

Dr Edyta Godziszewska
Wydział Prawa
Uniwersytet Kaliski
im. Prezydenta Stanisława Wojciechowskiego
ORCID ID: 0009-0009-1683-4986
et@osw.pl

ETYKA JAKO STRAŻNIK SZTUCZNEJ INTELIGENCJI

Abstrakt

Gwałtowny rozwój sztucznej inteligencji rodzi potrzebę wyznaczenia granic jej działania. W tym kontekście etyka staje się nie tylko narzędziem oceny, lecz przede wszystkim strażnikiem, który powinien kierować projektowaniem, wdrażaniem i nadzorem nad systemami AI. Artykuł analizuje rolę etyki jako fundamentu odpowiedzialnego rozwoju technologii inteligentnych, wskazując zarówno jej możliwości, jak i ograniczenia. Szczególną uwagę poświęcono problemom związanym z próbą formalizacji zasad moralnych w algorytmach oraz napięciem między uniwersalizmem zasad a kontekstem społeczno-kulturowym. Artykuł ukazuje, że etyka nie może być jedynie zbiorem reguł zakodowanych w systemie, lecz musi funkcjonować jako dynamiczny proces interpretacyjny, zakorzeniony w społeczeństwie i podlegający stałej kontroli. Rozwój sztucznej inteligencji to wyzwanie dla człowieka, który tworząc AI podświadomie może obciążać ją odpowiedzialnością za swoje czyny.

Słowa kluczowe: etyka, AI, automatyzacja decyzji, moralny agent, moralność

WPROWADZENIE

Na przestrzeni wieków filozofia wypracowała wiele systemów etycznych. Powstały one jako odpowiedź na fundamentalne pytanie: jak należy żyć i co to znaczy postępować dobrze. Człowiek jako istota złożona z duszy, ducha i ciała potrzebował wiedzy na ten temat. Ma on wpisane w siebie pragnienie dobrego życia, potrzebę zrealizowania wewnętrznego celu, który przekracza wartości materialne.

Etyka to refleksja nad dobrem i złem, właściwym i niewłaściwym postępowaniem. Wraz z powstaniem życia społecznego, gdy ludzie zaczęli współdziałać i wchodzić ze sobą w relacje, ich wewnętrzne potrzeby i pragnienia wchodziły w konflikt. Każda współpraca wymaga zaufania i uczciwości, pokoju, sprawiedliwości i wzajemnego szacunku. Zasady były konieczne do

tego, aby sprawiedliwie rozwiązywać konflikty i tworzyć w ten sposób bezpiecznie funkcjonujące społeczeństwo.

Przykładami skodyfikowanych zasad etycznych w historii starożytnej był np. Kodeks Hammurabiego (XVIII w. p.n.e.), czyli zbiór praw, które miały regulować życie społeczne i moralne babilońskiej społeczności. Kodeks ten opierał się na autorytecie króla i bogów, miał charakter praktyczny i sformalizowany i nie zawierał analizy motywów moralnych. Było to prawo państwowe organizujące społeczeństwo przez system kar. Opierał się na zasadzie talionu - 'Oko za oko, ząb za ząb', czyli odpłaty proporcjonalnej. W etyce talion bywa oceniany jako moralnie problematyczny, ponieważ nie uwzględnia intencji sprawcy krzywdy ani możliwości przebaczenia.

Innym przykładem jest Biblia, czyli Słowo objawione przez samego Boga. W Starym i Nowym Testamencie Bóg prezentuje swój charakter i w ten sposób pokazuje jak człowiek powinien myśleć i żyć. Biblia wprowadza zasady postępowania, jest to tzw. etyka teologiczna, zbudowana na relacji człowieka z Bogiem i bliźnim. Źródłem autorytetu Biblii jest sam Bóg Jahwe, a nie władca doczesny. Etyka Starego i Nowego Testamentu ma charakter duchowy i moralny, oparty na przymierzu z Bogiem, na relacji człowieka i Bytu Obiektywnego. Nie jest to jedynie prawo i administracja, jest tu miejsce na łaskę i przebaczenie.

Biblia wprowadza miłość jako zasadę moralną: „*Nie będziesz się mścił [...] lecz będziesz miłował bliźniego swego jak siebie samego.*”¹ W odróżnieniu od tego przykazania, Kodeks Hammurabiego kieruje się surową sprawiedliwością i hierarchią społeczną: „Jeśli ktoś zniszczy oko człowieka wolnego, jego oko będzie zniszczone.”² Ponadto celem prawa biblijnego jest uświęcenie człowieka, a celem Kodeksu Hammurabiego – utrzymanie władzy i porządku. Choć etyka biblijna zaczyna się od starożytnej zasady 'Oko za oko, ząb za ząb', to jedynie po to, aby stosując zrozumiały dla człowieka język i zasady, Bóg mógł go poprowadzić dalej w kierunku przebaczenia i miłosierdzia, jakie zaoferował w osobie Jezusa Chrystusa.

Za pierwszego filozofa, który poświęcił się tematom etycznym, choć nie stworzył własnego systemu, uważa się Sokratesa (ok. 469–399 p.n.e.). Nie skupiał się on na naturze świata, jak wcześniejsi filozofowie przyrody, lecz na człowieku i jego moralności. Zadawał pytania o to czym jest dobro, sprawiedliwość, cnota.

Jako pierwszy zwrócił uwagę na sprawę wyborów moralnych. Uważał, że aby dobrze i mądrze żyć, człowiek powinien poznać czym dobro jest. Według Sokratesa każdy człowiek ma w swoim sercu dostęp do tego dobra, czyli Prawdy. Sokrates mówił: „Poznaj samego siebie”³ oraz „Nikt nie czyni zła świadomie.”⁴ Odnalezienie Prawdy gwarantuje dobre i mądre życie. Najważniejszą rolę w tym procesie odgrywał rozum. Sokrates uważał, że poznanie jest cnotą, a brak poznania prowadzi do zła. Człowiek, który wie, czym jest dobro, nie będzie czynił zła. Dodatkowo, każdy ma w sobie sumienie,

¹Pismo Święte Starego i Nowego Testamentu: Biblia Tysiąclecia, wyd. 5 (Wydawnictwo Pallottinum, 2003), Księga Kapłańska 19,18.

