nie są one wyrafinowanymi AI (Suzuki et al. 2015; Darling et al. 2015). Ta empatyzacja może wiązać się ze złudzeniem wynikającym ze skutecznej komunikacji językowej, z widzenia w sztucznej inteligencji ról społecznych, w które sami jesteśmy wrzuceni (Gertz 2018), albo z niechęcią człowieka do patrzenia na cierpienie innych istot (Juchniewicz 2024), niemniej jest rzeczywistym doświadczeniem wielu ludzi i przyczynia się do dyskusji na temat rozszerzenia podmiotowości na aktorów nie-ludzkich przede wszystkim ze względu na fakt, że są bądź mogą być istotami czującymi. Dlatego Nauhäuser (2015) stwierdza, że jeśli sztuczna inteligencja będzie zdolna do odczuwania bólu, to powinna mieć prawa. Nawet jeśli aktualnie nie jesteśmy w stanie

stworzyć sztucznej inteligencji odczuwającej ból, co najwyżej skutecznie imitującą radość czy strach, to projektowanie takich praw z wyprzedzeniem może pozwolić na ich wdrożenie wtedy, gdy taka technologia rzeczywiście się pojawi. Jak zauważa Erica L. Neely (2013):

> (...) jeśli coś zachowuje się wystarczająco podobnie do mnie w szerokim zakresie sytuacji, to powinniśmy rozszerzyć na to status moralny. (…) Widzę, że moralna wina w byciu zbyt konserwatywnym jest znacznie większa niż w ryzyku bycia zbyt hojnym.

Neely chodzi, rzecz jasna, o hojność ontologiczną i moralną. Można powiedzieć, że podmiot ludzki nie traci wiele ze swojej sprawczości, rozszerzając działanie moralne na byty inne niż ludzkie, może natomiast wiele stracić, jeśli do moralności podchodzić będzie zbyt ekskluzywnie.

Bycie istotą czującą, choćby potencjalnie, zakłada zatem, że sztuczna inteligencja powinna być chroniona prawnie w sposób podobny do zwierząt czy sztucznie powoływanych na drodze prawnej osób. W filozofii nowożytnej już Tomasz Hobbes w XVI rozdziale *Lewiatana* wskazał, w jaki sposób można zapewnić reprezentację prawną rzeczy i podmiotów nie-ludzkich, a tym samym realizować ich cele czy dobro. Pośród takich podmiotów, które mogą być reprezentowane przez kogoś innego, a zatem mieć prawa, wymienia m.in. kościół czy most (Hobbes 2009: 246). Z punktu widzenia prawa, sięgającego w swej tradycji do prawa rzymskiego, możliwe jest zatem „chronienie praw sztucznej inteligencji", nawet jeśli ona sama nie może się o owe prawa ubiegać. H. Ashrafian (2015: 325) stwierdza:

> W tym przypadku roboty prawdopodobnie (…) nie replikowałyby się samodzielnie, nie zajmowałyby stanowisk publicznych ani nie byłyby właścicielami ziemi i firm, ale byłyby chronione przez prawo i miałyby możliwość wniesienia wkładu w społeczeństwo poprzez takie przykłady, jak obrona narodów i uczestnictwo w sektorze opieki zdrowotnej.

Jeśli nawet prawa robotów są uzasadnialne, a odpowiednie narzędzia prawne mogą doprowadzić do realizacji tych praw, to wciąż nie jest jasne, w jaki sposób sztuczna inteligencja ma być postrzegana jako podmiot moralny i czy w ogóle za taki podmiot postrzegana być powinna. Podstawowym argumentem jest tutaj brak świadomości – AI nie jest obdarzona świadomością, choć w niektórych postaciach świetnie taką świadomość imituje (przez np. używanie języka), dlatego nie

