Edukacja Filozoficzna
ISSN 0860-3839
DOI: 10.14394/edufil.2022.0001
ORCID: 0000-0002-5745-5773

Etyka sztucznej inteligencji w dokumentach Unii Europejskiej w latach 2017–2020

Izabela Lipińska
(Uniwersytet Kardynała Stefana Wyszyńskiego w Warszawie,
Wydział Filozofii Chrześcijańskiej)

1. Wprowadzenie

Sztuczna inteligencja (SI) to jedna z głównych technologii XXI wieku. Olbrzymie nakłady finansowe i osobowe kierowane do prac nad SI pokazują, jak wielkie są oczekiwania wobec inteligentnych technologii. Równie wielkie są jednak obawy co do negatywnych konsekwencji ich użycia w zakresie demokracji, gospodarki, ekonomii oraz wobec wszystkich obszarów istotnych dla prawidłowego funkcjonowania państw, społeczeństw, grup i jednostek. Nieprzewidywalność kształtowania się SI, wynikająca z uczenia głębokiego, nakazuje, by poszukiwać także innych form zapewnienia bezpieczeństwa użytkowania niż tylko analiza ryzyka. Jedną ze skuteczniejszych form wydaje się prowadzenie niekomercyjnych badań w obszarze etyki SI i dążenie do egzekwowania ustaleń etycznych na gruncie prawnym. Szczególnie ważne zdaje się polityczne, naukowe i społeczne zaangażowanie w nadawanie wyższego priorytetu kwestiom etycznym aniżeli komercyjnym. Efektem politycznego zaangażowania organów Unii Europejskiej są dokumenty podejmujące kwestię etyki SI – te omówione w artykule organy unijne wskazują jako najistotniejsze. W celu lepszego zrozumienia podejścia proponowanego w treści dokumentów zarysowana zostanie kwestia wartości w kontekście SI oraz dokumentów UE.

2. Wartości a sztuczna inteligencja

Termin „wartość" wywodzi się od łacińskiego słowa *valere* oznaczającego „być wartościowym"[1], „być zdrowym, mieć się dobrze; mieć wpływ, znaczenie, móc"[2]. Niektórzy badacze nawiązują również do łacińskiego słowa *validus*, które znaczy „mocny, silny, żwawy, obronny, wpływowy, skuteczny", i definiują wartość jako coś, co „posiada pewną siłę, aby nas pociągnąć ku sobie"[3]. W języku potocznym zazwyczaj słowo to odnosi się do czegoś z różnych względów cennego, o co warto zabiegać[4]. Termin używany jest w filozofii, psychologii, socjologii, ekonomii czy matematyce i w każdej z tych dziedzin oznacza nieco co innego. Istnieje zagrożenie pomieszania poziomów rozumienia terminu „wartość", a co za tym idzie określania danego artefaktu technologicznego jako wartościowy w sensie aksjologicznym, podczas gdy stanowi on wartość na odmiennym gruncie.

Wartości uznawane są za coś, „co budzi oceniające uznanie człowieka"[5], i w tym rozumieniu możemy usytuować je w wartościującej części normatywności. Tak ujmowane służą człowiekowi do oceny bytów w kategoriach dobra i zła. Kwestią problematyczną są racje, na podstawie których dokonywana jest ocena. Jeśli wartość rozumiana jest jako wynik wartościowania, wtedy najwyższą wartością staje się to, co cenione jest najbardziej. Nie zawsze jednak pozytywna ocena pokrywa się z istnieniem normatywnych przyczyn uznania wartości. W efekcie uzyskujemy cztery kategorie: wartość ceniona i mająca normatywne przyczyny uznania wartości, wartość ceniona a niemająca normatywnych przyczyn uznania wartości, wartość nieceniona a mająca normatywne przyczyny uznania wartości oraz wartość nieceniona i niemająca normatywnych przyczyn dla uznania wartości. Ibo van de Poel tak ujmuje tę kwestię:

> istnieje pewna zgodność między normatywnymi racjami wartościowania a faktem, że coś ma wartość. Tak więc, jeśli coś ma wartość, istnieją racje normatywne, aby to docenić, ale to nie znaczy, że zawsze jest ono rzeczywiście doceniane, ponieważ ludzie mogą nie doceniać na podstawie racji normatywnych. I odwrotnie, nie oznacza to również, że jeśli coś jest cenione, to ma war-

[1] S. Kowalczyk, *Filozoficzne koncepcje wartości,* „Collectanea Theologica" 1986, nr 1 (222), s. 37.
[2] G. Żuk, *Edukacja aksjologiczna. Zarys problematyki*, Lublin 2016, s. 18.
[3] Por. tamże.
[4] Por. M. Krąpiec, *Wartość*, w: *Powszechna Encyklopedia Filozofii*, http://www.ptta.pl/pef (dostęp: 19.02.2022).
[5] Tamże.

> tość, ponieważ ludzie czasami cenią na podstawie niewłaściwych powodów lub w ogóle nie mają (normatywnych) powodów[6].

W kontekście sztucznej inteligencji, którą można zaprojektować ku określonym wartościom, jednym z najważniejszych wyzwań staje się zatem ustalenie, które z wartości są tymi, które powinny być cenione ze względu na ich istotę, i przełożenie tych ustaleń na akceptowalny społecznie oraz atrakcyjny komercyjnie projekt SI.

Wartości badane są na każdym z trzech podstawowych gruntów etyki: na gruncie metaetyki, etyki stosowanej i etyki normatywnej. Etyka normatywna ma kluczowe znaczenie dla zrozumienia i zastosowania zasad moralnych przy projektowaniu sztucznych systemów[7]. Najistotniejsze kierunki w ramach etyki normatywnej to konsekwencjalizm, deontologia i etyka cnót. W projektowaniu SI korzysta się z osiągnięć każdej z nich. Zaprogramowanie etycznej sztucznej inteligencji, czyli praktyczne zastosowanie opisanych koncepcji i zasad, jest obecnie jednym z wyzwań technologicznych.

SI znajduje się aktualnie w fazie rozwoju określanej jako słaba (*weak AI*)[8]. Ogólna SI do tej pory nie została stworzona i nikt nie potrafi z całą pewnością stwierdzić, czy – a jeśli tak, to kiedy – to się stanie. Oczywiście nie oznacza to, że nie należy już dziś badać kwestii etyki ogólnej SI, opierając się na technologicznych prognozach co do jej własności. Znaczy to natomiast, że etyka musi przede wszystkim odpowiadać na wyzwania wynikające z realnych możliwości technologicznych. Etyka SI nie może być postrzegana jako „futurystyczny cel"[9], ale jako solidna dziedzina naukowa mająca realny wkład w tworzenie inteligentnych maszyn. Najistotniejsze wydaje się prowadzenie badań w ramach etyki stosowanej. Jak pisze Thilo Hagendorff:

> Etyka musi częściowo przekształcić się w „mikroetykę". Oznacza to, że w pewnych momentach musi nastąpić zasadnicza zmiana poziomu abstrakcji, o ile etyka ma na celu wywarcie określonego wpływu i to wpływu w dyscyplinach technicznych i praktyce badań i rozwoju sztucznej inteligencji […]. Na drodze od etyki do „mikroetyki" musi nastąpić transformacja od etyki do etyki tech-

[6] I. van de Poel, *Embedding Values in Artificial Intelligence (AI) Systems*, „Minds and Machines" 2020, t. 30, s. 388.

[7] Por. tamże, s. 37.

[8] W literaturze spotkać można także określenie „wąska SI" (*narrow AI*).

[9] A. Martinho i in., *Perspectives about Artificial Moral Agents*, „AI Ethics" 2021, nr 1, s. 487.

> nologii, etyki maszyn, etyki komputerów, etyki informacji i etyki danych. Dopóki etycy powstrzymują się od tego, pozostaną widoczni w opinii publicznej, ale nie w społecznościach zawodowych[10].

Badacze z Delft University idą nieco dalej i twierdzą, że „mikroetyka" powinna być badana w odniesieniu do konkretnej dziedziny, na przykład opieki zdrowotnej czy transportu. „Etyka dotycząca systemu AI jest stosowana i kształtowana do dziedziny, w której będzie działać, a nie odwrotnie"[11].

Organy unijne postawiły sobie za cel zbudowanie ekosystemu sztucznej inteligencji opartego na wspólnocie europejskich wartości, wśród których godność człowieka jest wartością centralną. W związku z tym europejskie podejście do SI nazywane jest *human-centered AI*, co można przetłumaczyć jako „SI ukierunkowana na człowieka". Unia chce stać się liderem w etycznym podejściu do technologii i wyróżniać się tym na arenie światowej. W tym celu: „Europa musi określić wizję normatywną dotyczącą przyszłości stojącej pod znakiem SI, którą chce zrealizować"[12]. Jednym z kroków prowadzących do tego celu było przedstawienie 27 stycznia 2017 roku *Sprawozdania z posiedzenia Parlamentu Europejskiego zawierającego zalecenia dla Komisji w sprawie przepisów prawa cywilnego dotyczących robotyki*.