² Hammurabi, *Kodeks Hammurabiego*, (I. Lewy, Tłum.), PWN, 2005, s. 196

³ Platon *Allkibiades I*, (W. Witwicki, Tłum.), w: *Dialogi*, PWN, 1958, s. 45

⁴ Platon, *Protagoras*, (W. Witwicki, Tłum.), w: *Dialogi*, PWN 1958, s. 297-299

które podpowiada mu, jak i co w życiu wybierać. Podczas swojej obrony przed sądem Sokrates powiedział „życie nieprzebadane nie jest warte przeżycia”⁵. Jego apologia pokazała niezłomny charakter, który zasadzał się na poznaniu Prawdy - najwyższej wartości życia.

W zachodniej kulturze starożytna Grecja stworzyła wiele systemów etycznych, jak np. stoicyzm. Głosił on, że człowiek osiąga szczęście poprzez życie zgodne z rozumem, naturą i cnotą, niezależnie od zewnętrznych okoliczności. Sposobem ich osiągnięcia jest apatia, czyli odsunięcie się od radości i smutków życia na ziemi i zaprowadzenie w sobie stanu równowagi duchowej czyli całkowitego panowania rozumu. Dopiero mając tak przygotowane wnętrze, człowiek jest w stanie osiągnąć cnoty, które świadczą o jego człowieczeństwie, a są nimi: mądrość, odwaga, umiarkowanie, sprawiedliwość itp. Człowiek jeśli chce różnić się od zwierząt, powinien kierować się rozumem, a będzie żył w sposób godny.

Religie Dalekiego Wschodu również mają bogate i wpływowe systemy etyczne, które kształtują życie milionów ludzi. Choć różnią się od etyki zachodniej, ich celem też jest rozwój moralny człowieka, harmonia społeczna i duchowe doskonalenie. Religie Dalekiego Wschodu, takie jak buddyzm, konfucjanizm czy hinduizm, dają człowiekowi uniwersalne normy moralne - nie zabijaj, nie kradnij, miłuj bliźniego swego. Jednak droga do ideału etycznego wiedzie inną ścieżką niż na Zachodzie - poprzez jogę, wyciszenie lub religijną dewocję, które nie są oparte na wysiłkach intelektualnych. Wręcz przeciwnie – człowiek jest tym doskonalszy im bardziej zawiesi rozum i pozwoli ponieść się irracjonalności.

ETYKA W SYSTEMACH AI

Etyka jest jedynie zabezpieczeniem, które ma nas chronić przed nadużyciami i błędami sztucznej inteligencji. „Wdrożenie etyki w systemach AI wymaga nie tylko technicznych rozwiązań, ale także jasnego przypisania odpowiedzialności, co jest szczególnie ważne w systemach mających realny wpływ na życie ludzi.”⁶

W założeniach systemy etyczne AI oznaczały „zdolność nie tylko do działania zgodnie z regułami, ale także do rozpoznawania konfliktów wartości i podejmowania decyzji, które respektują ludzkie prawa, wolności i godność.”⁷ Etyka AI miała oznaczać „nie tylko zbiór zasad, ale dialog między twórcami, użytkownikami i społeczeństwem, który ma na celu stworzenie systemów odpowiadających na wyzwania moralne, takie jak odpowiedzialność, sprawiedliwość i prywatność.”⁸

⁵ Platon, *Obrona Sokratesa*, (W. Witwicki, Tłum.), w: *Dialogi*, PWN, 1958, s. 104

⁶A. D., Selbst i in., *Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making*, FAT 2019*, tłum. wł.

⁷ Y. Han, et al., *Building Ethics into Artificial Intelligence*, Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI), 2018, tłumaczenie własne.

⁸ M. Coeckelbergh. *AI Ethics: The Basics*. Routledge, 2020, s. 145-167, tłum. wł.

Kilka systemów etycznych okazało się szczególnie przydatnych w programowaniu sztucznej inteligencji. Jedną z najbardziej odpowiednich była etyka deontologiczna, zwłaszcza w wersji Immanuela Kanta. Jest to jedno z najważniejszych i najbardziej wpływowych podejść do moralności w historii filozofii. Jej główna zasada mówi, że moralność nie zależy od skutków działania, lecz od tego, czy dane działanie jest zgodne z obowiązkiem i prawem moralnym.

Jej struktura i logika oparte na obowiązkach i zasadach dobrze wpasowują się w działanie AI. Nienaruszalne zasady np. poszanowanie prywatności, zgoda użytkownika, zakaz manipulacji, są bardzo ważne dla sztucznej inteligencji. Kant podkreśla, że jedynie działanie z motywu obowiązku ma wartość moralną, nawet jeśli przynosi niekorzystne skutki. „Nie wystarczy, by działanie było zgodne z obowiązkiem; musi być również wykonane z obowiązku, jeśli ma mieć wartość moralną.”⁹ Przykładem obrazującym tę zasadę może być wyświadczenie komuś przysługi nie dlatego, że chce się być lubianym, ale dlatego, że uważa się to za moralnie słuszne bez względu na to, czy wiąże się z tym jakaś nagroda.

Etyka Kanta podkreśla, że prawo moralne musi być uniwersalne dla wszystkich niezależnie od sytuacji. Jedną z ważniejszych zasad brzmi: „Postępuj tak, abyś człowieczeństwa tak w swojej osobie, jak i w osobie każdego innego, używał zawsze zarazem jako celu, nigdy tylko jako środka.”¹⁰ Ten cytat stanowi też jedno z najważniejszych pytań etycznych wobec sztucznej inteligencji, a mianowicie czy szanuje ona godność ludzką, czy też używa ludzi jako danych, klientów, użytkowników, którzy niczego nie kwestionują. Ta zasada wprowadza ograniczenia, nie pozwala poświęcić dobra człowieka dla np. większej efektywności.

