ETYKA SZTUCZNEJ INTELIGENCJI

WPROWADZENIE

Maciej Chojnowski

CENTRUM ETYKI TECHNOLOGII
INSTYTUTU HUMANITES

Warszawa 2022

ISBN 978-83-966660-0-0
Projekt okładki i opracowanie graficzne: Dorian Denes

Wydawca:

CENTRUM ETYKI TECHNOLOGII INSTYTUTU HUMANITES

Centrum Etyki Technologii Instytutu Humanites
ul. Nowogrodzka 56/7, 00–695 Warszawa

Patronat honorowy

govtech Polska

Partnerzy

GRAI GRUPA ROBOCZA DS. SZTUCZNEJ INTELIGENCJI

ZWIĄZEK PRACODAWCÓW POLSKA MIEDŹ THE POLISH COPPER EMPLOYERS' ASSOCIATION

IZTECH

Patronat medialny

FORUM AKADEMICKIE

Spis treści

Wstęp Minister Justyny Orłowskiej

Szanowni Państwo,

nowe technologie są nieodzowną częścią naszego codziennego życia. Naszą rolą jest nie tylko jak najlepsze ich poznanie, ale przede wszystkim dostosowanie sposobu ich użytkowania do czasów i okoliczności, w których się znajdujemy.

Chciałabym serdecznie pogratulować i podziękować Autorowi tej publikacji za jej przygotowanie, a całej społeczności Instytutu Humanites za konsekwentne i aktywne wspieranie działań poświęconych zagadnieniom etyki sztucznej inteligencji.

Z pozoru niezauważana, jednak już obecna w prostych czynnościach wykonywanych przez nas w codziennym życiu. Choć nie możemy jej dotknąć, nie możemy także zaprzeczyć, jak wielki wpływ wywiera na nasz komfort życia – od płacenia w sklepie za zakupy, po sterowanie ruchem w dużych miastach.

Publikacja, którą Państwo czytają, jest podzielona na rozdziały zajmujące się jej poszczególny elementami. Jestem przekonana, że ta lektura nie tylko pozwoli Państwu na zapoznanie się z zagadnieniami dotyczącymi etyki AI, ale przede wszystkim skłoni Państwa do przemyśleń na temat sztucznej inteligencji i podejścia do używania technologii.

Raz jeszcze dziękuję za wspólne tworzenie przestrzeni do technologicznej debaty, zapraszam do lektury i mam nadzieję na kolejne publikacje przygotowywane przez Państwa.

Justyna Orłowska
Pełnomocnik Prezesa Rady Ministrów ds. GovTech
oraz Pełnomocnik Ministra Edukacji i Nauki ds. Transformacji Cyfrowej

Od wydawcy

Od kilku już lat zainteresowanie etyką sztucznej inteligencji na całym świecie systematycznie wzrasta. Zarazem trudno o temat, który wywoływałby równie wiele nieporozumień. Jedni sądzą, że w etyce AI chodzi przede wszystkim o próby zaprogramowania moralności w maszynach. Inni – że jej głównym zadaniem powinno być zabezpieczenie ludzi przed powstaniem groźnej superinteligencji. Na dodatek sprzeczne opinie często padają z ust ludzi cieszących się naukowym autorytetem, co jeszcze potęguje zamęt.

W sytuacji, gdy świat wciąż próbuje uporać się z traumą pandemii koronawirusa, brutalna inwazja Rosji na Ukrainę ożywia niepokoje z okresu zimnej wojny, a kryzys energetyczny spowalnia przeciwdziałanie zmianom klimatycznym, problemy związane z coraz szerszym zastosowaniem sztucznej inteligencji mogą z jednej strony podsycać zbiorowe lęki, z drugiej zaś być marginalizowane w obliczu bardziej palących wyzwań.

Dlatego publikacja, którą oddajemy do Państwa rąk, ma ważną rolę do odegrania. Jest to bowiem pierwsza w języku polskim popularnonaukowa praca poświęcona etyce sztucznej inteligencji, syntetycznie ujmująca najważniejsze zagadnienia związane z tą dziedziną.