3. Problematyka etyczna w sferze robotyki

Sprawozdanie z posiedzenia Parlamentu Europejskiego rozpoczyna się od dość specyficznych jak na dokument urzędowy słów:

> mając na uwadze, że od postaci Frankensteina stworzonej przez Mary Shelley po antyczny mit o Pigmalionie poprzez historię praskiego golema i robota Karla Čapka (autora pojęcia „robot") ludzkość zawsze snuła fantazje na temat możliwości stworzenia inteligentnych maszyn, najczęściej androidów o cechach ludzkich; mając na uwadze, że w związku z tym, że ludzkość stoi obecnie u progu ery, w której coraz bardziej zaawansowane roboty, komputery, androidy i inne wcielenia sztucznej inteligencji wydają się dawać początek

[10] T. Hagendorff, *The Ethics of AI Ethics: An Evaluation of Guidelines*, „Minds and Machines" 2020, t. 30, s. 111.

[11] A. Martinho i in., *Perspectives*…, dz. cyt., s. 487.

[12] *Wytyczne w zakresie etyki dotyczące godnej zaufania sztucznej inteligencji*, Bruksela 2019, s. 11.

> nowej rewolucji przemysłowej, która prawdopodobnie nie ominie żadnej warstwy społecznej, niezmiernie ważne jest, by przepisy uwzględniały prawne i etyczne implikacje i skutki tych zmian bez hamowania innowacji[13].

Szerokie nawiązanie kulturowe pokazuje, że sztuczna inteligencja odnosi się nie tylko do czysto technicznie ujętej poprawy funkcjonowania ludzkości, ale także do głęboko zakorzenionych w człowieku marzeń o przekraczaniu własnych słabości oraz o niczym nieograniczonej mocy twórczej. Niestety zarówno powieści *Frankenstein* oraz *R.U.R.*, jak i mit o praskim golemie nie kończą się dla człowieka pozytywnie. Historii króla Cypru także nie można uznać za szczęśliwą, bo choć bierze on ślub z wyrzeźbioną przez siebie idealną kobietą, to jednak powodem jej stworzenia było jego rozczarowanie człowiekiem. Czy zasadne jest zatem wspominanie tych dzieł kultury tylko w kontekście rewolucji przemysłowej? Rewolucja wynikająca ze stworzenia sztucznej inteligencji dotyczyć będzie jeszcze, a może przede wszystkim, kondycji inter- i intrapersonalnej każdego z nas.

Schemat zapisu: „ważne jest, by prawne i etyczne implikacje i skutki tych zmian bez hamowania innowacji" odpowiada strukturze dylematu Collingridge'a. Już w 1980 roku David Collingridge wskazał na paradoks kontroli w odniesieniu do rozwoju technologii. Na początkowym etapie rozwoju technologii kontrola może bowiem nie mieć sensu ze względu na brak możliwości niepożądanych konsekwencji jej użycia. W momencie odkrycia tych konsekwencji kontrola staje się jednak niezwykle trudna ze względu na integralność z systemami ekonomicznym i społecznym[14]. Na szczególną wagę tego dylematu w odniesieniu do technologii, jaką jest SI, wskazuje między innymi węgierski filozof Mihály Héder[15]. Oczywiście należy mieć na względzie trudności generowane przez tę zależność, niedopuszczalne natomiast jest wykorzystywanie jej jako racji uzasadniającej zaniechania w zakresie ochrony fundamentalnych wartości w procesie projektowania SI.

Autorzy sprawozdania podkreślają konieczność przestrzegania zasad bezpieczeństwa i etyki, a w razie ich naruszenia – pociągnięcie twórców do odpowiedzialności prawnej[16]. „Odpowiedzialność" to jedna z ważniejszych kategorii

---

[13] *Sprawozdanie zawierające zalecenia dla Komisji w sprawie przepisów prawa cywilnego dotyczących robotyki*, Bruksela 2017, s. 3.

[14] W. Leibert, J.C. Schmidt, *Collingridge's Dilemma and Technoscience*, „Poiesis Prax" 2010, t. 7, s. 57.

[15] M. Héder, *A Criticism of AI Ethics Guidelines*, „Információs Társadalom" 2020, t. 20, nr 4, s. 61.

[16] Por. *Sprawozdanie zawierające zalecenia dla Komisji*…, dz. cyt., s. 5.

wymienianych w analizowanym dokumencie. Wiąże się z autonomicznością maszyny, czyli jej zdolnością do „podejmowania decyzji i do realizacji ich w świecie zewnętrznym, niezależnie od kontroli zewnętrznej lub wpływu z zewnątrz”[17]. Autonomia ta ma jednak charakter techniczny, to znaczy – zależy od sposobu zaprogramowania przez człowieka, co stanowi szansę na zidentyfikowanie osoby odpowiedzialnej za konkretny algorytm. Autorzy sprawozdania podkreślają konieczność określenia zdolności prawnej oraz statusu inteligentnych maszyn w celu zapewnienia przejrzystości i pewności prawnej zarówno dla projektantów, producentów, jak i użytkowników w całej UE. Wydaje się, że postulat prawnej pewności w odniesieniu do maszyny, która miałaby posiadać zdolność prawną, jest niestety nietrafiony. Prawną zdolność posiadałaby bowiem, o ile w ogóle, ogólna SI. Waga prawnej odpowiedzialności za działania ogólnej SI jest tymczasem nieadekwatna do nieprzewidywalności co do własności ogólnej SI (a co za tym idzie – co do charakteru podejmowanych przez nią działań) wynikającej z etapu rozwoju technologii. Wdrożenie tego postulatu na obecnym etapie mogłoby skutkować zahamowaniem innowacji.

W sprawozdaniu autorzy wezwali komisje unijne do „zaproponowania wspólnej unijnej definicji systemów cyberfizycznych, systemów autonomicznych, inteligentnych robotów autonomicznych oraz ich podkategorii przy uwzględnieniu następujących cech inteligentnych robotów: zdobywanie autonomii, zdolność samokształcenia, przynajmniej minimalna forma fizyczna, dostosowywanie swoich zachowań i działań do otoczenia, brak funkcji życiowych w sensie biologicznym”[18], a także do zaproponowania nowych zasad etycznych, to jest etycznych ram dla sztucznej inteligencji[19]. Powinny one zostać stworzone w oparciu na zasadach: przynoszenia korzyści, nieszkodliwości, autonomii i sprawiedliwości, a także z zachowaniem wartości ujętych w art. 2 Traktatu o Unii Europejskiej oraz w Karcie Praw Podstawowych Unii Europejskiej. Te wartości to: „godność ludzka, równość, sprawiedliwość i równouprawnienie, brak dyskryminacji, świadoma zgoda, ochrona życia prywatnego i rodzinnego oraz ochrona danych”[20]. Zachować należy także wartości i zasady stanowiące podstawę prawa UE: „brak stygmatyzacji, przejrzystość, autonomię, odpowiedzialność

[17] Tamże, s. 7.
[18] Tamże, s. 9.
[19] Tamże, s. 11.
[20] Tamże.

jednostki”[21]. Postuluje się sformułowanie zasad etycznych dla wszystkich grup mających wpływ na SI oraz tych, na które SI może mieć wpływ. Te grupy to w szczególności: naukowcy i badacze, projektanci, użytkownicy oraz członkowie komitetów etycznych oceniających procesy tworzenia inteligentnych maszyn. W sprawozdaniu autorzy zawarli swoją propozycję kodeksów postępowania dla wymienionych grup[22].

Nieco ponad rok później ukazał się kolejny istotny dokument, ukazujący, jak etyka SI powoli umacniała swoją pozycję w unijnych strukturach, a przede wszystkim w świadomości rządzących. Dokument ten nosi tytuł *Statement on Artificial Intelligence, Robotics and 'Autonomous' Systems*. Autorami są członkowie Europejskiej Grupy do spraw Etyki w Nauce i Nowych Technologiach.

4. Autonomia, godność i odpowiedzialność

Opublikowane w marcu 2018 roku przez Europejską Grupę do spraw Etyki w Nauce i Nowych Technologiach *Oświadczenie w kwestii Sztucznej Inteligencji, robotyki oraz autonomicznych systemów* pokazuje, jak szybko muszą ewoluować badania nad etyką SI. Minął nieco ponad rok od ukazania się przedstawionego w poprzedniej części sprawozdania, a technologia poczyniła niesamowite postępy między innymi w zakresie uczenia maszynowego, w szczególności uczenia głębokiego, a także zaawansowanej mechatroniki. Szybki rozwój obserwuje się również w zakresie tworzenia coraz ściślejszego związku człowiek–maszyna[23].

W związku z tym pojawia się konieczność postawienia ważnych i trudnych pytań moralnych. Po pierwsze, należy zapytać o bezpieczeństwo, to znaczy – jak sprawić, aby świat, gdzie obok siebie żyją ludzie i inteligentne, autonomiczne maszyny, był bezpieczny? Po drugie, jaka jest odpowiedzialność moralna człowieka? W jaki sposób, komu i w jakim sensie przypisać odpowiedzialność moralną za nieprzewidziane skutki użycia technologii? Po trzecie, w jaki sposób sprawować kontrolę i zarządzać ekosystemem autonomicznych systemów? Jak przeprojektować instytucje i przepisy, aby służyły budowaniu dobrostanu, a jednocześnie zapewniały bezpieczeństwo? Po czwarte, jak zapewnić wolność samostanowienia?

[21] Tamże.
[22] Por. tamże, s. 25–29.
[23] Por. *Statement on Artificial Intelligence, Robotics and 'Autonomous' Systems*, Brussels 2018, s. 7.