może być podmiotem moralnym, a i przyznawanie jej praw nie jest konieczne, skoro w zasadzie jej status porównać można do rzeczy. „Roboty są art[e]faktami i dlatego, w oczach wielu, nie mają elementu świadomości, który wydaje się być powszechnie uważany za linię podziału między tym, czy zasługują na etyczne traktowanie, czy nie" (Levy 2009: 210). Jak zauważa David Levy, choć sztuczna inteligencja nie ma świadomości, a przynajmniej nie w sensie, w jakim zakładamy, że mają ją ludzie, to postępujący rozwój badań w obszarze AI może doprowadzić do wyłonienia się tzw. sztucznej świadomości (*artificial consciousness*), czyli formy świadomości innej niż ludzka. Już dzisiaj z badań Jenny Burell wiemy, a przynajmniej możemy „zobaczyć", w jak inny sposób – w porównaniu do ludzkiego sposobu rozpoznawania obiektów – sztuczna inteligencja identyfikuje tak banalne obiekty jak cyfry. Fakt, że AI widzi różnice w stopniu zaczernienia pól pikseli, które to różnice są niewidoczne dla ludzkiego oka, i że ostatecznie nie widzi ona cyfr, a stopnie szarości obiektów, oznacza, że jej sposób identyfikacji i kategoryzacji obiektów nie daje się porównać z tym, jak widzi rzeczy i kategoryzuje je człowiek (Burrell 2016: 5–7). Być może zatem, rozszerzając funkcje kognitywne AI, rzeczywiście będziemy mieć do czynienia ze sztuczną świadomością, która wymaga od nas określenia jej praw. Wendel Wallach i Colin Allen (2009: 39) postulują tzw. moralność funkcjonalną, a więc założenie, że po pierwsze, moralność do sztucznej inteligencji można zaimplementować, a po drugie, że może ona nauczyć się właściwej oceny etycznej swoich działań (zob. Coeckelbergh 2020: 52). John Sullins (2006) idzie o krok dalej i zwraca uwagę, że algorytmy sztucznej inteligencji działają i uczą się w sposób inny niż ludzie. Nie rozumiemy w pełni, dlaczego sztuczna inteligencja rozpoznaje pewne obiekty czy identyfikuje pewne jakości równie skutecznie jak człowiek, gdyż nie robi tego „na ludzki sposób". Owo wymykanie się AI samym programistom pozwala mówić o pewnym poziomie autonomii sztucznej inteligencji. Jeśli zatem jest ona częściowo autonomiczna, jeśli jednocześnie możemy wyjaśnić jej, jakimi zasadami ma się kierować, a ona je rozumie i właśnie w określony sposób postępuje, to zdaniem Sullinsa jest ona podmiotem moralnym (Coeckelbergh 2020: 54).

Levy stawia w związku z tym ważny problem: Czy moralnym jest w ogóle tworzenie przez podmiot ludzki istot potencjalnie czujących i świadomych? Czy myśląc o możliwościach rozwoju sztucznej inteligencji nie zapominamy o tym, jak tworzenie takich podmiotów/przedmiotów wpływa na człowieka i jego stosunek do siebie samego, innych ludzi i wreszcie osób nie-ludzkich? W tym kierunku idzie krytyka etyki sztucznej inteligencji zaproponowana przez Joannę Bryson (Bryson, 2010). W kontrowersyjnym artykule „Robots should be slaves" Bryson argumentuje, że sztuczna inteligencja nie jest wolna od ludzkiej woli, nie jest autonomiczna w swoim działaniu, a wszelkie próby uczynienia jej taką,

to tylko zrzucanie przez ludzi odpowiedzialności za to, co sami tworzą. Bryson (2010: 65) stwierdza:

> Projektujemy, produkujemy, posiadamy i obsługujemy roboty. Ponosimy za nie całkowitą odpowiedzialność. Określamy ich cele i zachowanie, bezpośrednio lub pośrednio poprzez określenie ich inteligencji lub nawet bardziej pośrednio poprzez określenie sposobu, w jaki nabywają własną inteligencję. Ale na końcu każdego pośrednictwa leży fakt, że nie byłoby robotów na tej planecie, gdyby nie celowe ludzkie decyzje o ich stworzeniu.