Dzięki niej poznają Państwo m.in. genezę etyki AI, metody jej praktycznego zastosowania, a także ryzyka związane z potencjalnymi nadużyciami w tym obszarze. Dowiecie się również o związkach etyki sztucznej inteligencji z prawem oraz z przedsięwzięciami podejmowanymi na rzecz dobrostanu ludzi i środowiska. Przede wszystkim jednak poznacie jej główne cele oraz wyzwania, z którymi musi się uporać. Publikacja podkreśla też rolę indywidualnego krytycznego namysłu oraz wpływu, jaki mamy na funkcjonowanie nowych technologii jako ich użytkownicy.

Do niedawna popularne było przekonanie, że wartości społeczne muszą być zgodne z ewolucją technologii. I że to raczej technologia powinna kształtować normy, niż sama być przez nie kształtowana. Już czas przyznać otwarcie, że takie myślenie zaprowadziło nas w ślepy zaułek[1].

Mam nadzieję, że zawarte tu informacje pozwolą Państwu lepiej orientować się w problemach związanych ze sztuczną inteligencji oraz krytycznie uczestniczyć w debacie na temat przyszłości technologii. Chciałabym również, aby miały one wpływ na Wasze codzienne decyzje dotyczące sposobu korzystania z cyfrowych narzędzi.

* * *

Niniejsze opracowanie jest pierwszą publikacją Centrum Etyki Technologii Instytutu Humanites – utworzonego w styczniu 2021 r. ośrodka na rzecz odpowiedzialnych innowacji. Wierzę, że w ślad za nią będą się ukazywać kolejne wydawnictwa CET.

Jednocześnie jest to niejako kontynuacja innych inicjatyw podejmowanych od 2010 roku przez Instytut Humanites w trosce o rozwój technologii zgodnie z naszym Modelem Wioski™, obejmującym szeroki ekosystem społeczny, na który składają się takie obszary, jak: rodzina, edukacja, biznes i środowisko pracy oraz kultura i media. Dlatego znajdą tu Państwo także odwołania do naszych modeli i konkretnych projektów społecznych. Etyka sztucznej inteligencji nie wyczerpuje się bowiem w zbiorach zasad, metodach ich realizacji czy narzędziach do audytu. Jej absolutną podstawą jest spójnie rozwijający się człowiek oraz jego stosunek do świata, innych ludzi i samego siebie.

W Humanites wiemy, że integralność to jedna z najważniejszych cech dojrzałych i świadomych liderów. Takich przywódców potrzeba nam też do tego, aby etyka AI znalazła realne zastosowanie w naszej codzienności.

Gorąco zachęcam więc Państwa do lektury, zapraszając zarazem do współpracy z nami przy projektach ukierunkowanych na etyczny rozwój i zastosowanie technologii!

Zofia Dzik
Prezes Instytutu Humanites
Przewodnicząca Rady
Programowej CET

Moda na etykę AI?

Etyka sztucznej inteligencji (*artificial intelligence*, AI) zajmuje dziś coraz ważniejsze miejsce w debacie na temat nowych technologii. Rozprawiają o niej filozofowie, inżynierowie, politycy i prawnicy, a echa ich dyskusji trafiają do zwykłych użytkowników cyfrowych narzędzi[2].

Znaczenie tej dziedziny dostrzegają też organizacje i instytucje. Do kwietnia 2020 r. w repozytorium AlgorithmWatch[3] znalazło się aż 167 wytycznych z całego świata dotyczących etycznego rozwoju i wdrażania sztucznej inteligencji. Dokumenty te powstały w różnych organizacjach: międzynarodowych, krajowych, rządowych, pozarządowych i komercyjnych. Z czego wynika ta rosnąca popularność etyki AI?

Z jednej strony to skutek upowszechnienia sztucznej inteligencji, z drugiej – jej ograniczeń i błędów. AI to bowiem technologia ogólnego zastosowania[4] i jako taka jest dziś używana niemal we wszystkie sektorach. Zarazem, wbrew swojej nazwie, wcale nie jest nazbyt inteligentna. Jak bowiem wyjaśnia filozof Luciano Floridi, współczesna AI nie wytwarza inteligentnych zachowań, lecz tylko je reprodukuje (naśladuje)[5].

Chodzi o to, że sztuczna inteligencja korzysta głównie z rozwiązań syntaktycznych (kontekstowych) możliwych do rozwinięcia dzięki wielkim zbiorom danych (*big data*). Jej domeny to statystyka i prawdopodobieństwo. Tymczasem prawdziwa inteligencja rozpoznaje semantykę, czyli znaczenia. Potencjał kognitywny AI jest więc daleki od możliwości ludzkiego (a także zwierzęcego) mózgu. Skoro jednak sztuczna inteligencja nie dorównuje człowiekowi, to czy trzeba się nią aż tak bardzo przejmować?