Czy w dobie napędzanych sztuczną inteligencją social mediów i wyszukiwarek internetowych możemy jeszcze mówić o decyzjach wolnych, autonomicznych i świadomych? Po piąte, jak zagwarantować przejrzystość i wytłumaczalność?[24] Należy przyjrzeć się temu, jakim wartościom służą systemy SI. Czy są to te wartości, w które chcieliśmy je wyposażyć? Czy w ogóle wiemy, jakie wartości chcemy zaimplementować do SI? Czy wiemy, dookoła jakich wartości chcemy budować społeczeństwo, a jakie wartości poświęcimy na rzecz rozwoju technologii?[25]

Z perspektywy etycznej istotne według autorów dokumentu jest to, aby prawidłowo rozumieć trzy fundamentalne wartości: godność, autonomię i odpowiedzialność moralną człowieka. Autorzy stoją na stanowisku, że godność ludzka jest przyrodzona. Człowiek jest autonomiczny, zatem ma prawo do wyznaczania własnych standardów i wyboru własnych celów życiowych. Procesy poznawcze wspierające dokonywanie wyborów „należą do najściślej utożsamianych z godnością osoby ludzkiej oraz ludzką sprawczością i działaniem *par excellence*. Zwykle niosą za sobą cechy samoświadomości i sprawczości, opierając się na racji i wartościach. Autonomię w etycznie istotnym znaczeniu tego słowa można zatem przypisać jedynie istotom ludzkim"[26]. W tym miejscu można postawić pytania: czym procesy poznawcze tworzące reprezentacje w umyśle człowieka różnią się od tych tworzących reprezentacje w inteligentnym systemie? Jak procesy poznawcze SI łączą się z jego autonomią, a tym samym z odpowiedzialnością? Czym różni się autonomia człowieka od autonomii ogólnej SI? Czym różnią się człowiek i ogólna SI? W końcu warto postawić kluczowe dla antropologii filozoficznej pytanie: kim jest człowiek? Podważanie zasadności tego pytania jest bowiem jedną z przyczyn współczesnego kryzysu antropologicznego. Stanowi on zaś zagrożenie dla zdefiniowania fundamentalnych cech ludzkich, a także cech SI, które w pewnym zakresie mają być odwzorowaniem ludzkich. Brak definicyjnej precyzji skutkuje zamieszaniem pojęciowym, które stwarza istotny problem w aktualnej debacie dotyczącej SI.

Warto mieć na względzie, że autonomia przypisywana inteligentnym maszynom nie jest tożsama z autonomią ludzką. Autonomia SI polega bowiem na zdolności do wypracowania własnego wzorca decyzyjnego. Działanie SI wynika z aktualnego stanu algorytmów decyzyjnych, w tym z ilości danych posiadanych

[24] Por. tamże, s. 8.
[25] Por. tamże, s. 8–9.
[26] Tamże, s. 9.

przez system, ilości funkcji kodu oraz stopnia ich złożoności. Stan algorytmów decyzyjnych jest wynikiem akcji podejmowanych przez projektanta w procesie modyfikacji systemu oraz przez sam system w procesie uczenia się. Przejrzystość systemu jest zaś wprost proporcjonalna do procenta akcji podejmowanych przez projektanta vs. akcje podejmowane przez system. Im wyższy procent akcji podejmowanych przez system, tym poziom przejrzystości może stawać się niższy. Przez przejrzystość rozumiem nie tylko wsteczne odtworzenie kodu tak, aby dotrzeć do pierwotnej zasady lub normy zaimplementowanej do systemu przez projektanta, ale przede wszystkim możliwość zrozumienia, w jaki sposób w procesie uczenia się zaimplementowana zasada doprowadziła do konkretnej decyzji podjętej przez system. Héder wskazuje, że taki „test zrozumiałości" powinien zostać doprecyzowany o charakterystykę osoby, która miałaby zrozumieć system, oraz o przyjęcie określonego sposobu pomiaru[27].

Opierając się na przedstawionych wartościach, autorzy dokumentu wysuwają propozycję podstawowych zasad dotyczących projektowania SI. Pierwszą z nich jest zasada godności człowieka. Postuluje się wprowadzenie egzekwowanych prawnie wymogów informowania człowieka o tym, „czy i kiedy wchodzi w interakcję z maszyną", a także ograniczenia w przetwarzaniu i wykorzystywaniu informacji dotyczących istot ludzkich[28]. Tu można postawić pytanie o charakter tych ograniczeń, biorąc pod uwagę, że rozwój inteligentnego systemu zależy od ilości posiadanych przez niego danych.

Drugą zasadą jest zasada autonomii człowieka. Autonomia zakłada wolność; wolność przekłada się na odpowiedzialność, a ta z kolei – na powinność sprawowania kontroli nad działaniem inteligentnych systemów. SI nie może ograniczać swobody człowieka w określaniu własnych standardów i norm. Co więcej, SI musi umieć działać w zgodzie z tymi standardami, a nawet je wspierać. SI musi szanować ludzką zdolność i prawo do decydowania i samostanowienia, co oznacza, że inteligentne systemy mogą przedstawić użytkownikowi różnorodne treści i co najwyżej, zasugerować te, które według systemów są optymalne. Do sugestii powinno się dołączyć informację, na jakiej podstawie system uznał daną treść za optymalną. Do zilustrowania tej kwestii można posłużyć się działaniem wyszukiwarki internetowej. Systemy SI muszą zatem być przejrzyste i przewidywalne

[27] M. Héder, *A Criticism*…, dz. cyt., s. 71.

[28] Por. *Statement on Artificial Intelligence*…, dz. cyt., s. 16.

tak, aby każdy użytkownik był w stanie zatrzymać ich działanie, jeśli uzna to za moralnie konieczne[29].

Trzecią zasadą jest zasada odpowiedzialności. Możliwości SI powinny być wykorzystywane tylko w celach służących globalnemu dobru społecznemu i środowiskowemu. Należy przy tym pamiętać o potencjalnych negatywnych skutkach wykorzystania SI. Istotna według autorów dokumentu jest świadomość ryzyka oraz działanie charakteryzujące się wysokim stopniem ostrożności[30]. SI charakteryzuje się wysokim stopniem nieprzewidywalności, dlatego wydaje się, że postulowana świadomość ryzyka jest raczej życzeniem niż uwzględniającą realia praktyczną wskazówką.

Zasada czwarta głosi, że „sztuczna inteligencja powinna przyczyniać się do globalnej sprawiedliwości i równego dostępu do korzyści i możliwości, jakie mogą przynieść sztuczna inteligencja, robotyka i systemy autonomiczne". Należy zapobiegać dyskryminacji i stronniczości. Autorzy postulują stworzenie nowych modeli sprawiedliwej dystrybucji, które uwzględniałyby przemiany gospodarcze wywołane przez automatyzację, cyfryzację oraz sztuczną inteligencję. Zalecają także powszechny dostęp do informacji na temat nowych technologii oraz edukację cyfrową, dzięki którym będzie można zminimalizować wykluczenie, a także wspierać budowanie zaufania do SI[31].

Zasada piąta dotyczy demokracji. Autorzy są zdania, że kluczowe decyzje co do regulacji oraz rozwoju SI powinny być rezultatem demokratycznej debaty opartej na szczerym zaangażowaniu społecznym. „Zasady godności ludzkiej i autonomii obejmują prawo człowieka do samostanowienia za pomocą środków demokracji. Kluczowe znaczenie dla naszych demokratycznych systemów politycznych ma pluralizm wartości, różnorodność i pogodzenie różnych koncepcji dobrego życia obywateli"[32] – piszą autorzy. SI nie może w żaden sposób odbierać człowiekowi możliwości decydowania i wyrażania opinii, dokonując wyboru za niego, odbierając mu głos czy wywierając na niego jakikolwiek wpływ[33].

Zasada szósta odnosi się do praworządności, która warunkuje demokrację i ochronę praw człowieka. Należy opracować regulacje zapewniające człowiekowi ochronę przed zagrożeniami wynikającymi ze stosowania autonomicznych

[29] Por. tamże.
[30] Por. tamże, s. 16–17.
[31] Por. tamże, s. 17.
[32] Tamże.
[33] Por. tamże, s. 17–18.

systemów, prawo do zadośćuczynienia, a także do rzetelnego procesu sądowego. Trzeba precyzyjnie określić wymogi dotyczące projektowania i ścieżki odpowiedzialności dla naukowców, projektantów, programistów, inwestorów, sprzedawców, użytkowników i innych grup zaangażowanych w rozwój i wykorzystanie SI. Ponadto postuluje się stworzenie skutecznych systemów łagodzenia szkód[34].

Zasada siódma dotyczy bezpieczeństwa oraz integralności cielesnej i psychicznej. Można ją rozpatrywać w trzech wymiarach: bezpieczeństwa dla środowiska i użytkowników, niezawodności i wewnętrznej odporności systemu na ataki zewnętrzne oraz bezpieczeństwa emocjonalnego jednostek, w szczególności, gdy maszyna ma ewidentne cechy humanoidalne[35]. Z tej zasady wynika jednocześnie zasada ósma – konieczność ochrony danych i prywatności. Autorzy dokumentu odwołują się tu do prawa w zakresie ochrony danych, ale zwracają także uwagę na prawo do prywatności, rozumiane jako prawo do bycia wolnym od technologii wpływających między innymi na rozwój osobisty, opinie czy relacje. „W świetle obaw dotyczących wpływu systemów «autonomicznych» na życie prywatne i prywatność, można zwrócić uwagę na trwającą debatę na temat wprowadzenia dwóch nowych praw: prawa do konstruktywnego kontaktu z ludźmi oraz prawa do niebycia profilowanym, do bycia wolnym od mierzenia, analizowania, trenowania lub wywierania wpływu"[36] – czytamy w dokumencie.