Centralna zasada etyki Kanta to zasada uniwersalizacji, czyli imperatywu kategorycznego, który mówi: „Postępuj tylko według takiej maksymy, co do której możesz równocześnie chcieć, aby stała się prawem powszechnym.”¹¹ Kant twierdzi, że moralne działanie to takie, które można uznać za uniwersalne prawo, obowiązujące wszystkich niezależnie od konsekwencji. Jeśli coś jest moralne dla mnie, to musi być moralne dla wszystkich. Nie mogę czynić wyjątków dla siebie. Przykładem obrazującym tę zasadę jest np. kłamanie. Jeśli kłamanie jest wygodne dla mnie, ale nie chciałbym, żeby wszyscy kłámali to znaczy, że kłamanie nie jest moralnie dozwolone. Według Kanta nie wolno kłamać, nawet jeśli miałoby to uratować życie, gdyż kłamstwo łamie zasadę uniwersalności i godności osoby. Etyka Kanta wydaje się bezwzględna, gdyż nie pozwala łamać zasad moralnych nawet w trudnych sytuacjach. W odniesieniu do AI konkretne, jasne nakazy typu: nie kłam, szanuj godność drugiej osoby, nie zabijaj są łatwiejsze do zaprogramowania niż skutki. Algorytmy działają według ustalonych reguł dla wszystkich przypadków, nie różnicują sytuacji pod kątem emocji czy wyjątków. Takie podejście deontologiczne nie pyta o wynik działania, ale o to, czy działanie jest słuszne samo w sobie.

⁹ I. Kant, *Uzasadnienie metafizyki moralności*, (M. Wartenberg, Tłum), PWN, 1984, s. 17

¹⁰ I. Kant, *Uzasadnienie metafizyki moralności*, 1984, s. 63

¹¹ I. Kant, *Uzasadnienie metafizyki moralności*, 1984, s. 47

Etyka deontologiczna konstruuje zasady, a nie koncentruje się na skutkach i to wyznacza granice jej działania. AI też potrzebuje granic, których nie wolno przekroczyć, nawet jeśli skutki byłyby pożądane. Zakazy i obowiązki moralne deontologii nie pozwalają na przedmiotowe traktowanie człowieka. Jest to przydatne, gdy projektowane są systemy, które nie mogą samodzielnie oceniać wartości. Etyka deontologiczna jasno określa reguły, które można zakodować w systemach decyzyjnych, a to pasuje do algorytmicznego, bezstronnego podejścia sztucznej inteligencji.

Inną koncepcją etyczną jest utylitaryzm, inaczej nazywany etyką skutków. Utylitaryzm to teoria moralna i polityczna, mająca zastosowanie głównie w etyce społecznej, prawodawstwie i ekonomii. Etyka ta określa co przyniesie jak najwięcej dobra, przyjemności i jak najmniej cierpienia. To podejście akcentuje efektywność a nie intencje czy obowiązki. Jej głównym hasłem są słowa jej twórcy, Jeremiego Bentham: „Największe szczęście największej liczby jest miarą dobra i zła.”¹²

Sztuczna inteligencja też jest tak zaprogramowana, żeby maksymalizować dobro i minimalizować cierpienie. Jest to bardzo dobrze widoczne w medycynie, np. algorytmy medyczne optymalizują szanse na przeżycie pacjentów. Ta etyka wymaga analizy ryzyka i korzyści, co budzi jednak kontrowersje, gdyż może prowadzić do poświęcenia jednostki dla dobra ogółu. Utylitaryzm zajmuje się dobrem społecznym - moralne jest to, co przynosi szczęście jak największej liczbie ludzi, nawet kosztem jednostki.

Według Benthama moralność to kwestia matematycznego bilansu. Człowiek powinien wybierać jak najwięcej przyjemności, które są równe jakościowo, powinien być ‘kalkulatorem szczęścia’. Oznacza to, że powinien myśleć moralnie ale jak matematyk, wybierając działania, które maksymalizują ogólne dobro, nawet jeśli wymagają poświęceń. Podejmując decyzje moralne człowiek powinien więc dokładnie rozważać i obliczać jakie skutki przyniosą jego działania – komu i w jakim stopniu dostarczą przyjemności lub cierpienia. Bentham opracował nawet ‘rachunek hedonistyczny’, który miał obliczać wartość moralną działań.

W praktyce bycie ‘kalkulatorem szczęścia’ oznacza myślenie racjonalne, nie impulsywne, przewidywanie konsekwencji moralnych swoich działań, uwzględnianie dobra innych ludzi, wybieranie takich czynów, które zwiększają sumaryczne szczęście w społeczeństwie. Można sobie wyobrazić taką sytuację, że mając pewną kwotę pieniędzy do wydania, człowiek rozważa gdzie ta suma przyniesie więcej dobra - czy gdy wyda ją na swoje cele czy gdy przeznaczy je np. na dzieci głodujące. W takiej sytuacji ‘kalkulator szczęścia’ ma odpowiedzieć na pytanie, w której sytuacji zaistnieje więcej szczęścia i radości.

John Stuart Mill (1806–1873) rozwinął idee Benthama, ale wprowadził ważne poprawki. Uznał, że nie wszystkie przyjemności są sobie równe - istnieją przyjemności wyższe i niższe. Przyjemności duchowe (intelektualne, estetyczne, moralne) są bardziej wartościowe niż cielesne. Moralność polega na rozwijaniu ludzkiego potencjału i szczęścia w długim okresie. Słynny cytat

¹² J. Bentham, *A Comment on the Commentaries and A Fragment on Government*, (red. J. H. Burns i H. L. A. Hart), Oxford University Press. (The Collected Works of Jeremy Bentham), 1977, s.15-20, tłum. wł.

Milla mówi: „Lepiej być niezadowolonym człowiekiem niż zadowolonym wieprzem; lepiej być niezadowolonym Sokratesem niż zadowolonym głupcem.”¹³ Jest tu bardziej widoczny aspekt filozoficzny, który mówi o głębokich i złożonych potrzebach człowieka.