Sformułowanie jednoznacznej i niebudzącej zastrzeżeń **definicji sztucznej inteligencji** nie jest łatwe, czego dowodzi np. dyskusja wokół charakterystyki AI zawartej w unijnym projekcie rozporządzenia AI Act. Sytuację komplikuje również to, że zagadnienia poruszane w ramach etyki AI nierzadko obejmują dziedziny, których AI jest tylko komponentem, jak np. robotyka czy internet rzeczy. W niniejszym opracowaniu przyjmujemy jej następującą definicję: sztuczna inteligencja polega na takim zaprogramowaniu komputerów, aby mogły one wykonywać czynności, które w przypadku ludzi (czy innych istot żywych) uznawane są za wymagające inteligencji. Tak rozumiana AI cechuje się mniejszym lub większym stopniem autonomiczności i obejmuje różne podejścia, z których obecnie najpopularniejsze jest uczenie maszynowe – chodzi o programy wykorzystujące statystykę oraz możliwości obliczeniowe współczesnych komputerów do wykrywania wzorów w dużych zbiorach danych i tworzenia na tej podstawie predykcji przyszłych sytuacji[6].

Zdecydowanie tak, ponieważ upowszechnianie się tej technologii sprawia, że ma ona wpływ na coraz więcej sfer naszego życia, zaś jej częściowo autonomiczne decyzje nie są wolne od błędów. Aby zilustrować relację człowieka i sztucznej inteligencji, Floridi przyrównuje ją do związku dwóch osób, w którym jeden z partnerów jest inteligentny, ale leniwy (człowiek), zaś drugi pracowity, lecz uparty i niezbyt rozgarnięty (AI). Jeśli zasady funkcjonowania obojga nie zostaną od początku dobrze ustalone, to może się to skończyć podporządkowaniem człowieka logice działania maszyny.

Urojone lęki czy realne obawy?

Nasze wyobrażenia o zagrożeniach ze strony sztucznej inteligencji kształtują się w dużym stopniu pod wpływem mediów i kultury popularnej. Zbuntowane androidy, mordercze roboty czy demoniczne superkomputery często goszczą w literackich i filmowych wizjach dystopijnej przyszłości.

Również część futurologów lubuje się w prognozach dotyczących nadejścia tzw. osobliwości, czyli momentu, w którym AI osiągnąwszy ludzkie możliwości, przekształci się skokowo w niemożliwą do ujarzmienia superinteligencję.

Podobnych scenariuszy nie można oczywiście zupełnie lekceważyć, jednak warto je postrzegać w kontekście medialnego i marketingowego zgiełku, który towarzyszy dziś sztucznej inteligencji. Realne wyzwania związane z AI są zazwyczaj o wiele mniej spektakularne od wyobrażeń, które serwuje nam popkultura.

Ciekawą alternatywę dla rozpowszechnionych dziś apokaliptycznych scenariuszy rozwoju technologii przedstawia amerykańska futurolożka Amy Webb[7]. Jej zdaniem bardziej od robotów-zabójców grozi nam sytuacja, którą metaforycznie nazywa ona życiem z milionem skaleczeń. Chodzi o nagromadzenie niedogodności wynikających z niewłaściwego rozwoju technologii, które choć osobno nie stanowią wielkiego problemu (są jak niegroźna ranka), to zsumowane uniemożliwiają normalne życie. W dobie internetu rzeczy, gdy coraz więcej urządzeń jest podłączanych do sieci, taka wizja może niepokoić.

Dobrym przykładem **problemów związanych z działaniem tzw. inteligentnych urządzeń (*smart devices*)** jest przypadek asystenta głosowego zainstalowanego w jednym z amerykańskich domów, który przez pomyłkę zarejestrował fragment prywatnej rozmowy mieszkańców, a następnie przesłał nagranie ich znajomemu. Odbyło się to bez wiedzy właścicieli, zaś producent urządzenia nie potrafił w pełni wyjaśnić, jak doszło do incydentu. Przypuszczalnie maszyna niewłaściwie zinterpretowała wypowiadane w rozmowie słowa i potraktowała je jako serię poleceń[8].