Dziewiąta i ostatnia z zasad zaproponowanych przez Europejską Grupę do spraw Etyki w Nauce i Nowych Technologiach dotyczy zrównoważonego rozwoju. Rozwój SI musi bowiem współgrać z ludzkim obowiązkiem dbania o środowisko w celu prawidłowego funkcjonowania ludzkości oraz zachowania optymalnych warunków życia dla przyszłych pokoleń. Należy pracować nad stworzeniem strategii zapobiegających szkodliwemu wpływowi technologii na istoty ludzkie oraz przyrodę tak, aby zapewnić priorytet ochrony środowiska oraz zrównoważonego rozwoju[37].

Dokument kończy się wezwaniem do podjęcia przez Komisję Europejską kroków prowadzących do stworzenia globalnych norm etycznych i prawnych w kwestii SI i robotyki[38].

---

[34] Tamże, s. 18.
[35] Por. tamże, s. 18–19.
[36] Tamże, s. 19.
[37] Por. tamże.
[38] Tamże, s. 20.

5. Zasady etyczne w świecie sztucznej inteligencji

W celu wypracowania fachowych ekspertyz dotyczących nowych technologii Komisja Europejska powołała trzy grupy eksperckie: Niezależną Grupę Ekspertów Wysokiego Szczebla do spraw Sztucznej Inteligencji, Niezależną Grupę Ekspertów Wysokiego Szczebla do spraw Wpływu Transformacji Cyfrowej na Rynki Pracy UE oraz Grupę Ekspertów do spraw Odpowiedzialności i Nowych Technologii. Kwestiami etyki SI zajęła się pierwsza z nich.

Niezależna Grupa Ekspertów Wysokiego Szczebla do spraw Sztucznej Inteligencji (AI HLEG) została powołana w czerwcu 2018 roku. 52 ekspertów pracowało dwa lata (mandat europejski grupy wygasł w lipcu 2020 roku) zgodnie z europejskim mottem: „zjednoczeni w różnorodności"[39]. Przez ten czas dostarczyli cztery istotne dokumenty: *Wytyczne w zakresie etyki dotyczące godnej zaufania sztucznej inteligencji* (*Ethics Guidelines for Trustworthy AI*, 10 kwietnia 2019 roku), *Zalecenia dotyczące polityki i inwestycji w godną zaufania sztuczną inteligencję* (*Policy and Investment Recomendations for Trustworthy AI*, 26 czerwca 2019 roku), *Listę kontrolną dla godnej zaufania sztucznej inteligencji* (*Assessment List for Trustworthy AI*, 17 lipca 2020 roku) oraz *Rozważania sektorowe dotyczące polityki i rekomendacji inwestycyjnych* (*Sectoral Considerations on the Policy and Investment Recommendations*, 23 lipca 2020 roku). Posłużyły one za punkt wyjścia inicjatyw podejmowanych przez Komisję Europejską oraz państwa członkowskie.

Krytyce poddany został jednak skład grupy eksperckiej. Krytykowano między innymi zbyt duży, bo wynoszący prawie 50%, udział przedstawicieli przemysłu[40] oraz zbyt mały udział etyków[41]. W przypadku tych drugich faktycznie dysproporcja jest rażąca: 4 etyków i 48 nie-etyków. Thomas Metzinger, niemiecki filozof i członek AI HLEG, określił to jako próbę zbudowania najnowocześniejszego komputera AI z „48 filozofami, jednym hakerem i trzema informatykami", z czego „dwóch z nich jest zawsze na wakacjach"[42].

[39] *Wytyczne w zakresie etyki*…, dz. cyt., s. 5.

[40] Por. S. Larsson, *On the Governance of Artificial Intelligence through Ethics Guidelines*, „Asian Journal of Law and Society" 2020, t. 7, nr 3, s. 443.

[41] Por. T. Metzinger, *Ethics Washing Made in Europe*, https://www.tagesspiegel.de/politik/eu-guidelines-ethics-washing-made-in-europe/24195496.html (dostęp: 22.02.2022).

[42] Por. tamże.

Najważniejszym w kontekście etyki sztucznej inteligencji dokumentem dostarczonym przez AI HLEG są *Wytyczne w zakresie etyki dotyczące godnej zaufania sztucznej inteligencji*. Dokument powstał między innymi na podstawie zasad etycznych zaproponowanych – przedstawionych w poprzedniej części – przez Europejską Grupę do spraw Etyki w Nauce i Nowych Technologiach. AI HLEG uznała te zasady, dopracowała je i przedstawiła w wytycznych. Dodatkowo eksperci opierają się na Karcie Praw Podstawowych, traktatach UE, konwencjach ONZ dotyczących praw człowieka, konwencjach Rady Europy, między innymi Konwencji o ochronie praw człowieka i podstawowych wolności, oraz na przepisach ustawowych państw członkowskich.

Celem wytycznych jest „promowanie godnej zaufania sztucznej inteligencji"[43]. Aby dany system został uznany za godny zaufania, musi spełnić trzy warunki: być zgodny z prawem, etyczny i solidny. Zgodność z prawem oznacza poszanowanie wszystkich obowiązujących na danym gruncie przepisów. Etyczność ma gwarantować przestrzeganie wymogów, a w efekcie zgodność z zasadami etycznymi. Solidność, rozpatrywana zarówno z technicznego, jak i społecznego punktu widzenia, zapewnia natomiast, że maksymalnie zminimalizowano prawdopodobieństwo wystąpienia szkód, nawet tam, gdzie zastosowanie SI wydaje się ewidentnie korzystne[44]. Etyczność i solidność są ze sobą ściśle powiązane. Według autorów tylko etycznie zaprojektowana i rozwijana SI może bowiem gwarantować solidność[45].

Określenie SI jako godnej zaufania zostało skrytykowane. Co ciekawe, krytyka została wysunięta przez jednego z autorów wytycznych, wspomnianego już Thomasa Metzingera, według którego to nie maszyny są godne zaufania, tylko ludzie, a nazwanie sztucznej inteligencji jako godnej zaufania wynika z korzyści takiej retoryki w odniesieniu do strategii inwestycyjnych[46].

W wytycznych omówiono drugą cechę sztucznej inteligencji – etyczność. Treść podzielono na trzy części i uporządkowano od najwyższego do najniższego poziomu abstrakcji. Część pierwsza określa podstawy; przedstawia prawa podstawowe jako fundament godnej zaufania sztucznej inteligencji, a zarazem punkt wyjścia czterech zasad etycznych. Część druga dotyczy wdrażania; prezentuje siedem wymogów, których należy przestrzegać w procesie tworzenia inteli-

[43] Por. *Wytyczne w zakresie etyki*…, dz. cyt., s. 2.
[44] Por. tamże, s. 6.
[45] Por. tamże, s. 9.
[46] Por. T. Metzinger, *Ethics Washing Made in Europe*, dz. cyt.

gentnych systemów. Część trzecia dotyczy oceny. Znajdziemy tu konkretną listę kontrolną służącą do oceny, czy dany system spełnia wymogi godnej zaufania sztucznej inteligencji.

Treść wytycznych, w szczególności zaproponowane zasady etyczne, została jednak skrytykowana jako zbyt ogólna w rozumieniu braku specyficznego odniesienia do SI[47]. Krytycy wskazują, że większość zasad wystosowanych względem SI z powodzeniem można odnieść do innych obszarów inżynieryjnych. Héder proponuje zastosowanie metody, którą nazywa testem bojlera wodnego. Metoda polega na zamianie słów „sztuczna inteligencja" (SI) na „bojler wodny" i zweryfikowaniu, czy zapisy nadal mają sens. Okazuje się, że treść wytycznych nie zdaje testu, ponieważ zawarte w treści zdania pozostają tak samo istotne w kontekście bojlerów jak w wypadku SI[48]. Można jednak wyselekcjonować właściwości specyficzne tylko dla SI, a w efekcie zasady etyczne, które powinny być przestrzegane. Taką właściwością jest autonomia, a zasadami – przejrzystość, wytłumaczalność i nadzór ludzki[49].

Inne zarzuty dotyczą braku klarowności w kwestii tego, do kogo lub czego odnoszą się wytyczne – do deweloperów czy do sztucznych systemów? Jednym z przykładów jest nakaz unikania stronniczości, który w jednych zapisach tyczy się działania systemu, w innych pracy dewelopera, w jeszcze innych obydwu jednocześnie, a czasem zbyt trudno jednoznacznie to określić[50]. Nie wiadomo również, dlaczego wiele z zapisów wytycznych, szczególnie tych dotyczących zagrożeń wynikających z użycia technologii, sugeruje odniesienie do ogólnej SI, podczas gdy definicja wskazuje na SI w ujęciu słabym[51].