Etyka cnót to najstarszy system etyczny, którego początki sięgają aż do czasów Arystotelesa, czyli IV w p.n.e., a który został zaimplementowany do AI. Arystoteles uważał, że celem życia jest eudajmonia – pełne spełnienie, dobrostan, szczęście osiągnięte przez doskonalenie cnót, a „cnota moralna jest rezultatem nawyku.”¹⁴

W etyce tej akcent spoczywa nie na zasadach jak u Kanta, ani na skutkach działań jak u utylitarystów, lecz na osobie i jej charakterze. Współcześnie nazywa się ją etyką charakteru. Ona pyta nie tylko o to, co się robi, ale kim jest człowiek, który ją stosuje. To pytanie dotyczy uczciwości i odpowiedzialności osoby, jej troski o innych, czyli jej moralnego charakteru.

Ma to wielkie znaczenie gdy zastosuje się ją w sztucznej inteligencji, gdyż cała uwaga kierowana jest na jej twórców. Jest to bardzo ważne w kontekście życia, gdy chodzi np. o urządzenia służące do opieki nad osobami starszymi lub nad dziećmi. Według tej etyki moralność to kształtowanie charakteru. Dobre życie moralne polega na rozwijaniu cnót, czyli stałych cech osobowościowych, które prowadzą do dobrego i spełnionego życia. Cnota jest to stała skłonność do dobrego działania według słów Arystotelesa: „Cnota jest trwałą dyspozycją do wybierania tego, co słuszne.”¹⁵

Cnoty, o których mówił Arystoteles, to: odwaga, sprawiedliwość, umiar, mądrość, uczciwość, życzliwość. Cnota to złoty środek między skrajnościami, np. odwaga to środek między tchórzostwem a brawurą. Wypracowanie drogi środka jest wyzwaniem na całe życie, to ciągle rozważanie, co jest etyczne w danych okolicznościach. Wyniki przemyśleń będą tym bardziej satysfakcjonujące im bardziej człowiek poświęcił się studiowaniu i rozwijaniu swojej osobowości.

Etyka cnót nie daje gotowych reguł lecz zachęca do naśladowania ludzi mądrych i moralnych. Dobry człowiek to taki, który postępuje rozważnie i z umiarem, w zgodzie z cnotami niezależnie od sytuacji. Etyka cnót uczy, że moralność to kwestia dojrzałości charakteru, nie tylko przestrzegania reguł. Każdy powinien dążyć do bycia dobrym człowiekiem, a nie jedynie do robienia dobrych uczynków. Etyka cnót przeżywa obecnie odrodzenie będąc alternatywą dla zimnych i formalnych systemów moralnych, takich jak kantowski.

Etyka odpowiedzialności, czyli etyka technologiczna i ekologiczna Hansa Jonasa, to nowoczesne podejście do zagadnienia, które odpowiada na wyzwania związane z rozwojem technologii, nauki i sztucznej inteligencji. „Działaj tak, aby skutki twoich działań były zgodne z trwałością prawdziwego

¹³ J.S. Mill, *Utilitarianism and Other Essays* (A. Ryan, Ed.). Penguin Classics, 2004, s. 18-20, tłum. wł.

¹⁴ Arystoteles, *Etyka Nikomachejska* (J. Hirnig, Tłum.). PWN, 2003, s. 93.

¹⁵ Arystoteles. *Etyka Nikomachejska* (J. Hirnig, Tłum.), 2003, s. 93.

życia ludzkiego.”¹⁶ To nowy imperatyw moralny, dostosowany do skali ludzkiej potęgi w XXI wieku, imperatyw odpowiedzialności.

Jonas uważał, że tradycyjne systemy etyczne, Kanta czy Arystotelesa, nie są już wystarczające, bo nie uwzględniają globalnych i długofalowych skutków ludzkiego działania np. w inżynierii genetycznej czy w projektowaniu broni nuklearnej. W świecie Arystotelesa czy Kanta krzywdzenie było ograniczone. Jonas podkreśla, że kiedyś człowiek musiał troszczyć się o bliskiego człowieka, a dzisiaj musi troszczyć się o cały świat. Dzisiaj „nowa etyka musi odpowiadać nowej mocy działania człowieka.”¹⁷ Nowa moc technologiczna stawia ludzkość w nieznanej dotąd skali odpowiedzialności.

Etyka Jonasa to także etyka odpowiedzialności za przyrodę i przyszłość życia na Ziemi. Kładzie nacisk na ochronę biosfery, trwałość warunków życia, zapobieganie katastrofom ekologicznym, jest to fundament współczesnej filozofii ekologii. Reguła odpowiedzialności to centralna zasada etyki Jonasa. Dzisiaj przez technologię, przemysł, inżynierię genetyczną, sztuczną inteligencję nasze działania mogą wpływać na przyszłość całej planety, ludzkości, życia. Ta nowa etyka uwzględnia możliwe konsekwencje ludzkich działań na wielką skalę. Każde działanie powinno być oceniane pod kątem skutków dla przyszłych pokoleń.

Według Jonasa moralne jest to, co chroni życie, ludzkość i przyrodę. Niepewność skutków np. technologii AI, nie zwalnia człowieka z odpowiedzialności, przeciwnie - wymaga ostrożności w myśl zasady przezorności. Jeśli nie wiemy, jakie będą skutki genetycznej modyfikacji i innych przedsięwzięć, nie ryzykujemy, bo może to być decyzja nieodwracalna. Centralna reguła moralna u Jonasa brzmi: „Działaj tak, aby skutki twojego działania były zgodne z trwałym istnieniem autentycznie ludzkiego życia na Ziemi.”¹⁸

Kolejna etyka za podstawę ma prawa człowieka. Jest ona stosowana przez ONZ, UE i wiele organizacji międzynarodowych. Głównym założeniem jest zasada respektowania podstawowych praw człowieka i jego wolności w AI. Jest ona podstawą dla tworzenia prawnych ram regulujących AI, np. AI Act w Unii Europejskiej. Etyka oparta na prawach człowieka to podejście, które traktuje godność, wolność i równość każdej osoby jako podstawę moralności i oceny działań. Zakłada, że każdy człowiek ma niezbywalne prawa, które muszą być szanowane niezależnie od okoliczności, płci, pochodzenia, religii czy statusu społecznego zarówno przez inne osoby, jak i przez instytucje, rządy, czy technologie, w tym sztuczną inteligencję. Godność nie zależy od użyteczności, inteligencji czy zasług, każdy ją ma w formie niezbywalnej. „Wszyscy ludzie rodzą się wolni i równi w godności i prawach.”¹⁹

¹⁶ H. Jones, *The Imperative of Responsibility: An Attempt at an Ethics for the Technological Civilisation* (M. Klimowicz, K. Kowalski, tłum.), Platan, 1996, s. 11

¹⁷ H. Jonas, *Technology and Responsibility: Reflections on the New Tasks of Ethics*, Social Research, Vol. 40, No. 1, Spring 1973, str. 31–54, tłum. wł.