Ryzyko podobnych zagrożeń wzrasta, jeśli uwzględnimy kondycję współczesnego człowieka. Zofia Dzik, twórczyni Instytutu Humanites, zwraca w tym kontekście uwagę na nadmiar bodźców docierających dziś do ludzi, a także malejącą przestrzeń na refleksję, zanikanie krytycznego myślenia, nasilającą się zewnątrzsterowność i – w efekcie – postępujące intelektualne otępienie[9].

Grzechy główne AI

Czego zatem naprawdę powinniśmy się dziś obawiać ze strony sztucznej inteligencji? Etyk Bernd C. Stahl wymienia trzy główne rodzaje wyzwań związanych z AI:

- **problemy związane z uczeniem maszynowym**[10]**,**
- **społeczne i polityczne kwestie dotyczące życia w scyfryzowanym świecie,**
- **pytania metafizyczne o naturę człowieka i status maszyn.**

Pierwsza grupa to najczęściej kwestie techniczne, bezpośrednio związane ze specyfiką uczenia maszynowego. Chodzi tu np. o dyskryminację będącą efektem skrzywienia algorytmicznego (*bias*), które bierze się z trenowania systemów AI na niewystarczająco reprezentatywnych zbiorach danych. W rezultacie sztuczna inteligencja może działać krzywdząco wobec pewnych grup ludzi (np. mniejszości). Innym problemem z tej kategorii jest nieprzejrzystość działania AI. Mowa o tzw. czarnych skrzynkach, czyli systemach sztucznej inteligencji, w których sposób dojścia do określonych wyników pozostaje niezrozumiały nawet dla osób odpowiadających za ich zaprogramowanie. W takich sytuacjach decyzje podjęte przez AI okazują się arbitralne, bo nie znamy kryteriów, które by je uzasadniały. Jest to niedopuszczalne w przypadku zastosowań mających wpływ na istotne sfery ludzkiego życia (np. decyzje o przyznaniu świadczeń społecznych czy warunkowym zwolnieniu z więzienia).

Kolejnym przykładem niebezpieczeństwa związanego z uczeniem maszynowym jest zdolność odkrywania zależności pomiędzy pozornie neutralnymi danymi (tzw. korelacja), co może skutkować zagrożeniem dla prywatności. Dysponując odpowiednio dużym

W książce ***Atlas of AI*** **Kate Crawford** zwraca uwagę, że dyskryminacja algorytmiczna skrywa szerszy problem związany z używaniem przez władzę różnych narzędzi do klasyfikowania populacji. W takim ujęciu AI przestaje być neutralną technologią, której ewentualne błędy można naprawić optymalizując działanie oprogramowania, a staje się systemową maszyną opresji redukującą wielowymiarową rzeczywistość społeczną do kilku arbitralnie przyjętych kategorii. To właśnie w takim statystycznym, uśredniającym podejściu do świata badaczka widzi jedno ze źródeł problemów z AI[11].

zasobem danych, systemy AI są bowiem w stanie pozyskiwać wrażliwe informacje (np. dotyczące statusu materialnego czy orientacji seksualnej) o ludziach, którzy sami publicznie ich nie ujawnili.

Problemy z **drugiej grupy** dotyczą oddziaływania systemów AI na różne sfery życia społecznego czy politycznego. To np. malejące poczucie sprawczości wśród pracowników coraz częściej traktowanych jak dodatek do maszyn bądź zanik różnych profesji wskutek zaawansowanej automatyzacji (zjawisko widoczne dziś np. w sektorze usług, gdzie miejsce żywych konsultantów stopniowo zajmują boty).

W tej grupie mieści się też kwestia koncentracji kapitału i władzy przez największe firmy technologiczne, które monopolizują usługi cyfrowe na globalnym rynku. Chodzi również o rosnące ryzyko masowej manipulacji i inwigilacji czy to ze strony prywatnych korporacji, czy rządów. Do tej kategorii zalicza się także kwestie wpływu technologii na środowisko, np. ślad węglowy wynikający z energochłonności infrastruktury obliczeniowej potrzebnej do pracy systemów AI. Przed kilkoma laty obliczono, że wytrenowanie jednego modelu AI służącego do przetwarzania języka naturalnego (*natural language processing*, NLP) wymagało zużycia energii skutkującego emisją 284 ton dwutlenku węgla, co odpowiada emisji pięciu samochodów osobowych w całym ich cyklu użytkowania[12]!