Celem godnej zaufania, czyli ukierunkowanej na człowieka, sztucznej inteligencji jest rozwijanie jego potencjału oraz poprawa dobrostanu jednostek i społeczeństwa. W rezultacie SI ma przyczynić się do dobra ogółu, między innymi poprzez wspieranie innowacyjności i postępu naukowego. Unia liczy zwłaszcza na udział SI w usprawnieniu celów zrównoważonego rozwoju wyznaczonych przez ONZ: promowania równowagi płci, przeciwdziałania zmianie klimatu, racjonalizacji sposobu korzystania z zasobów naturalnych, działań podejmowanych w ramach dążenia do poprawy stanu zdrowia obywateli, mobilności i procesów

47 Por. M. Héder, *A Criticism*…, dz. cyt., s. 71.
48 Tamże, s. 66.
49 Tamże, s. 69–70.
50 Tamże, s. 63.
51 *Definicja SI: główne funkcje i dyscypliny naukowe*, Bruksela 2019, s. 6.

produkcji. Liczy się również, że SI wspomoże sposoby monitorowania postępów, opierając się na wskaźnikach zrównoważonego charakteru i spójności społecznej[52].

Punktem wyjścia rozumowania zaproponowanego w wytycznych są wartości i prawa uznane przez Unię Europejską. Podejście do etyki SI oparto na prawach podstawowych ujętych art. 2 i 3 Traktatu o Unii Europejskiej, Karcie Praw Podstawowych Unii Europejskiej oraz międzynarodowym prawie dotyczącym praw człowieka[53]. Karta Praw Podstawowych wymienia cztery wartości: godność, wolność, równość i solidarność, a także dwie zasady: demokracji i państwa prawnego, w ramach których wartości są realizowane. Traktat z Lizbony nazywa je wartościami: godności, wolności, demokracji, równości, praworządności oraz praw człowieka. Oparte na wartościach prawa podstawowe ustanowiono fundamentem „abstrakcyjnych zasad etycznych i wartości, które można wdrożyć w kontekście SI"[54]. „Wspólna podstawa" łącząca te prawa jest według autorów dokumentu „zakorzeniona w poszanowaniu godności ludzkiej" i odzwierciedla to, „co określamy mianem «podejścia ukierunkowanego na człowieka», w którym człowiek cieszy się wyjątkowym i niezbywalnym moralnym statusem pierwszeństwa w wymiarze cywilnym, politycznym, gospodarczym i społecznym"[55]. Można zatem powiedzieć, że wartość godności człowieka jest fundamentem praw, które są z kolei fundamentem wartości i zasad możliwych do wdrożenia w ramach SI. Na podstawie wartości i zasad określa się konkretne wymogi względem systemu SI. Na podstawie zgodności architektury i działania systemu SI z wymogami możliwe ma być stwierdzenie, czy dany system jest etyczny, czy też nie.

Godność ujęta jest w wytycznych jako „wrodzona wartość", która „nie powinna być ograniczana, naruszana lub tłumiona […]. W kontekście SI poszanowanie godności ludzkiej oznacza, że wszyscy ludzie są traktowani z szacunkiem, jaki im się należy jako podmiotom moralnym, a nie jako zaledwie przedmioty, które mają być przesiewane, sortowane, oceniane, gromadzone, warunkowane lub manipulowane"[56]. Godność człowieka i powiązane z nią prawa podstawowe i zasady „mają charakter bezwzględny i nie mogą być przedmiotem wyważa-

[52] Por. *Wytyczne w zakresie etyki*…, dz. cyt., s. 5.
[53] Por. tamże, s. 12.
[54] Tamże.
[55] Tamże.
[56] Tamże, s. 13.

nia racji"[57]. Systemy SI należy zatem projektować tak, aby szanowały psychiczną i fizyczną integralność każdego człowieka, jego tożsamość osobową i kulturową, a także gwarantowały zaspokojenie podstawowych ludzkich potrzeb[58].

Wolność rozumiana jest jako „możliwość samodzielnego podejmowania życiowych decyzji", co oznacza zobowiązanie, leżące między innymi po stronie organów unijnych i państwowych, do przekazywania jednostce coraz większej kontroli nad własnym życiem. SI powinna tę wolność wzmacniać na przykład poprzez dostarczanie rzetelnych danych czy analiz wspierających autonomiczne procesy decyzyjne. Nie może wolności odbierać przez wprowadzanie w błąd czy nieuczciwą manipulację. Człowiek musi zachować niezależność i pełną świadomość działania w zakresie pozyskiwania informacji, kreowania poglądów oraz wolności wypowiedzi. W tym celu SI musi szanować prywatność jednostki, w tym prawo do życia prywatnego i wolności od technologii[59].

Systemy SI powinny wspierać procesy demokratyczne, praworządność i sprawiedliwość między innymi poprzez poszanowanie pluralizmu wartości oraz obowiązujących przepisów ustawowych i wykonawczych. Nie mogą podważać równości jednostek względem prawa. Nie mogą celowo manipulować informacjami, wpływając tym samym na decyzje jednostek i potencjalnie także na wyniki głosowań w demokratycznych wyborach[60].

Systemy SI powinny też wspierać równość, solidarność i niedyskryminację. Nie można dopuścić do dyskryminacji wynikającej ze stronniczości systemu. Dopuszcza się jedynie możliwość poszukiwania wzorców na podstawie obiektywnych przesłanek, jednak w sytuacji, gdy rezultat nosi znamiona stronniczości, można wnioskować, że algorytm został napisany błędnie. Należy dbać o to, aby dane wprowadzane do samouczącego się systemu były jak najbardziej reprezentacyjne, przekrojowe, rzetelne i aby jak najprecyzyjniej odzwierciedlały stan faktyczny. Trzeba zwrócić szczególną uwagę na grupy znajdujące się w wyjątkowej sytuacji, jak na przykład osoby z niepełnosprawnościami, dzieci, osoby starsze, kobiety, mniejszości i inne[61].

Ostatnią z wartości, którą według ekspertów powinna wspierać sztuczna inteligencja, są prawa obywateli. SI stosowana w sektorze publicznym, w instytucjach

[57] Tamże, s. 19.
[58] Tamże, s. 13.
[59] Tamże.
[60] Tamże.
[61] Por. tamże.

rządowych czy samorządowych może mieć bowiem negatywny wpływ na obywatelskie wolności. Należy zatem kontrolować prawidłowość działania inteligentnych systemów w tych obszarach i objąć prawa obywateli większą ochroną[62].

Opierając się na wskazanych wartościach i ich rozumieniu, eksperci AI HLEG zaproponowali cztery zasady etyczne, których twórcy SI powinni przestrzegać. Wymieniane są one w kolejności, w jakiej prawa, których są odzwierciedleniem, występują w Karcie Praw Podstawowych. Te zasady to: poszanowanie autonomii człowieka, zapobieganie szkodom, sprawiedliwość oraz możliwość wyjaśnienia. Sformułowane zostały w formie etycznych imperatywów, aby zachęcić wszystkie strony zaangażowane w budowanie ekosystemu inteligentnych maszyn do ich przestrzegania[63].

Owo zachęcanie do przestrzegania norm zostało tymczasem skrytykowane jako słabe w rozumieniu braku procedur egzekwujących[64]. Mechanizmem urealniającym przedstawione zasady miałaby stać się samoregulacja w odniesieniu do twórców SI oraz do uczących się systemów SI, co może spowodować, że etyka będzie wykorzystywana jako narzędzie prowadzące do wygenerowania społecznego zaufania względem technologii oraz do korzyści ekonomicznych, nie mając jednocześnie wiele wspólnego z uczciwym stosowaniem się do zasad etycznych w procesie projektowania[65]. Na inny aspekt tej kwestii zwraca uwagę Héder, który pisze:

> jednym z obszarów rozwoju sztucznej inteligencji jest ustalenie, jakie decyzje należy delegować do systemów autonomicznych i jak uzyskać najlepsze wyniki. Dlatego wydaje się, że wymagalibyśmy, aby artefakty wytwarzane przez programistów AI (systemy autonomiczne) były „zgodne etycznie" i aby programiści odkryli, jak to zapewnić, jednocześnie nakazując, aby sami byli „zgodni etycznie". Ta podwójna pośredniość oznacza, że wszystko, co mogą zrobić AIGU, to pokierować programistami w „kierowaniu" systemami AI. To powoduje, że AIGU mają tendencję do abstrakcji do punktu nieefektywności[66].

[62] Por. tamże, s. 13–14.
[63] Por. tamże, s. 14.
[64] Por. T. Hagendorff, *The Ethics of AI Ethics*, dz. cyt., s. 90.
[65] Por. S. Larsson, *On the Governance*…, dz. cyt., s. 442.
[66] M. Héder, *A Criticism*…, dz. cyt., s. 71.

Pierwsza z zasad, dotycząca poszanowania autonomii człowieka, związana jest z prawami do godności i wolności ujętymi w art. 1 i 6 Karty Praw Podstawowych[67]. Nakazuje, by to człowiek, a nie SI, zachował kontrolę nad całością procesów zachodzących w inteligentnym systemie. Tylko w ten sposób człowiek utrzyma także pełną zdolność do świadomego dokonywania wyborów, samostanowienia, sprawczość oraz poczucie realnego wpływu na kształtowanie rzeczywistości, w której żyje. Inteligentne systemy nie mogą „bezpodstawnie podporządkowywać, przymuszać, oszukiwać, kształtować lub kontrolować ludzi ani nimi manipulować"[68]. Nadzór ludzki wymaga więc balansowania między różnymi znaczeniami pojęcia „kontrola". Wydaje się, że różnice te nie zostały tu jednak uchwycone. Jak pisze Héder:

> Potrzebujemy sztucznej inteligencji, aby autonomicznie sprawowała kontrolę nad sytuacją, w której się znajduje, oczekując jednocześnie, że zachowamy kontrolę w tym sensie, że potrzebujemy sztucznej inteligencji, aby uniknąć niepożądanych konsekwencji – których nie możemy z góry wyliczyć, ponieważ wiele z nich jest nieprzewidywalnych[69].