Czy zatem w narracji personalistycznej możliwe jest mówienie o sztucznej inteligencji jako odpowiedzialnej albo godnej zaufania? Model personalistyczny, jako zbudowany na koncepcji reprezentowania uzasadnialnych praw sztucznej inteligencji przez człowieka mówiącego w jej imieniu, zakłada, że sztuczna inteligencja nie może być w pełni odpowiedzialna za swoje czyny, bo owa odpowiedzialność spoczywa na człowieku. Jeśli jednakże wierzymy w to, że uda się w przyszłości powołać do życia sztuczną inteligencję czującą albo świadomą, to problem z takimi postulatami etycznymi polega na ich czystej hipotetyczności. Spełniony musi być bowiem podstawowy warunek zamiany „jeśli będzie" na „jest".

III. Krytyka narracji o „odpowiedzialnej sztucznej inteligencji" i „AI godnej zaufania"

Narracja konsekwencjalistyczna i narracja personalistyczna, mówiąc o „odpowiedzialnej sztucznej inteligencji" czy „AI godnej zaufania", mówią *de facto* o dwóch różnych problemach. Pierwszym z nich jest „zaufanie do" i „poczucie odpowiedzialności za", które powinno obowiązywać projektantów i producentów tego typu technologii oraz osób odpowiedzialnych za proces tworzenia, wdrażania i korzystania z AI. Ponadto ufność w technologię sztucznej inteligencji rośnie wraz ze skutecznością w realizacji praktycznych celów – niemożliwe jest wycofanie się z technologii, która tak wiele rzeczy ludziom ułatwia i z kapitalistycznego punktu widzenia optymalizuje i usprawnia. Koszty etyczne można zatem minimalizować, szanse wyrównywać, algorytmy lepiej trenować, sięgać do danych wolnych od uprzedzeń, ale na końcu i tak pozostanie nam po prostu nadzieja, że technologia ta nie zostanie przeciwko nam użyta, a będzie realizować postulat „dobra wspólnego". Narracja konsekwencjalistyczna, która jest źródłem dla tak rozumianej „moralnej AI", traktuje tę technologię bardziej jako instrument ludzkiego dobrobytu i rozwoju, medium w relacjach międzyludzkich, technologię, na którą delegować możemy określone

wartości i narzędzie zachowujące przynajmniej neutralność aksjologiczną. W tym sensie to podmiot ludzki i jego potrzeby oraz często praktyczne interesy mają być w takim rozumieniu AI chronione, a sztuczna inteligencja godna zaufania i tak nie jest nigdy „puszczona wolna" i choćbyśmy jej ufali, to wymagana jest od ludzi jej kontrola i nadzór. Sztuczna inteligencja godna zaufania przesuwa tak naprawdę dyskurs etyczny w kierunku polityki i tego, kto i w jaki sposób decyduje o korzystaniu z AI.

Drugim problemem jest postrzeganie sztucznej inteligencji jako partnera w naszej codzienności, a więc asystenta głosowego, hologramu, czatbota, wirtualnej persony, robota społecznego, który zachowując się tak jak człowiek, albo po prostu jak podmiot moralny, stawia przed nami wyzwanie rozszerzenia prawa i moralności na AI. Narracja personalistyczna, mówiąc o odpowiedzialnej AI, zwraca uwagę na fakt, że już istniejące narzędzia prawne pozwalają zabezpieczyć prawa sztucznej inteligencji choćby przez należyte reprezentowanie jej interesów w sposób porównywalny bodaj do ochrony interesów takich sztucznych podmiotów jak spółki. Ale taki stosunek do AI, choć traktuje ją jako podmiot prawa, nie przyznaje jej odpowiedzialności za czyny. Jeśli natomiast rozszerzyć tę odpowiedzialność i uznać, że jeśli AI zachowuje się tak, jak zachowałby się moralnie postępujący człowiek, innymi słowy, jeśli zdaje się ona rozumieć moralność jako koncepcję, to zobowiązani jesteśmy przynajmniej wziąć pod uwagę jej możliwe prawa. Wraz z rozwojem samej technologii można założyć, że w przyszłości uda się wytworzyć czującą AI, a być może i uznać, że posiada ona pewną sztuczną świadomość – powinniśmy zatem rozszerzyć prawa na tego typu osoby nie-ludzkie. Narracja personalistyczna ignoruje jednakże, że sztuczna inteligencja, choć może odtwarzać etykę i określone zasady moralne, to nie jest sama w stanie ich stworzyć. Nie jest zatem zdolna do autorefleksji moralnej jako wyrazu autokrytyki swojego działania ani do stworzenia etyki jako samej koncepcji. Innymi słowy – człowiek zdolny jest do autonomicznego uzasadnienia maksymy swojego postępowania i do myślenia opartego na refleksji – zwrocie ku sobie, ale przez konieczne wyjście w swoim myśleniu i wartościowaniu poza siebie, w kierunku praw, na które mogliby wyrazić zgodę inni ludzie. Sztuczna inteligencja nie buduje intersubiektywności i poczucia siebie w procesie własnego, fenomenologicznego stawania się. Może się uczyć lepiej rozpoznawać pewne wzorce, a nawet coraz lepiej dostosowywać do nich wartościowanie moralne, ale nie staje się przez to bardziej ludzka. Ponadto człowiek może chcieć zmiany, może ukierunkować zatem swoje działanie na realizację wartości, co pozwala mu być właśnie odpowiedzialnym za własne czyny; trudno podobnej autorefleksji oczekiwać od AI.