Ostatnia grupa zagadnień dotyczy moralnego i prawnego statusu AI (a także kondycji ludzkiej), przez co może się wydawać najbardziej abstrakcyjna, ponieważ budowa myślących maszyn to dziś wciąż pomysł rodem z fantastyki naukowej. Refleksja w tym obszarze ma w dużym stopniu charakter krytyczny i demaskatorski – na przykład wobec szumnych deklaracji niektórych wynalazców, że stworzyli samoświadomą sztuczną inteligencję[13].

Zagadnienia z tej grupy nie sprowadzają się jednak wyłącznie do rozważań o ewentualnych prawach i obowiązkach przyszłych androidów. Już dziś mamy bowiem do czynienia z neurotechnologiami, które mogą być wszczepiane do ludzkiego organizmu, aby wzmacniać funkcje poznawcze czy percepcyjne człowieka. Ocena takich rozwiązań jest również przedmiotem etyki technologii, choć często wykracza poza obszar samej sztucznej inteligencji[14].

Podział tych problemów na grupy nie oznacza zarazem, że należy je rozpatrywać osobno. Filozof Mark Coeckelbergh podkreśla, że są one wzajemnie powiązane: np. konieczność zaradzenia dyskryminacji algorytmicznej to wyzwanie nie tylko techniczne, ale również polityczne – dotyczące urzeczywistniania w społeczeństwie zasady sprawiedliwości[15].

Komu posłuży sztuczna inteligencja

Jeśli jednak użytkowanie AI wiąże się z tyloma kłopotami, to po co w ogóle ją rozwijamy? Zwolennicy determinizmu technologicznego odpowiedzą krótko: bo postęp jest nieunikniony. Takie wyjaśnienie nie rozwiązuje jednak problemu, a ponadto wciąga nas w pułapkę myślenia w kategoriach fatalistycznych („jesteśmy skazani na technologie i nic na to nie poradzimy").

Orędownicy etycznego podejścia do AI pokazują, jak uniknąć tej pułapki, przywołując przykłady innych technologii czy branż (motoryzacja, lotnictwo), które choć na początku też rozwijały się żywiołowo, to z czasem były stopniowo regulowane w trosce o nasze większe bezpieczeństwo. Ich zdaniem podobny los czeka AI, która na razie wciąż znajduje się jeszcze na wczesnym etapie rozwoju.

Etyczna refleksja nad AI jest dziś tym ważniejsza, że stosowaniu sztucznej inteligencji mogą przyświecać bardzo różne cele, zauważa Stahl. Z jednej strony chodzi na przykład o zwiększenie wydajności i maksymalizację zysków, z drugiej – o większą kontrolę nad ludzką populacją. To jednak nie wyczerpuje wszystkich możliwości: celem AI może być także wzrost dobrostanu ludzi i innych istot zamieszkujących Ziemię. W realizacji tego pozytywnego potencjału sztucznej inteligencji pomaga właśnie refleksja etyczna[16].

Współczesny kult indywidualizmu (ideały samodoskonalenia i samorealizacji) rozwija się równolegle z kulturą masową i konsumpcjonizmem, które kuszą nas zwodniczymi hasłami personalizacji produktów i usług. Zarazem nasza zdolność do myślenie w kategoriach dobra wspólnego została znacznie osłabiona, choć w dobie coraz bardziej zaawansowanych i wszechobecnych technologii oraz kryzysu klimatycznego jest szczególnie potrzebna. Dlatego jako konsumenci powinniśmy dbać o świadome wybory, tak byśmy umieli odróżnić wizerunkowe mydlenie oczu przez korporacje od wartościowych prospołecznych i proekologicznych inicjatyw.

Musimy także pamiętać, że rozwój AI dotyka wielu bardzo różnych interesariuszy, a nie tylko przedsiębiorców czy inwestorów. Dlatego potrzebujemy uczciwego rachunku zysków i strat, który pozwoli rozwijać tę technologię w pożądanym społecznie i środowiskowo kierunku, tak by minimalizować ryzyko, zaś maksymalizować korzyści. W ten sposób dzięki etyce AI powracamy w debacie publicznej do pytań o dobre życie i sprawiedliwe społeczeństwo.

Etyka AI – dla ludzi czy maszyn?