Zasada zapobiegania szkodom związana jest z ochroną integralności fizycznej lub psychicznej ujętą w art. 3 Karty Praw Podstawowych[70]. Zasada nakazuje, aby systemy SI nie powodowały ani w żaden sposób nie przyczyniały się do zwiększania negatywnych efektów indywidualnego lub zbiorowego działania występujących w środowisku społecznym, kulturowym czy politycznym. SI nie może być podatna na wykorzystywanie w złym zamiarze, nie może powodować lub pogłębiać niekorzystnego wpływu wywołanego przez asymetrię władzy czy informacji. Musi uwzględniać w swoich działaniach środowisko naturalne i wszystkie żyjące istoty, ze szczególnym uwzględnieniem grup wrażliwych, zagrożonych wykluczeniem czy wykluczonych. Należy projektować i produkować systemy tak, aby były solidne pod względem technicznym, bezpieczne i pewne[71].

Proces projektowania i produkcji musi również opierać się na zasadzie sprawiedliwości. Zasada ta związana jest z prawem do niedyskryminacji, solidarno-

[67] Por. *Wytyczne w zakresie etyki*…, dz. cyt., s. 14.
[68] Tamże, s. 15.
[69] M. Héder, *A Criticism*…, dz. cyt., s. 63.
[70] Por. *Wytyczne w zakresie etyki*…, dz. cyt., s. 14.
[71] Por. tamże, s. 15.

ści i sprawiedliwości ujętym w art. 21 i następnych Karty Praw Podstawowych[72]. Istotne jest, aby pojęcie sprawiedliwości było postrzegane jednolicie. Rozumienie podzielane przez ekspertów zakłada, że sprawiedliwość ma wymiar materialny oraz proceduralny. Wymiar materialny związany jest z równym i sprawiedliwym podziałem korzyści oraz kosztów. Gwarantuje brak stronniczości, dyskryminacji oraz stygmatyzacji. Wymiar proceduralny zapewnia natomiast możliwość kwestionowania decyzji SI oraz skuteczne dochodzenie roszczeń w sytuacji, gdy decyzje okażą się krzywdzące dla człowieka. W kontekście SI należy zadbać o przejrzystość, wytłumaczalność i precyzyjne przypisanie odpowiedzialności tak, aby możliwe było zrozumienie oraz uzasadnienie decyzji podjętych przez system, a także uzyskanie zadośćuczynienia[73].

Ostatnią zasadą zaproponowaną przez AI HLEG jest zasada możliwości wyjaśnienia, związana z prawami odnoszącymi się do sprawiedliwości ujętymi w art. 47 Karty Praw Podstawowych[74]. Możliwość wyjaśnienia oznacza, że każda osoba, dla której jakakolwiek decyzja SI będzie niezrozumiała, błędna czy krzywdząca, będzie mogła prześledzić kroki, które do niej doprowadziły. Pozwoli to ocenić, czy decyzja jest wynikiem niezamierzonego błędu, zaniechania czy intencjonalnie niemoralnego działania. Twórcy muszą zatem zagwarantować, że procesy SI będą przejrzyste i w jak największym stopniu zrozumiałe dla wszystkich, na których SI ma bezpośredni czy pośredni wpływ. Cele i procesy zachodzące w SI muszą być otwarcie komunikowane i wytłumaczalne, a jej decyzje możliwe do prześledzenia i uzasadnienia[75]. Przejrzystość musi iść jednak w parze z wytłumaczalnością w znaczeniu zrozumiałości.

Na podstawie czterech opisanych zasad sformułowano siedem konkretnych wymogów, które mają być praktycznym wsparciem w tworzeniu i wdrażaniu godnej zaufania sztucznej inteligencji. Są to: przewodnia i nadzorcza rola człowieka; techniczna solidność i bezpieczeństwo; ochrona prywatności i zarządzanie danymi; przejrzystość; różnorodność, niedyskryminacja i sprawiedliwość; dobrostan społeczny i środowiskowy oraz odpowiedzialność. Według autorów dokumentu wszystkie wymogi są jednakowo ważne i wzajemnie się uzupełniają. Powinny ponadto obowiązywać w całym cyklu życia systemu SI. W sytuacji konfliktu pomiędzy nimi należy przeanalizować i ocenić, który z wymogów

[72] Por. tamże, s. 14.
[73] Por. tamże, s. 15.
[74] Por. tamże, s. 14.
[75] Por. tamże.

lepiej przyczynia się do realizacji zasad etycznych i nadać mu wyższy priorytet. Przejrzystość, wytłumaczalność i rozliczalność systemów SI są warunkiem niezbędnym do ustalenia hierarchii wartości, a tym samym do zaproponowania etycznego rozwiązania. Jeśli określenie dopuszczalnych pod względem etycznym kompromisów jest niemożliwe, oznacza to, że kształt systemu jest nieprawidłowy. Należy go zmienić lub całkowicie zatrzymać jego działanie. Dodatkowo podmiot odpowiedzialny za skutki działania SI powinien zapewnić skuteczną ścieżkę dochodzenia roszczeń dla wszystkich poszkodowanych. Wiedza co do sposobu dochodzenia roszczeń powinna być ogólnodostępna, jako że ma to kluczowe znaczenie dla budowania zaufania do sztucznej inteligencji[76].

Komisja 19 lutego 2020 roku opublikowała *Białą Księgę w sprawie sztucznej inteligencji*. Treść obejmuje dwa główne zagadnienia: „ramy polityczne określające środki służące połączeniu wysiłków na szczeblu europejskim, krajowym i regionalnym"[77], które mają przysłużyć się stworzeniu „ekosystemu doskonałości", oraz „kluczowe elementy przyszłych ram regulacyjnych dotyczących sztucznej inteligencji w Europie"[78], mające pomóc w stworzeniu „ekosystemu zaufania". Od 19 lutego 2020 do 14 czerwca 2020 roku przeprowadzone zostały konsultacje publiczne. Komisja Europejska zebrała ponad 1250 opinii na tematy przedstawione w księdze. Z wynikami konsultacji można się zapoznać w raporcie *Public Consultation on the AI White Paper. Final Report*, opublikowanym przez Komisję Europejską w listopadzie 2020 roku.

Członkowie Komisji przyjęli wszystkie siedem wymogów etycznych zaproponowanych przez AI HLEG. Niestety wymogi te nie mają prawnie wiążącego charakteru. Ta dobrowolność w połączeniu z potężnymi inwestycjami i jeszcze większymi potencjalnymi zyskami ze sprzedaży zaawansowanej SI może tymczasem skutkować niskim wskaźnikiem stosowania się do wymogów, a w rezultacie powodować zagrożenia dla człowieka. Komisja postuluje zatem stworzenie wiążących ram regulacyjnych uwzględniających kwestie etyczne. Ramy regulacyjne miałyby koncentrować się na sposobach zminimalizowania zagrożeń, a w efekcie szkód wynikających z działania SI[79].

[76] Por. tamże, s. 24–25.

[77] *Biała Księga w sprawie sztucznej inteligencji. Europejskie podejście do doskonałości i zaufania*, Bruksela 2020, s. 3.

[78] Tamże.

[79] Por. tamże, s. 12.

6. Definicja pojęcia „godność człowieka" przyjęta w wytycznych i jej filozoficzne podstawy

Z uwagi na to, że wartość godności człowieka została ujęta w wytycznych jako fundament, w którym zakorzeniona jest podstawa praw człowieka, poszanowanie godności „odzwierciedla to, co nazywamy mianem podejścia ukierunkowanego na człowieka", a „w podejściu ukierunkowanym na człowieka człowiek cieszy się wyjątkowym i niezbywalnym moralnym statusem pierwszeństwa w wymiarze cywilnym, gospodarczym i społecznym"[80], warto przyjrzeć się, jaką definicję pojęcia „godność" proponują członkowie AI HLEG, a co za tym idzie – jaką definicję uznaje Unia Europejska.

Ujęcie pojęcia „godność człowieka" jako fundamentu praw człowieka sugeruje relację wynikania, która według wielu autorów jest nieuzasadniona. W sensie logicznym z „pojęcia nic nie wynika, ponieważ wynikanie jest (dedukcyjną) relacją między zdaniami (sądami)"[81]. Tak rozumiane wynikanie podważa również Eric Hilgendorf, na którego koncepcję godności powołują się autorzy wytycznych. W jaki sposób Hilgendorf uzasadnia zatem stwierdzenie, że godność człowieka jest fundamentem praw, a tym samym radzi sobie z trudnością wynikania? Na jakich konceptach filozoficznych opiera się filozof i czy będą one tymi, na których posadowiona zostanie także „normatywna wizja Europy"?