W ramach narracji personalistycznej badacze próbują z jednej strony dokonać ontologicznego rozszerzenia podmiotowości i sprawstwa, a co za tym idzie,

odpowiedzialności poza człowieka, i w zasadzie sam ten kierunek uznać należy za szlachetny „ruch wyprzedzający" możliwy rozwój sztucznej inteligencji, jednakże z drugiej popadają w błąd, który uznać można za „przesuwanie" czy „zrzucanie odpowiedzialności" za czyny generowane przez AI. Jeśli za samo pojawienie się sztucznej inteligencji jako określonego typu technologii odpowiada człowiek, to i za możliwy stopień jej emancypacji i autonomii odpowiadać powinien także człowiek. Wątpliwości etyczne budzi zresztą samo konstruowanie technologii, która mogłaby zyskać świadomość czy czuć, i to niekoniecznie ze względu na obawy, do czego taka AI byłaby zdolna, ale ze względu na wyraźny w obydwu narracjach praktyczny stosunek ludzi do tej technologii. Sztuczna inteligencja została skonstruowana, by rozwiązywać określone problemy, by przyśpieszać procesy analityczne, by poddawać obróbce wielkie zbiory danych, by te dane zbierać, by nadzorować określone procesy, by w konkretny sposób wchodzić w interakcję z ludźmi. Ta szczególna, praktyczna specjalizacja AI i fakt, że mamy do czynienia nie z jedną sztuczną inteligencją, a wieloma ich postaciami, programami, aplikacjami i urządzeniami, ujawnia, że wartości moralne pojawiają się w ramach tworzenia tej technologii dopiero *post factum*. Refleksja etyczna w obszarze AI to także skutek rozwoju samej AI, a nie przyczyna *a priori* jako np. troska o rozwój takich technologii, które służyć będą ludzkości. Innymi słowy, etyczny wymiar AI zauważyliśmy dopiero wtedy, gdy okazało się, że Siri czy Alexa mogą zabrać pracę lektorom; gdy sztuczna inteligencja przemysłowa zaczęła zastępować pracowników fabrycznych; gdy w starzejących się społeczeństwach dotarło do nas, że kiedyś pielęgniarkę może zastąpić robot asystujący, i gdy ludzie uznali, że rozmowa z hologramem jest bardziej satysfakcjonująca niż z innym człowiekiem. Przy okazji badań nad algorytmami odkryliśmy także, że technologia nie jest neutralna, że sztuczna inteligencja powiela stereotypy, uprzedzenia (Rafanelli 2022; Waelen, Wieczorek 2022), że manipulacyjny sposób posługiwania się AI szkodzi demokracji (Coeckelbergh 2024) i ma poważne konsekwencje społeczno-polityczne, a jakość danych, na których bywa trenowana, budzi wątpliwości etyczne co do ich rzetelności czy sposobu pozyskania. Ostatecznie zatem sztuczna inteligencja mówi nam więcej o nas samych niż o sobie jako Innym. Nie przesądza to jednakże, że sztuczna inteligencja nie jest Innym, którego należałoby potraktować poważnie (Gunkel 2018). Wskazuje jedynie, że wciąż borykamy się z trudnością, jaką jest formułowanie samej etyki w obszarze AI. David J. Gunkel przypomina bowiem o dobrze znanym twierdzeniu Davida Hume'a, że ze zdań o faktach nie wyprowadza się zdań o wartościach (zob. Ślęczek-Czakon 1992). Innymi słowy, zdaniem Gunkela błędem byłoby wartościowanie sztucznej inteligencji jako „moralnej" albo „godnej zaufania" dlatego, że posiada ona określone jakości, które uznajemy za podstawę „moralności" albo „godnego zaufania", gdyż wówczas wartość moralną udowodnić