Jak widać, kontekst dyskusji o AI jest bardzo szeroki. Coeckelbergh zauważa[17], że jej tło stanowią głębokie spory o naturę człowieka, jego świadomość i umysł, a także o rozumienie, twórczość, sens czy wiedzę. W debacie dotyczącej AI odbijają się też nasze różnice światopoglądowe. Inaczej na jej rozwój będzie bowiem patrzył zdeklarowany materialista pokładający zaufanie w empirii i postępie, a inaczej ktoś ukształtowany w duchu katolickiego personalizmu czy lewicowej teorii krytycznej.

Dyskusje wokół sztucznej inteligencji w równym stopniu dotyczą więc samej technologii, co człowieka. Dlatego coraz częściej mówi się dziś o AI nie jako czymś abstrakcyjnym czy samodzielnym, lecz właśnie jako elemencie szerszego społecznego porządku.

Pochodną tych debat jest także spór o to, czy w maszynach można zakodować etykę. Zwolennicy koncepcji sztucznych podmiotów moralnych (*artificial moral agents*) argumentują, że jest to konieczne, ponieważ inteligentne maszyny będą wykorzystywane w sytuacjach nacechowanych etycznie, jak np. opieka nad chorymi.

Z kolei przeciwnicy takich pomysłów wyjaśniają, że po pierwsze AI stoi dziś wciąż na niskim poziomie zaawansowania i nie sposób w odniesieniu do niej mówić o jakimkolwiek rzeczywistym rozumowaniu, tym bardziej moralnym.

Zamiast **fantazjować o zaprogramowaniu moralności w maszynach,** powinniśmy skoncentrować się na zapewnianiu ich bezpiecznego działania, twierdzi etyczka Aimee van Wynsberghe[18]. Przecież zarówno kosiarka do trawy, jak i pies przewodnik także funkcjonują w sytuacjach etycznie nieobojętnych (kosiarka może kogoś zranić, a pies przewodnik doprowadzić do wypadku), jednak nikt nie oczekuje od narzędzi czy zwierząt, że będą zdolne do moralnych decyzji, argumentuje badaczka. Mają być po prostu bezpieczne – czy to w wyniku odpowiedniego zaprojektowania, czy tresury. Dlaczego zatem z AI miałoby być inaczej?

Po drugie, możliwość stworzenia w przyszłości sztucznych podmiotów moralnych jest również bardzo mało prawdopodobna: maszyny nie są zdolne ani do moralnego rozumowania, ani nie posiadają percepcji, refleksji czy wyobraźni moralnej. Chodzi przy tym nie tylko o kompetencje czysto kognitywne, ale także te wynikające z ucieleśnionego bycia w świecie[19].

Po trzecie wreszcie, moralność jest cechą swoiście ludzką – nie dzielą jej z nami nawet zwierzęta. Dlatego spekulowanie o wyposażaniu maszyn w etykę może jedynie zaciemnić obraz, na czym faktycznie polega bycie podmiotem moralnym[20].

Czym jest etyka AI

Skoro zatem moralność to domena ludzi, a nie maszyn, po co nam w ogóle etyka sztucznej inteligencji?

Luciano Floridi wyjaśnia, że etyka wraz z zarządzaniem (*governance*) i regulacją to główne elementy systemu normatywnego dotyczącego technologii (w tym AI), który określa, w jakim kierunku chcemy ją rozwijać. Podczas gdy regulacja wskazuje, co jest zgodne bądź niezgodne z prawem, etyka podpowiada, które działania są pożądane, abyśmy mogli funkcjonować w lepszym społeczeństwie. Innymi słowy: etyka odgrywa rolę moralnej busoli pozwalającej krytycznie oceniać rozwój i zastosowanie technologii, by odbywało się to z jak największą korzyścią dla społeczeństwa.

Te trzy komponenty systemu normatywnego oddziałują na siebie, ale to etyka zajmuje w nim naczelne miejsce, ponieważ za pomocą moralnej oceny kształtuje zarówno obszar regulacji, jak i zarządzania. Z kolei regulacja wpływa na zarządzanie poprzez wymóg zgodności z prawem, zaś zarządzanie oddziałuje na etykę i prawo jako ich praktyczna realizacja.