W tekście pod tytułem *Problem Areas in the Dignity Debate and the Ensemble Theory of Human Dignity*[82], do którego przekierowują autorzy wytycznych[83], Hilgendorf twierdzi, że w odniesieniu do pojęcia „godność ludzka" „nie istnieje żadne ustalone znaczenie, które moglibyśmy odkryć"[84]. Bezzasadne i prowadzące do „normatywnych konfliktów"[85] na gruncie prawnym próby zdefiniowania pojęcia na podstawie istoty rzeczy są podejściem błędnym. Znaczenie pojęcia nie jest w żaden sposób predeterminowane[86], a zatem pytanie „Czym jest godność

[80] Por. *Wytyczne w zakresie etyki*…, dz. cyt., s. 12.
[81] A. Bronk, *Kategoria godności człowieka*, „Annales UMCS" 2010, t. 35, nr 1, s. 81.
[82] E. Hilgendorf, *Problem Areas in the Dignity Debate and the Ensemble Theory of Human Dignity*, w: *Human Dignity in Context*, red. D. Grimm, A. Kemmerer, Ch. Mollers, Baden-Baden 2018.
[83] Por. *Wytyczne w zakresie etyki*…, dz. cyt., s. 13.
[84] E. Hilgendorf, *Problem Areas*…, dz. cyt., s. 329.
[85] Tamże, s. 328.
[86] Por. tamże, s. 329.

człowieka?" jest błędne. Należy skonstruować znaczenie pojęcia. Filozof proponuje wyjść od ustalenia kontekstu, w jakim używany jest dany termin.

Stwierdza, że pojawienie się w dokumentach założycielskich pojęcia godności ludzkiej było reakcją „na brutalne zbrodnie przeciwko ludzkości i ekstremalne naruszenie ludzkich interesów" i „zgodnie z tym rozumieniem godność ludzka dostarcza normatywną kotwicę w obliczu wciąż rozszerzającego się na świecie pluralizmu"[87]. Pojęcie godności człowieka ma więc posłużyć jako „normatywna kotwica, którą akceptują wszyscy lub prawie wszyscy członkowie społeczeństwa, niezależnie od ich pochodzenia kulturowego i bez względu na ich konkretne systemy wartości"[88]. Pytanie, które powinniśmy postawić, nie brzmi zatem „Czym jest godność człowieka?", ale „Jak zdefiniować koncept godności człowieka w najbardziej efektywny sposób, aby osiągnąć cel, a mianowicie stworzyć normatywną podstawę akceptowalną dla wszystkich lub prawie wszystkich ludzi w naszym pluralistycznym społeczeństwie?"[89], aby w efekcie zapewnić ludziom ochronę przed ekstremalnym naruszaniem ich interesów. Według Hilgendorfa taką ochronę można uzyskać przez zagwarantowanie jednostce praw „absolutnych", to znaczy takich, które muszą być bezwzględnie przestrzegane niezależnie od okoliczności[90].

Filozof sugeruje, aby zdanie „X posiada godność człowieka" zrównać ze zdaniem „X posiada pewne prawa". Prawa, jakie proponuje, są odzwierciedleniem obserwowalnych w społeczeństwie „podstawowych ludzkich potrzeb", które mogą być uznane za uniwersalne mimo różnic kulturowych czy przedziału czasowego. Potrzeby te nazywa „minimalnym prawem naturalnym"[91]. Prawa zaś wynikające z potrzeb to: prawo do materialnego minimum egzystencji, prawo do bycia wolnym od ekstremalnego bólu – zakaz tortur, prawo do podstawowej integralności psychicznej – zakaz „prania mózgu", prawo do minimum autonomicznego rozwoju indywidualnego, prawo do kontrolowania najbardziej intymnych danych dotyczących naszej osoby, prawo do równości wobec prawa – zakaz niewolnictwa, a także prawo do minimum szacunku. Naruszenie któregokolwiek z tych praw zostanie uznane za naruszenie godności człowieka, jako że prawa

87 Tamże, s. 328.
88 Tamże, s. 329.
89 Tamże, s. 331.
90 Por. tamże, s. 332.
91 Por. tamże, s. 336–337.

i godność są tożsame[92]. W ten sposób, jak twierdzi Hilgendorf, przezwyciężony zostaje problem wynikania.

Wyjściowo Hilgendorf określa zatem cel użycia terminu „godność". Opierając się na tym celu, pyta: jakie powinno być znaczenie pojęcia „godność człowieka", aby mogło ono przełożyć się na zapisy blokujące możliwość zbrodniczych działań analogicznych do podejmowanych podczas II wojny światowej? Stwierdza, że zbrodnicze działania naruszały podstawowe potrzeby człowieka, a więc należy te potrzeby zidentyfikować i zabronić ich naruszania. Potrzeby człowieka, ustalone na drodze empirycznej, przekłada na zdania; zdania dotyczące potrzeb przekłada na zdania tyczące się praw człowieka. Następnie utożsamia je z wartością godności człowieka, która stanowi wartość centralną, jako że prawo do poszanowania godności „nie może zostać naruszone, nawet jeżeli oznaczałoby to ograniczenie innego prawa"[93]. W ten sposób nie odkrywa, a tworzy znaczenie pojęcia „godność człowieka" zgodnie z ustalonym przez siebie celem. Filozof zaprzecza esencjalizmowi, opowiada się natomiast za podejściem konstruktywistycznym.

Konstruktywizm to podejście, z którym związane są takie nurty myślowe, jak: antyobiektywizm, antyesencjalizm, antydualizm, antyfundamentalizm poznawczy, a także relatywizm społeczny i historyczny, zaprzeczając przy tym uniwersalizmowi klasycznej definicji prawdy[94]. Konstruktywiści odrzucają tezę, że wiedza odzwierciedla obiektywną naturę rzeczywistości. Uznają, że jest ona konstruowana na gruncie relacji społecznych. Wiedza nie jest odkrywana, ale wytwarzana[95].

W przeprowadzonej analizie konstruktywizmu w odniesieniu do idei i wartości europejskich Tomasz Grosse stwierdza, że głównymi funkcjami metody konstruktywistycznej są: konstruowanie, maskowanie oraz legitymizacja rzeczywistości społecznej i politycznej. Konstruowanie odbywa się „zawsze w odniesieniu do określonych interesów społecznych lub politycznych". Przykładem takich konstrukcji są próby „formowania nowego Europejczyka oraz społeczeństwa europejskiego". Maskowanie polega na wykorzystywaniu skonstruowanych

[92] Por. tamże, s. 332.

[93] *Wyjaśnienia dotyczące Karty Praw Podstawowych (2007/C 303/02)*, Dziennik Urzędowy Unii Europejskiej 2007, https://eur-lex.europa.eu/ (dostęp: 10.09.2021).

[94] Por. J. Kostyszak, *Konstruktywizm jako metodologia badania współczesności*, „Zeszyty Naukowe Polskiego Towarzystwa Ekonomicznego w Zielonej Górze" 2019, nr 10, s. 110–111.

[95] Por. M. Wendland, *Perspektywa konstruktywistyczna jako filozoficzna podstawa rozważań nad komunikacją*, „Kultura i Edukacja" 2011, t. 83, nr 4, s. 32–33.

wartości i idei w celu „skrywania" realiów świata społecznego czy politycznego. Legitymizowanie natomiast jest „próbą zdobycia społecznego uznania" dla instytucji[96]. Jak pisze Grosse: „postrzeganie interesów poszczególnych grup może być rekonstruowane i modyfikowane. Wymaga to jedynie odpowiedniej interpretacji i odniesienia do nadrzędnych wartości lub idei"[97]. Czemu ma zatem służyć utrzymywanie w unijnej retoryce pojęcia „godność człowieka", mimo iż, jak zauważa Łuków: „odwoływanie się do godności ludzkiej nie wydaje się konieczne[98]; może wystarczyć międzynarodowy konsensus dotyczący praw człowieka"?[99] Czy posługiwanie się godnością jest działaniem służącym faktycznej ochronie człowieka czy raczej zdobyciu społecznej akceptacji wobec komercyjnej wizji sztucznej inteligencji?

Jak pisze Grosse:

> Dochodzimy w ten sposób do zasadniczego problemu metody konstruktywistycznej, a mianowicie tego, że angażuje ona określone grupy społeczne i ich interesy, a nawet intencjonalnie tworzy nowe instytucje i skupia wokół nich grupy interesów, aby aktywnie promowały i rozwijały określoną wizję […]. Dlatego tego typu metoda działania jest określana jako „strategiczny konstruktywizm". Idee i wartości torują drogę dla zmian instytucjonalnych mających w rezultacie przynieść korzyści strategiczne dla najbardziej uprzywilejowanych podmiotów, zarówno w wymiarze politycznym […], jak i ekonomicznym[100].

Czy znaczenie wartości powinno być konstruowane w taki sposób, aby odpowiadało na potrzeby ludzi, czyli na to, co w danym kontekście i przedziale czasowym ludzie uważają za wartościowe? Przywołując rozważania van de Poela w kontekście sztucznej inteligencji, którą można zaprojektować ku określonym wartościom, jednym z najważniejszych wyzwań staje się ustalenie, które z wartości są tymi, które powinny być cenione ze względu na przyczyny normatywne. Wyzwaniem jest także znalezienie sposobu zapobieżenia konstruowaniu znacze-

96 Por. T.G. Grosse, *Trzy oblicza konstruktywizmu w Europie: rozważania o kryzysie metody integracyjnej*, „Chrześcijaństwo – Świat – Polityka" 2014/2015, nr 17–18, s. 37–38.

97 Tamże, s. 38.

98 W kontekście dokumentów założycielskich i dokumentów odnoszących się do nich.

99 P. Łuków, *A Difficult Legacy: Human Dignity as the Founding Value of Human Rights*, „Human Rights Review" 2018, t. 19, nr 3, s. 327.