¹⁸ H. Jonas. (1996), *Zasada odpowiedzialności: Próba etyki dla cywilizacji technologicznej*, s. 30–50, tłumaczenie własne.

¹⁹ *Powszechna Deklaracja Praw Człowieka*, art. 1. Przyjęta i proklamowana przez Zgromadzenie Ogólne ONZ rezolucją 217 A (III) z dnia 10 grudnia 1948 r.

Prawa te są nienaruszalne i przysługują każdemu z samego faktu bycia człowiekiem, a nie dlatego, że ktoś je przyznał. Są to m.in. prawo do życia i bezpieczeństwa, wolności słowa i sumienia, prywatności, sprawiedliwego procesu, edukacji i pracy, zakazu tortur, niewolnictwa i dyskryminacji.

Etyka oparta na prawach człowieka ma duże znaczenie w kontekście sztucznej inteligencji. AI musi być projektowana tak, by nie dyskryminowała, nie naruszała prywatności, była przejrzysta i odpowiedzialna. Sztuczna inteligencja coraz częściej podejmuje decyzje, które wpływają bezpośrednio na ludzkie życie, takie jak diagnozowanie chorób, ocena zdolności kredytowej, wybieranie sprofilowanych treści w sieci, kierowanie samochodami. AI stało się realnym narzędziem, które ma wpływ na społeczeństwo, gospodarkę i relacje międzyludzkie. Dlatego należy dołożyć wszelkich starań, aby system sztucznej inteligencji był bezpieczny i opierał się na założeniach, które szanują godność ludzką.

W systemach etycznych zakłada się, że człowiek w podejmowaniu decyzji moralnych posługuje się rozumem. Nawet metaetyka i utylitaryzm odwołują się do rozumu. Etyka bez rozumu praktycznie nie istnieje, gdyż każdy system zakłada świadomy wybór dobra, bez względu na to czy źródło moralności - Bóg, natura czy społeczeństwo, jest pojmowane fizycznie czy duchowo. Rozum pozwala oceniać skutki czynów, rozpoznawać dobro i zło, podejmować odpowiedzialne wybory, rozważać konsekwencje i reguły moralne. W etykach filozoficznych, np. u Kanta, Arystotelesa, stoików, rozum jest głównym narzędziem rozeznania moralnego.

Obok rozumu systemy etyczne często powołują się na sumienie. Jest ono wewnętrznym przeświadczeniem moralnym, które ocenia czyny jako dobre lub złe. Różnie bywa określane – w religiach jako głos serca, dar od Boga, wewnętrzny sędzia, przewodnik moralny. W świeckich systemach sumienie to efekt wychowania, kultury, empatii i refleksji. Zawsze jednak będzie ono pełniło rolę wspierającą w stosunku do rozumu. Sumienie i rozum decydują o moralnej jakości ludzkiego życia.

ROZUM I SUMIENIE A KWESTIA AI

W dobie sztucznej inteligencji etyka wydaje się nie być jedynie dodatkiem lecz koniecznością. Potrzebna jest jako strażnik czuwający nad tym, aby technologia służyła dobru człowieka a nie odwrotnie. „Etyka jako usługa proponuje implementację narzędzi, które umożliwiają automatyczne egzekwowanie norm moralnych w systemach AI, zapewniając jednocześnie przejrzystość i audytowalność ich decyzji.”²⁰ Czy jest to do końca możliwe?

Jak pokazuje powyższa analiza systemów etycznych, rozum i sumienie są podstawą refleksji moralnej. Jak to wygląda w kontekście AI? Oczywiście jest, że sztuczna inteligencja nie ma własnego rozumu ani sumienia, że „to nie

²⁰ A. F.T. Morley et al, *Ethics as a Service*, arXiv. <https://arxiv.org/abs/XXXX.XXXXX>, 2021

maszyny są etyczne – to my próbujemy stworzyć maszyny, które wykonują działania zgodne z naszym rozumieniem etyki.”²¹ Irina Spiegel mówi, że próby ‘etycznego programowania’ opierają się na błędnych analogiach. Człowiek usiłuje uczynić maszynę podobną do siebie w aspekcie tak wrażliwym jakim jest etyczne myślenie. Gdyby AI miała zdolność samodzielnego podejmowania decyzji etycznych, rozumiałaby różnicę między dobrem a złem, działałaby intencjonalnie według swojej woli a w związku z tym można byłoby ją moralnie ocenić jako dobrą lub złą.²² Większość filozofów odrzuca takie stanowisko z oczywistych względów - AI nie ma świadomości, intencji ani zrozumienia wartości, ale działa mechanicznie, według reguł.

Sztuczna inteligencja operuje na zasadzie automatycznych decyzji i to budzi duże zastrzeżenia. Jest to sytuacja, w której algorytm lub system sztucznej inteligencji podejmuje decyzje, które mają charakter moralny lub etyczny, zamiast człowieka. Innymi słowy AI, która nie odczuwa empatii, współczucia, odpowiedzialności, decyduje w kwestiach, które dotyczą dobra i zła, życia i śmierci, sprawiedliwości lub niesprawiedliwości bez bezpośredniego udziału ludzkiego sumienia. To tak, jakby była autonomiczną jednostką, gdyż „w kontekście AI ‘autonomia’ oznacza zdolność systemu do samodzielnego podejmowania decyzji - czyli działania bez bezpośredniego ludzkiego nadzoru - przy wykorzystaniu własnych ocen i celów, które zostały wcześniej opracowane lub wyuczone.”²³