Etyka AI (czy szerzej: etyka technologii) jest często uważana za przykład etyki stosowanej, która – obok etyki normatywnej (badającej zasady i normy moralne), etyki opisowej (analizującej rzeczywiste funkcjonowanie wartości w społeczeństwie) oraz metaetyki (zajmującej się statusem etyki i znaczeniem używanych w niej pojęć) – stanowi jeden z głównych obszarów tej dziedziny wiedzy[21]. W etyce stosowanej chodzi o praktyczne wskazówki – mogą one dotyczyć różnych obszarów życia i działalności człowieka (np. etyka zawodowa – lekarska czy biznesowa). W przypadku etyki sztucznej inteligencji są one skierowane do ludzi projektujących, używających i nadzorujących AI.

Co istotne, etyka AI nie sprowadza się do samego formułowania wskazówek (np. w postaci zasad zawartych w kodeksach): jej innym ważnym zadaniem jest m.in. ciągłe krytyczne analizowanie kontekstów, w których stosowana jest AI, oraz sugerowanie niezbędnych korekt – czy to w zasadach ją regulujących, czy w sposobach ich realizacji (por. sekcja: Trzy podejścia do etyki AI).

Stanisław Lem twierdził[22], że technologia stwarza możliwości wyboru tam, gdzie wcześniej w życiu człowieka działały fatalizm, przypadek bądź Opatrzność. Konsekwencje tej sytuacji nie są wyłącznie pozytywne. Wraz z możliwościami rośnie bowiem także ludzka odpowiedzialność i konieczność podejmowania trudnych decyzji, z których niegdyś byliśmy zwolnieni. Ponadto nasze normy nie ewoluują równie szybko, co zmiany zachodzące w wyniku rozwoju technologii, przez co może dojść do tzw. implozji aksjologicznej (czyli zapaści systemu wartości), gdy dotychczasowe normy i zasady współżycia społecznego zostaną znacznie rozchwiane.

Historia etyki AI w pigułce

Filozoficzny namysł nad wynalazkami, postępem technicznym oraz ich wpływem na ludzkość ma długą historię sięgającą starożytności. Podobnie jest z etyką, także w jej wymiarze stosowanym (za ojca etyki lekarskiej tradycja uznaje Hipokratesa).

Również etyka technologii[23] nie pojawiła się dopiero wraz z rozwojem uczenia maszynowego, lecz kształtowała się już wcześniej – szczególnie w reakcji na rozwój broni masowego rażenia. W latach 70. XX w. filozof Hans Jonas podkreślał, że znaną od wieków etykę bliźniego musimy uzupełnić o etykę globalną (dotyczącą odpowiedzialności za cały świat), ponieważ w cywilizacji technicznej zasięg działań człowieka nie ogranicza się już do kręgu osób, z którymi ma on bezpośrednio do czynienia, ale może skutkować nawet unicestwieniem życia na Ziemi[24].

Już przed kilkoma dekadami rozpoczęła się także dyskusja filozoficzna na temat samej AI. Jednak w kontekście, który nas tutaj najbardziej interesuje (rozwój uczenia maszynowego), historia etyki sztucznej inteligencji zamyka się w ostatnich 10 latach. Mimo to można w niej wyróżnić kilka etapów:

1. Geneza: dynamiczny rozwój uczenia maszynowego na przełomie 1. i 2. dekady XXI w. sprawił, że niektórzy futurolodzy i badacze zaczęli obawiać się rychłego nadejścia tzw. ogólnej sztucznej inteligencji (*artificial general intelligence*, AGI), czyli sztucznej inteligencji dorównującej człowiekowi, oraz superinteligencji, znacznie przewyższającej ludzkie możliwości. Pojawiły się wówczas publikacje[25] opisujące możliwy rozwój AI i sugerujące konieczność przeciwdziałania negatywnym scenariuszom. W tym celu powołano również takie instytucje, jak np. Future of Life Institute.