100 T.G. Grosse, *Trzy oblicza konstruktywizmu w Europie*, dz. cyt., s. 42–43.

nia wartości tak, aby wspierały tylko cele komercyjnego rynku inteligentnych technologii. Pytania bardziej przekrojowe, jakie warto tu postawić, brzmią: czy możliwe jest zapobieżenie stosowania konstruktywizmu w odniesieniu do wartości, zwłaszcza tych, na których mają oprzeć się inteligentne systemy? Czy faktycznie chcemy odrzucić esencjalizm, a w efekcie naturę człowieka?

Jeśli Unia Europejska, podążając ścieżką konstruktywistyczną, decyduje o znaczeniu wartości, to znaczy definiuje wartości tak, aby uzasadniały realizację obranego przez nią celu, to czy możemy być pewni, że definicja człowieka, czyli zapis określający, kim jest każdy z nas, nie zostanie skonstruowany w podobny sposób? Co w niedalekiej przyszłości Unia nazwie godnością człowieka, a co człowiekiem i jak przełoży się to na inteligentne systemy, w szczególności na dopuszczone do użytku funkcjonalności? Już teraz użytkowanie funkcjonalności, takich jak na przykład systemy rekomendacji treści, wpływa destrukcyjnie na ludzką autonomię, mimo że pozornie wydaje się odwrotnie[101].

Pilne i ważne zdaje się przeprowadzenie pogłębionych analiz dotyczących rozwoju sztucznej inteligencji opartej na filozoficznych koncepcjach przyjętych przez Unię Europejską. Badania powinny w szczególności być przeprowadzane przez zespoły etyczno-technologiczne na gruncie etyki stosowanej. Jak pisze Metzinger, nadszedł czas,

> aby uniwersytety i społeczeństwo obywatelskie […] wyjęły samoorganizującą się dyskusję z rąk przemysłu. Wszyscy to czują: znajdujemy się w szybkim przejściu historycznym, które odbywa się na wielu poziomach jednocześnie. Okno możliwości, w ramach którego możemy przynajmniej częściowo kontrolować przyszłość sztucznej inteligencji i skutecznie bronić filozoficznych i etycznych fundamentów kultury europejskiej, za kilka lat się zamknie[102].

[101] Warto zapoznać się z artykułem M. Boets, *Autonomy and the Social Dilemma of Online Manipulative Behavior*, „AI Ethics" 2022, https://link.springer.com/article/10.1007/s43681-022-00157-5 (dostęp: 20.04.2022).

[102] T. Metzinger, *Ethics Washing Made in Europe*, dz. cyt.

Bibliografia

Literatura źródłowa

Biała Księga w sprawie sztucznej inteligencji. Europejskie podejście do doskonałości i zaufania, Komisja Europejska, Bruksela 2020.

Definicja SI: główne funkcje i dyscypliny naukowe, Niezależna Grupa Ekspertów Wysokiego Szczebla do spraw Sztucznej Inteligencji, Bruksela 2019.

Sprawozdanie zawierające zalecenia dla Komisji w sprawie przepisów prawa cywilnego dotyczących robotyki, Parlament Europejski, Bruksela 2017.

Statement on Artificial Intelligence, Robotics and 'Autonomous' Systems, European Group on Ethics in Science and New Technologies, Brussels 2018.

Trustworthy AI Assesment List, High-level expert group on Artificial Intelligence, Brussels 2020.

Wytyczne w zakresie etyki dotyczące godnej zaufania sztucznej inteligencji, Niezależna Grupa Ekspertów Wysokiego Szczebla do spraw Sztucznej Inteligencji, Bruksela 2019.

Literatura pomocnicza

Boets M., *Autonomy and the Social Dilemma of Online Manipulative Behavior*, „AI Ethics" 2022, https://doi.org/10.1007/s43681-022-00157-5.

Bronk A., *Kategoria godności człowieka*, „Annales UMCS" 2010, t. 35, nr 1, s. 77–96.

Europa na miarę ery cyfrowej: Komisja proponuje nowe przepisy i działania na rzecz doskonałości i wiarygodności sztucznej inteligencji, Komisja Europejska, Bruksela 2021.

Grosse T.G., *Trzy oblicza konstruktywizmu w Europie: rozważania o kryzysie metody integracyjnej*, „Chrześcijaństwo – Świat – Polityka" 2014/2015, nr 17–18, s. 35–50, https://doi.org/10.21697/csp.2014.18.1.03.

Hagendorff T., *The Ethics of AI Ethics: An Evaluation of Guidelines*, „Minds and Machines" 2020, t. 30, s. 99–120, https://doi.org/10.1007/s11023-020-09517-8.

Héder M., *A Criticism of AI Ethics Guidelines*, „Információs Társadalom" 2020, t. 20, nr 4, s. 57–73, https://doi.org/10.22503/inftars.XX.2020.4.5.

Hilgendorf E., *Problem Areas in the Dignity Debate and the Ensemble Theory of Human Dignity*, w: *Human Dignity in Context*, red. D. Grimm, A. Kemmerer, Ch. Mollers, Nomos, Baden-Baden 2018, s. 325–344, https://doi.org/10.5771/9783845264585-325.

Kostyszak J., *Konstruktywizm jako metodologia badania współczesności*, „Zeszyty Naukowe Polskiego Towarzystwa Ekonomicznego w Zielonej Górze" 2019, nr 10, s. 109–116, https://doi.org/10.26366/PTE.ZG.2019.152.

Kowalczyk S., *Filozoficzne koncepcje wartości*, „Collectanea Theologica" 1986, nr 1 (222), s. 37–51.

Krąpiec M., *Wartość*, w: *Powszechna Encyklopedia Filozofii*, http://www.ptta.pl/pef/.

Larsson S., *On the Governance of Artificial Intelligence through Ethics Guidelines*, „Asian Journal of Law and Society" 2020, t. 7, nr 3, s. 437–451, https://doi.org.10.1017/als.2020.19.

Leibert W., Schmidt J.C., *Collingridge's Dilemma and Technoscience*, „Poiesis Prax" 2010, t. 7, s. 55–71, https://doi.org/10.1007/s10202-010-0078-2.

Łuków P., *A Difficult Legacy: Human Dignity as the Founding Value of Human Rights*, „Human Rights Review" 2018, t. 19, nr 3, s. 313–327, https://doi.org/10.1007/s12142-018-0500-z.

Martinho A., Poulsen A., Kroesen M. i in., *Perspectives about Artificial Moral Agents*, „AI Ethics" 2021, nr 1, s. 477–490, https://doi.org/10.1007/s43681-021-00055-2.

Metzinger T., *Ethics Washing Made in Europe*, https://www.tagesspiegel.de/politik/eu-guidelines-ethics-washing-made-in-europe/24195496.html.

Poel I. van de, *Embedding Values in Artificial Intelligence (AI) Systems*, „Minds and Machines" 2020, t. 30, s. 385–409, https://doi.org/10.1007/s11023-020-09537-4.

Traktat z Lizbony, Parlament Europejski, https://www.europarl.europa.eu/.

Wendland M., *Perspektywa konstruktywistyczna jako filozoficzna podstawa rozważań nad komunikacją*, „Kultura i Edukacja" 2011, t. 83, nr 4, s. 30–47.

Wyjaśnienia dotyczące Karty Praw Podstawowych (2007/C 303/02), Dziennik Urzędowy Unii Europejskiej 2007, https://eur-lex.europa.eu/.

Żuk G., *Edukacja aksjologiczna. Zarys problematyki*, Wydawnictwo Uniwersytetu Marii Curie-Skłodowskiej, Lublin 2016.

Streszczenie

Przedmiotem artykułu jest zagadnienie etyki sztucznej inteligencji (SI) w kontekście najistotniejszych dokumentów Unii Europejskiej opublikowanych w latach 2017–2020. Celem tekstu jest zaprezentowanie treści odnoszących się do etyki, w tym wartości i zasad, jakich według organów unijnych należy przestrzegać

w procesach związanych z projektowaniem, tworzeniem i wdrażaniem produktów i usług opartych na inteligentnej technologii. Zasadnicze pytanie badawcze artykułu brzmi: w jaki sposób Unia Europejska ujmuje kwestię etyki sztucznej inteligencji? W poszukiwaniu odpowiedzi zadano pytania dodatkowe: jakie wartości są dla UE kluczowe do utrzymania? Jakimi zasadami powinni kierować się i jakie wymogi powinni spełniać twórcy inteligentnych maszyn? Jak rozumiana jest wartość godności człowieka, która w ujęciu Unii jest podstawą określania zasad i wartości w odniesieniu do SI? Jaką definicję pojęcia „godność człowieka" uznaje Unia Europejska i jakie koncepcje filozoficzne leżą u jej podstaw?

Słowa kluczowe: sztuczna inteligencja, SI, etyka, Unia Europejska, wytyczne

Summary

Ethics of Artificial Intelligence in European Union Documents in the Years 2017–2020

This article concerns the ethics of artificial intelligence (AI), within the scope and perspective held by the European Union, as expressed in documents published over the 2017–2020 period. The aim of the article is to present the content related to ethics, including the values and principles which, according to European Union authorities, should be followed in processes related to the design, development and implementation of products and services based on intelligent technology. The main research question of this article is: how does the European Union address the issue of the ethics of AI? In search of answers, additional questions were asked: what values are the key to maintaining for the European Union? What principles should be followed and what requirements should be met by the creators of intelligent machines? How does the European Union define the value of "human dignity," which in the European Union's understanding is the basis for creating the principles and values in relation to AI? What definition of the concept of human dignity does the European Union support and what philosophical concepts underlie it?

Key words: artificial intelligence, AI, ethics, European Union, guidelines