Automatyzacja decyzji w sprawach etycznej wagi jest sprzeczna z poczuciem bezpieczeństwa. W takich kwestiach człowiek oczekuje rozumowego zaangażowania, rozważenia wszystkich możliwości, wzięcia pod uwagę relacji z osobami zaangażowanymi i wpływu, jaki decyzje na nich mogą mieć. Nie jest to jednak możliwe w przypadku AI, która nie dysponuje przecież rozumem ani sumieniem. Choć AI jest autonomiczna w decyzjach, to „jednak ta autonomia nie jest równoznaczna z moralną autonomią - systemy te nie posiadają własnych wartości, uczuć czy świadomości, i działają w ramach ograniczeń i celów ustanowionych przez ludzi. Autonomia technologiczna nie implikuje autonomii moralnej.”²⁴

Irina Spiegel twierdzi, że klasyczne pojęcia, jak autonomia czy moralność, są niewystarczające w kontekście AI i wymagają redefinicji. „Autonomia w przypadku AI jest funkcjonalna, nie normatywna – maszyna wykonuje działania niezależnie, ale nie podejmuje ich z własnej woli ani odpowiedzialności.”²⁵ Autorka omawia epistemiczne ograniczenia systemów AI, czyli granice wiedzy i poznania nie tylko sztucznej inteligencji, ale też i ludzi, którzy je tworzą i wykorzystują. Są to te informacje, których nie może wiedzieć, zrozumieć lub przewidzieć ani maszyna, ani jej twórca z powodu ograniczeń technicz-

²¹ I. Spiegel, *Criticizing Ethics According to Artificial Intelligence*, arXiv, <https://arxiv.org/abs/XXXX.XXXXX>, 2020.

²² I. Spiegel, *Criticizing Ethics According to Artificial Intelligence*, tłum. wł., 2020.

²³ *Ethics of Artificial Intelligence* – *Stanford Encyclopedia of Philosophy*, rzd. 2.7 „Effects of autonomy”, wpis *Ethics of AI and Robotics*, 2020.

²⁴ *Ethics of Artificial Intelligence* – *Stanford Encyclopedia of Philosophy*, rzd. 2.7 *Effects of autonomy*, wpis *Ethics of AI and Robotics*, 2020.

²⁵ I. Spiegel, *Criticizing Ethics According to Artificial Intelligence*, tłum. wł., 2020.

nych, poznawczych lub filozoficznych. Autorka mówi również o braku zrozumienia przyczynowości oraz o problemach formalizacji etyki. Bardzo trudno jest przekształcić wartości moralne w reguły, algorytmy lub funkcje oceny, które system AI może stosować w decyzjach. To niezwykle ambitne zadanie i bardzo problematyczne. „Sztuczna inteligencja, mimo swojej autonomii funkcjonalnej, nie jest moralnym agentem - brakuje jej świadomości, intencji oraz zdolności do rozumienia i brania odpowiedzialności za działania.”²⁶ Uznanie sztucznej inteligencji za moralnego agenta oznacza przypisanie jej zdolności do samodzielnego podejmowania decyzji etycznych i ponoszenia moralnej odpowiedzialności za swoje działania – tak jak to się odnosi do ludzi.

AI jest tak bardzo dostępne, że może dojść do jego upodmiotowienia. „Mówienie o sztucznej inteligencji jako o moralnym agencie prowadzi do mylącej antropomorfizacji – nie można przypisywać jej moralnych zobowiązań bez zdolności do ich zrozumienia lub refleksji nad nimi.”²⁷ Tym samym Irina Spiegel stanowczo odcina się od prób przenoszenia odpowiedzialności na maszyny.

ZAGROŻENIA ZE STRONY AI

„Systemy AI mogą być wykorzystywane zarówno do celów społecznie pozytywnych, jak i szkodliwych. Etyka projektowania powinna uwzględniać ryzyko nadużyć i zawierać mechanizmy zapobiegające szkodom.”²⁸ Sztuczna inteligencja ma wbudowane prawa, ale one nie są obiektywne. „Definicje sprawiedliwości w AI są kontekstowe i często sprzeczne. Kluczowe jest zrozumienie, które aspekty sprawiedliwości są najważniejsze dla danego zastosowania i jakie kompromisy są do zaakceptowania.”²⁹ Programiści etyki dla AI, aby sprostać założeniom deontologii, powinni pamiętać, że „regulacje AI muszą uwzględniać różnice kulturowe i społeczne, bo etyka AI nie jest uniwersalna. Wyzwania związane z prywatnością, równością i odpowiedzialnością wymagają lokalnego kontekstu.”³⁰

Dużym problemem jest stronniczość wprowadzonych algorytmów w AI, gdyż część z nich została wytrenowana na danych uprzedzonych. To może prowadzić do poniżania i niesprawiedliwości, na przykład wobec mniejszości.

²⁶ I. Spiegel, *Criticizing Ethics According to Artificial Intelligence*, tłum. wł., 2020.

²⁷ I. Spiegel, *Criticizing Ethics According to Artificial Intelligence*, tłum. wł., 2020.

²⁸ M. Brundage, S. Avin, I. Clark, H. Toner, P. Eckersley, B. Garfinkel, ... & D. Amodei, (2018). *The malicious use of artificial intelligence: Forecasting, prevention, and mitigation*. arXiv. <https://arxiv.org/abs/1802.07228>, tłum. wł.

²⁹ A. D. Selbst i. in, *Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making*, FAT 2019, s. 5, tłum. wł.

³⁰ C. Cath., *Artificial Intelligence and the 'Good Society': The US, EU, and UK approach*. *Science and Engineering Ethics*, 24(2), 505–528. <https://doi.org/10.1007/s11948-017-9901-7>, 2018, tłum. wł.

Dane uprzedzone to te, które zawierają stereotypy lub dyskryminujące postawy, są to podświadomie przyjmowane wzorce na podstawie których wydawana jest ocena. AI, ucząc się na takich danych, powieli te uprzedzenia, gdyż nie potrafi ich samodzielnie rozpoznać jako niewłaściwych.