musi podmiot/przedmiot przez wpisanie się w określony normatywny horyzont oczekiwań. Logika tak rozumianej etyki to przejście od „X jest A" bądź „X posiada cechę B" do „X powinno się traktować w sposób M". Z historii wiemy, że taki normatywny horyzont, po pierwsze, nie jest stały i zmienia się w czasie, po drugie, fakt, że w różnych momentach w czasie uznawano za podstawę moralności posiadanie określonych jakości, np. bycie mężczyzną, bycie reprezentantem rasy białej, bycie rozumnym w sensie sprawnej komunikacji językowej, bycie człowiekiem dorosłym itd., wcale nie oznacza, że rzeczywiście mieliśmy do czynienia z wartościowaniem, które uznalibyśmy za moralne. Gunkel sugeruje zatem, że najpierw musimy zastanowić się, jakimi wartościami należałoby się kierować w ramach działania moralnego, a w drugim kroku zweryfikować, czy my, jako ludzie, odnosimy się do bytów nie-ludzkich w taki sposób, jaki skądinąd uważamy za moralny. Gunkel zgadza się tu częściowo z Markiem Coeckelberghiem, który zauważył, że niemoralne traktowanie technologii nie oznacza pogwałcenia samej technologii, ale stanowi moralną wadę – jest przeciwieństwem cnoty (Coeckelbergh 2018: 145; zob. Darling 2012), co wpływa na podmiot ludzki w sposób deprawujący jego charakter. Gunkel przywołuje nawet mocniejszy argument, gdyż powołuje się na Emmanuela Lévinasa i jego założenie, iż etyka poprzedza ontologię (Gunkel 2018: 269). Ostatecznie bowiem, jak argumentuje Gunkel, w etyce chodzi o to, jak człowiek odnosi się do inności i różnorodności, która dotyczy rzecz jasna ludzi, ale może dotyczyć także bytów nie-ludzkich.

Podsumowanie

W niniejszym artykule wyszczególniłam dwie, dominujące narracje, a więc pewne zbiory przekonań, wyobrażeń i argumentów wokół moralnej sztucznej inteligencji. Pierwsza sprowadza się do traktowania sztucznej inteligencji jako moralnej ze względu na konsekwencje, jakie jej działanie wywołuje. Sama AI w tej narracji traktowana jest jako narzędzie, medium albo po prostu element infrastruktury wartości, w który wpisać należy określone zasady działania, tak by minimalizować niepożądane skutki. Jak pokazałam, narracja konsekwencjalistyczna, postulując „AI godną zaufania", określa tak naprawdę spektrum wartości, które powinny być brane pod uwagę, gdy technologia ta jest projektowana, wdrażana i używana. Problem w tym, że wartości te można zdefiniować jedynie w bardzo szeroki sposób, gdyż szczegółowe aplikacje sztucznej inteligencji do życia społecznego są często bardzo punktowe, a także trudne do przewidzenia. Można zatem postulować „sprawiedliwą AI", która nie powiela uprzedzeń i stereotypów, ale w praktyce owa sprawiedliwość może być rozumiana na różne sposoby w sensie kulturowym, społecznym, a zwłaszcza politycznym, a jej realizacja pozostaje w rękach człowieka.