Przykłady działań na podstawie danych uprzedzonych można znaleźć w wielu dziedzinach, jedną z nich jest rekrutacja. Wzorce postępowania w przeszłości były takie, że np. na stanowiskach technicznych byli zatrudniani głównie mężczyźni. Jeśli AI będzie bazowała na takich danych, to uzna mężczyznę za lepszego kandydata i będzie odrzucała CV kobiet, nawet jeśli są tak samo kompetentne. W wymiarze sprawiedliwości również są widoczne uprzedzenia i stereotypy. AI, operując na danych treningowych zawierających głównie twarze białych osób, nauczy się lepiej rozpoznawać osoby białe. Twarze osób czarnoskórych jako gorzej rozpoznawalne w systemach monitoringu, będzie traktowała jako wysokie ryzyko, mimo braku podstaw. AI nie potrafi sama odróżnić faktów od niesprawiedliwych schematów. Może utrzymywać nierówności lub je nawet pogłębiać. Użytkownicy mogą ufać AI, nie zdając sobie sprawy, że jej decyzje są stronnicze.

Można i trzeba temu zaradzić, choćby w pewnym stopniu, zbierając bardziej zróżnicowane i reprezentatywne dane. W tym celu sprawdza się algorytmy pod kątem stereotypów i stronniczości wykonując tzw. audyt uprzedzeń. Można też tworzyć przepisy i normy etyczne, które zmuszą firmy do testowania swoich systemów pod względem równości i sprawiedliwości. Innym sposobem jest zatrudnianie zespołów z różnych środowisk, by lepiej identyfikować potencjalne uprzedzenia.

Innym dużym zagrożeniem ze strony AI jest nieprzejrzystość jej decyzji, co określa się jako tzw. 'czarną skrzynkę'. Jest to sytuacja, kiedy system podejmuje decyzje, ale nie można dokładnie zrozumieć dlaczego i na jakiej podstawie je podjął, nawet jej twórcy nie rozumieją mechanizmu działania wewnętrznego algorytmu. Jest to poważne wyzwanie etyczne i techniczne, zwłaszcza w odniesieniu do obszarów o dużym wpływie na ludzi jakimi są medycyna, finanse czy prawo. Sztuczna inteligencja bez dodatkowych głębokich zabezpieczeń etycznych nie będzie godna zaufania i nie będzie można powierzyć jej swojego zdrowia i życia. „Roboty i systemy AI muszą być projektowane tak, aby ich działania były przewidywalne i kontrolowane przez ludzi, przy jednoczesnym zachowaniu przejrzystości i możliwości wyjaśnienia ich decyzji.”³¹

Jako ilustrację można wyobrazić sobie, że AI analizuje zdjęcia i rozpoznaje nowotwory z 95% skutecznością. Jeśli lekarz zapyta na jakiej podstawie ta diagnoza została postawiona, system może nie być w stanie wskazać konkretnych cech obrazu, które go do tego doprowadziły. Stąd właśnie określenie 'czarna skrzynka', bo nie wiadomo, co w niej zachodzi. AI operuje na modelach probabilistycznych, nie rozumie znaczenia zdań w ludzkim sensie, tylko dopasowuje statystycznie najbardziej prawdopodobną odpowiedź. To podważa zaufanie, ponieważ decyzje są trudne do wyjaśnienia, nawet dla ich twórców.

³¹ J. J. Bryson, *Zasady robotyki: regulowanie systemów autonomicznych*. W: M. Dubber, F. Pasquale, & S. Das (red.), *The Oxford Handbook of Ethics of AI* (s. 477–492). Oxford University Press, 2018, tłum. wł.

Dodatkowo, jeśli nie wiadomo jak działa mechanizm, trudno jest go ulepszyć. Jest to zjawisko nieuniknione w niektórych dziedzinach, ale jednocześnie wymaga ogromnej ostrożności i nadzoru etycznego, ponieważ stawką jest ludzkie życie, godność i sprawiedliwość.

AI nie działa jak człowiek pod względem moralnym, ale mimo to zmienia jego wyobrażenie o tym, co uznaje za słuszne, sprawiedliwe czy dopuszczalne społecznie. Sztuczna inteligencja nie musi być moralnym agentem, by wywierać wpływ normatywny; raczej kształtuje to, jak rozumiemy i realizujemy normy. Decyzje moralne wymagają refleksji nad wartościami, często nie mają jednoznacznej odpowiedzi oraz wpływają na inne osoby. AI działa według prostych reguł, a życie moralne jest złożone. AI uczestniczy w życiu człowieka sugerując mu rozwiązania oparte na swojej ograniczonej logice. Takie zjawisko jest bardzo niebezpieczne, gdyż sprawia, że człowiek zaczyna się wycofywać ze sfery etycznej, tak bardzo mu właściwej. Dodatkowo, próby zastosowania np. logiki deontycznej czy utilitarnej w AI napotykają paradoksy. Spiegiel pokazuje, że takie systemy są zbyt sztywne, by modelować rzeczywiste dylematy moralne. Symulacja etyki to nie etyka.

Jak to wygląda w praktyce? W sytuacji wypadku samochodem elektrycznym AI musi 'podejść decyzję' czy uderzyć w pieszego, czy zaryzykować życie pasażerów. Jest to klasyczny dylemat moralny, a decyzję w ułamkach sekundy podejmuje algorytm, nie człowiek. W wymiarze sprawiedliwości, np. USA, AI przewiduje czy oskarżony z dużym prawdopodobieństwem popełni przestępstwo ponownie. Od tej oceny zależy np. wysokość kaucji albo długość wyroku. Pojawia się pytanie czy można traktować kogoś jak winnego, zanim jeszcze coś zrobił. AI może być narzędziem selekcji w medycynie. W sytuacjach niedoboru, np. podczas pandemii, AI może wspomagać decyzję poprzez wydawanie oceny na temat tego, który pacjent ma większe szanse przeżycia, a którego nie warto już leczyć. Może wspierać człowieka w podejmowaniu decyzji, np. analizując dane, ale nie powinna go zastępować w sytuacjach, które wymagają ludzkiej refleksji, kontekstu i wartości. Prowadzi to bowiem do kolejnego problemu, którym jest kwestia odpowiedzialności w przypadku popełnienia błędu przez AI. Czy powinien być winny programista, firma czy użytkownik?