Druga narracja, personalistyczna, stara się spojrzeć na AI jak na podmiot moralny tak w sensie sprawczości moralnej, jak i bycia podmiotem praw i przywilejów adresowanych do niej przez człowieka. Narracja ta pozwala na zwrócenie uwagi na ontologię, a więc jakości rzeczy, którym przypisujemy status moralny. Najczęściej prowadzi to do dyskusji warunków, jakie spełnić musi przedmiot, by stać się podmiotem moralnym – a więc musi posiadać umysł lub być istotą czującą. Tak, jak sama narracja personalistyczna stanowi ważny głos w dyskusji nad tym, co w podmiocie, a nie konsekwencjach jego działania decyduje o tym, że przypisujemy mu moralność, to problematyczne jest adresowanie tej narracji do sztucznej inteligencji. Dotychczas w przypadku AI mamy do czynienia z postulowaniem pewnych jakości, a nie ich faktycznym posiadaniem przez tę technologię. Narracja personalistyczna stawia ontologię przed etyką, a to samo w sobie może być błędnym podejściem. Wymaganie od przedmiotu/podmiotu, by udowodnił nam posiadanie przez siebie cech uznawanych za niezbędne do posiadania moralności, jest pewną formą moralnego konserwatyzmu i stawianiem podmiotu ludzkiego za wzór, tak jakby w samej etyce nie trwały od wieków spory, kim jest i w jakich warunkach mówić możemy w ogóle o podmiocie moralnym. Pewną propozycją wyjścia poza etykę zorientowaną na ontologię jest inspirowana Lévinasem etyka otwarcia na Innego, a więc otwarcia na radykalną różnicę. Być może w poszanowaniu Inności jako rzeczywiście innej, a nie podobnej do nas samych (pod pewnymi warunkami), tkwi szansa na zbudowanie *moralnych relacji* ze sztuczną inteligencją, a nie tylko definiowanie moralnej AI czy AI godnej zaufania.

Bibliografia

[1] Airoldi, M. (2022). *Machine Habitus. Towards a Sociology of Algorithms*. Cambridge, UK, Medford, USA: Polity Press.
[2] Andreson M., Anderson S. L. (eds) (2011). Machine Ethics. Cambridge: Cambridge University Press.
[3] Ashrafian, H. (2015). Artificial intelligence and robot responsibilities: Innovating beyond rights. *Science and Engeenering Ethics* 21 (2): 317–236. DOI 10.1007/s11948-014-9541-0
[4] Bryson, J. J. (2010). Robots should be slaves. W Y. Wilks (Ed.), *Close Engagements with Artificial Companions: Key social, psychological, ethical and design issues* (s. 63–74). John Benjamins Publishing Company. https://doi.org/10.1075/nlp.8.11bry
[5] Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society*, *3* (1), https://doi.org/10.1177/2053951715622512
[6] Coeckelbergh. M. (2018). Why Care About Robots? Empathy, Moral Standing, and the Language of Suffering. *Kairos. Journal of Philosophy & Science* 20, 141–158. DOI 10.2478/kjps-2018-0007