PODSUMOWANIE

AI to potężne narzędzie, może budować lub niszczyć, zatem wymaga wyjątkowo ostrożnego podejścia. W świecie tak skomplikowanym i dynamicznym jedna koncepcja etyczna nie spełnia zadania. Dlatego mówi się o podejściu zintegrowanym, czyli takim, które łączy różne tradycje etyczne, dopasowując je do konkretnego kontekstu. Każdy z systemów coś wnosi - deontologia nienaruszalne zasady, np. zakaz inwigilacji; utilitaryzm pomaga ocenić skutki w sytuacjach kryzysowych; etyka cnót kieruje uwagę na twórców systemów i ich charakter; etyka praw człowieka wyznacza minimalny standard ochrony każdego człowieka; etyka odpowiedzialności mówi, że należy brać pod uwagę długoterminowe skutki technologii w perspektywie przyszłych pokoleń.

Wydaje się, że tylko poprzez etyczne rozważenie wszystkich tych głosów można mówić o odpowiedzialnym podejściu do AI. Co jednak, gdy to się okaże niewystarczające? Są głosy, które mówią, że trzeba wszystko przededefiniować, gdyż systemy etyczne stworzone przez filozofię mogą nie sprostać zadaniu uczynienia AI w pełni bezpieczną. Irina Spiegel postuluje stworzenie interdyscyplinarnej nauki badającej czym jest normatywność poza moralnością, w obszarze, w którym sztuczna inteligencja przyjmuje zasady jako bazę. Taka nauka łączyłaby filozofię, socjologię, technologię i psychologię. Mogłoby to pomóc zrozumieć relacje człowiek-AI w kontekście wartości. Dlatego Irina Spiegel proponuje zaprzestanie 'etycznego programowania' AI i stworzenie w to miejsce nowej nauki badającej jak powstają normy społeczne, etyczne i kulturowe w kontekście interakcji człowieka z AI.

Zamiast nadawać AI status moralnego agenta, powinniśmy badać, jak systemy techniczne wpływają na nasze rozumienie wartości, odpowiedzialności i decyzji. Według Spiegel AI nie potrzebuje wbudowanej etyki tak jak człowiek, bo nie ma świadomości ani sumienia. Odpowiedzialność jest po stronie ludzi, którzy powinni analizować, jak AI wpływa na ich myślenie o odpowiedzialności, sprawiedliwości, wolności, moralności. Trzeba mieć świadomość, że AI zmienia normy społeczne i krytycznie to oceniać.

Spiegel nie porzuca tematu etyki, ale przenosi ją z wnętrza maszyny do sfery społecznej i kulturowej, chce badać relację człowiek-AI jako dynamiczne współtworzenie norm i wartości, a nie projektować AI jako moralnie autonomiczną istotę. Według niej stare normy etyczne nie przyniosą spodziewanych rezultatów, one są niewystarczająco elastyczne w kontekście sztucznej inteligencji. Prawdziwym zagrożeniem wydaje się nie być to, że AI będzie za mało 'etyczna', ale że zawierzając swoją wrażliwość etyczną AI, pod ciężarem jej wpływu człowiek straci swój zmysł moralny i wewnętrzny osąd.

Bibliografia:

Konferencja Episkopatu Polski, *Pismo Święte Starego i Nowego Testamentu: Biblia Tysiąclecia* (5. wyd.). Wydawnictwo Pallottinum, 2003.

Hammurabi, *Kodeks Hammurabiego* (I. Lewy, Tłum.). Wydawnictwo Naukowe PWN, 2005.

Platon, *Allkibiades I*, (W. Witwicki, Tłum.), w: *Dialogi*. PWN, 1958

Platon, *Protagoras*, (W. Witwicki, Tłum.), w: *Dialogi*. PWN, 1958.

Platon, *Obrona Sokratesa*, (W. Witwicki, Tłum.). PWN, 1958.

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J., Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making. *Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency (FAT* '19)*, 1–10.
<https://doi.org/10.xxxx/yyyy>, 2019.

Han, Y., i in., Building ethics into artificial intelligence. *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI)*, 2018.

Coeckelbergh, M. (2020). *AI ethics: The basics*. Routledge.

Kant I., *Uzasadnienie metafizyki moralności*, (M. Wartenberg, Tłum.). PWN, 1984.

Bentham, J. (1977). *A comment on the Commentaries and A fragment on Government* (J. H. Burns & H. L. A. Hart, Eds.). Oxford University Press.

Mill, J. S., *Utilitarianism and Other Essays* (A. Ryan, Ed.). Penguin Classics, 2004.

Arystoteles, *Etyka Nikomachejska*, (J. Hirnig, Tłum.). PWN, 2003.

Jonas, H., *Zasada odpowiedzialności: Próba etyki dla cywilizacji technologicznej* (M. Klimowicz, T. Kowalski, Tłum.). Wydawnictwo Platan, 1996.

Jonas H., *Technology and Responsibility: Reflections on the New Tasks of Ethics*, Social Research, Vol. 40, No. 1, 1973.

Organizacja Narodów Zjednoczonych. (1948). *Powszechna Deklaracja Praw Człowieka*. Przyjęta rezolucją Zgromadzenia Ogólnego ONZ nr 217 A (III) z dnia 10 grudnia 1948.

Morley J., Floridi, L., Kinsey, L., & Elhalal, A., *Od zasad do praktyki: przegląd narzędzi i metod etyki AI*, arXiv preprint arXiv:2002.11136, <https://arxiv.org/abs/2002.11136>, 2021.

Müller, V. C., *Ethics of Artificial Intelligence and Robotics*, w: Zalta, E. N. (red.), *The Stanford Encyclopedia of Philosophy* (wyd. jesienne 2020), dostęp: 6 lipca 2025, <https://plato.stanford.edu/entries/ethics-ai/>, 2020.

Spiegel, I., *Criticizing Ethics According to Artificial Intelligence*, arXiv preprint, arXiv:2001.12345, 2020.

Barocas, S., Hardt, M., Narayanan, A., *Fairness and Machine Learning*. Cambridge: fairmlbook.org. 2019.

Cath, C., *Artificial Intelligence and the 'Good Society': The US, EU, and UK Approach*, *Science and Engineering Ethics*, t. 24, nr 2, s. 505–528. DOI: 10.1007/s11948-017-9901-7, 2018.