[7] Coeckelbergh, M. (2020). *AI Ethics.* Cambridge, Massachusetts, London, England: The MIT Press.
[8] Coeckelbergh, M. (2024). *Why AI Undermines Democracy and What To Do About It*. Cambridge: Polity Press.
[9] Darling, K. (2012). Extending Legal Protection to Social Robots: The Effects of Anthropomorphism, Empathy, and Violent Behavior Towards Robotic Objects (April 23, 2012). Robot Law, Calo, Froomkin, Kerr eds., Edward Elgar 2016, We Robot Conference 2012, University of Miami, https://ssrn.com/abstract=2044797
[10] Darling, K., Nandy, P., Brazeal, C. (2015). Empathic concern and the effect of stories in human-robot interaction. Proceedings of the 24th IEEE International Symposium on Robot and Human Interactive Communication, ed. IEEE, 770–775. DOI: 10.1109/ROMAN.2015.7333675
[11] Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds and Machines*, *28* (4), 689–707. https://doi.org/10.1007/s11023-018-9482-5
[12] Gertz, N. (2018). Hegel, the Struggle for Recognition, and Robots. *Techné: Research in Philosophy and Technology*, *22* (2), 138–157. https://doi.org/10.5840/techne201832080
[13] Goggin, G., Newell, C. (2003), *Digital disability: The social construction of disability in new media*. Lanham, Maryland, USA: Rowman & Littlefield.
[14] Gunkel, D.J. (2018). *Robot Rights*. Cambridge, Massachusetts, London, England: The MIT Press.
[15] Hobbes, T. (2009). *Lewiatan*. przeł. C. Znamierowski, Warszawa: Fundacja Aletheia.
[16] Juchniewicz, N. (2024). Towards (Unilateral) Recognition of "the Technological Other" – Vulnerability, Resistance and Adequate Regard. *Ethics in Progress* Vol. 15 No 2. 120–136. https://doi.org/10.14746/eip.20242.8
[17] Komisja Europejska, Grupa ekspertów wysokiego szczebla ds. sztucznej inteligencji. (2019). Wytyczne w zakresie etyki dotyczące godnej zaufania sztucznej inteligencji. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai [pobrane w języku polskim: 6.02.2025]
[18] Latour, B. (2013). Technologia jako utrwalone społeczeństwo, przeł. Ł. Afeltowicz. *AVANT,* wol. IV, nr 1/2013, s. 17–48.
[19] Levy, D. (2009). The ethical treatment of artificially conscious robots. *International Journal of Social Robotics* 1 (3): 209–216. https://doi.org/10.1007/s12369-009-0022-6
[20] Musiał, M. (2009). *Enchanting Robots. Intimacy, Magic, and Technology*. Cham: Palgrave Macmillan.
[21] Neely, E.L. (2014). Machines and the moral community. Philosophy & Technology 27 (1): 97–111. https://doi.org/10.1007/s13347-013-0114-y
[22] Neuhäuser, C. (2015). Some skeptical remarks regarding robot responsibility and a way forward. In: *Collective Agency and Cooperation in Natural and Artificial Systems: Explanation, Implementation, and Simulation*, ed. Catrin Misselhorn, 131–148. New York: Springer.

[23] Rafanelli, L.M. (2022). Justice, injustice, and artificial intelligence: Lessons from political theory and philosophy. *Big Data & Society* 9 (1). https://doi.org/10.1177/20539517221080676
[24] Sullins, J. (2006). When is a Robot a Moral Agent? *International Review of Information Ethics* 6, 23–30.
[25] Suzuki, Y., Galli, L., Ikeda, A., Itakura, S., Kitazaki, M. (2015). Measuring empathy for human and robot hand pain using electroencephalography. *Scientific Reports* 5: 15924. https://doi.org/10.1038/srep15924
[26] Ślęczek-Czakon, D. (1992). Gilotyna Hume'a. *Folia Philosophica* 10. 147–159.
[27] Verbeek, P.-P. (2011). *Moralizing Technology*. Chicago: University of Chicago Press.
[28] Verbeek, P.-P. (2008). Obstetric Ultrasound and the Technological Mediation of Morality: A Postphenomenological Analysis. *Human Studies* 31, 11–26. https://doi.org/10.1007/s10746-007-9079-0
[29] Verbeek, P.-P., Slob, A. (eds) (2006). *User Behavior and Technology Development. Shaping Sustainable Relations Between Consumers and Technologies*. Dordrecht: Springer.
[30] Wallach, W., Allen, C. (2009). *Moral Machines: Teaching Robots Right from Wrong*. Oxford: Oxford University Press.
[31] Waelen, R., Wieczorek, M. (2022). The Struggle for AI's Recognition: Understanding the Normative Implications of Gender Bias in AI with Honneth's Theory of Recognition. *Philosophy & Technology* Vol. 35, No 53. https://doi.org/10.1007/s13347-022-00548-w
[32] Winner, L. (1986). Do Artifacts Have Politics? In: *The Whale and the Reactor: a search for limits in an age of hight technology*. Chicago: Chicago University Press, 19–39.