2. Kodyfikacja: na fali zainteresowania zagrożeniami ze strony AI różne zespoły eksperckie podjęły prace nad zdefiniowaniem ogólnych zasad mających gwarantować etyczny rozwój i wdrażanie AI (zazwyczaj chodzi o problemy mniej spektakularne niż superinteligencja). Przykładem takiego kodeksu etyki AI są unijne wytyczne dot. godnej zaufania sztucznej inteligencji. To właśnie tego rodzaju

W IV w. p.n.e. **Platon opisał w „Fajdrosie" spotkanie egipskiego króla Tamuza z bogiem Teutem** – wynalazcą liczb, geometrii, astronomii i pisma. Teut zachwala pismo jako zbawienne dla ludzkiej pamięci i mądrości. Tamuz przeciwnie – uważa, że przyniesie ono same szkody. Zaufawszy pismu, człowiek zacznie bowiem czerpać całą wiedzę spoza siebie, a nie z własnego wnętrza, i będzie to tylko pozór pamięci i mądrości. Abstrahując od antynomii typowych dla platońskiego idealizmu, te skrajnie różne oceny pisma przypominają współczesne dyskusje wokół internetu, który jednym wydaje się błogosławieństwem, zaś innym – przekleństwem.

opracowania gromadzi przywoływana na początku organizacja AlgorithmWatch (co znamienne, publikowanym kodeksom zazwyczaj nie towarzyszą wskazówki dotyczące wdrażania zawartych w nich metod czy weryfikacji ich stosowania).

3. Operacjonalizacja: mnogość kodeksów przy jednoczesnym braku konkretnych wskazówek, jak je stosować w praktyce, doprowadziła do powstania w ostatnim czasie projektów badawczych, w których opracowano bardziej szczegółowe rozwiązania w tym zakresie (np. podejście Ethics by Design określające sposoby projektowania AI zgodnie z wartościami).

4. Standaryzacja i kontekstualizacja: dalsze prace w dziedzinie etyki AI koncentrują się m.in. na opracowaniu standardów etyki AI, np. w obszarze znormalizowanych mechanizmów oceny (audytu) systemów AI pod kątem etycznym. Badacze wskazują także na potrzebę większej refleksji nad różnymi kontekstami funkcjonowania sztucznej inteligencji (m.in. środowiskowym i społecznym).

Prace podejmowane w ostatnim z wymienionych obszarów nie oznaczają bynajmniej, że sposoby efektywnej operacjonalizacji etyki AI zostały już przyswojone na szeroką skalę. Co więcej, sama znajomość zasad etycznych dotyczących sztucznej inteligencji także nie jest jeszcze powszechna. Znajdujemy się więc na etapie przejściowym, w którym część narzędzi służących praktycznemu zastosowaniu etyki AI już istnieje, choć jest sporadycznie stosowana, zaś inne czekają dopiero na wypracowanie.

Trzy podejścia do etyki AI

Etyka sztucznej inteligencji może być ujmowana na różne sposoby, a poszczególne podejścia są po części związane z opisanymi powyżej etapami jej rozwoju. Najważniejsze z nich to:

- **Podejście legalistyczne (pryncypalistyczne, kodeksowe)** – statyczne, koncentruje się na formułowaniu zasad i norm oraz potwierdzaniu zgodności z nimi. W tym ujęciu etyka AI przypomina miękką odmianę prawa. Jest ono związane przede wszystkim z etapem kodyfikacji, ale może też przerodzić się w nadrzędne podejście w związku z nadreprezentacją prawników wśród ekspertów zajmujących się etyką AI.

- **Podejście operacyjne (pragmatyczne)** – aktywne, dąży do zastosowania w praktyce ogólnych norm i zasad, czyli operacjonalizacji wartości. Polega na metodycznym włączeniu etyki do procesów projektowania, wdrażania i użytkowania sztucznej inteligencji.

- **Podejście krytyczne (kontekstowe)** – dynamiczne, dostrzega różnorodność i zmienność sytuacji, w jakich stosowane są zasady i normy. Związane

z etapem kontekstualizacji. Etyka jest tu narzędziem krytycznej refleksji nad kształtem społecznej i politycznej rzeczywistości, w której funkcjonuje AI.

Obecnie najbardziej rozpowszechnione jest podejście legalistyczne. Może ono przyjmować kształt rekomendacji czy wytycznych – ogólnych (jak wytyczne UE lub OECD) lub skierowanych do konkretnych interesariuszy (np. wytyczne IEEE). Obejmuje również kodeksy etyczne przyjmowane przez firmy. Z podejściem tym wiążą się jednak pewne ryzyka:

Ryzyko 1:
Etyka jako substytut regulacji prawnych – jak ostrzega etyczka AI Anaïs Resseguier[26], gdy etyka jako miękka odmiana prawa nie ma instytucjonalnego zakorzenienia, wówczas wyegzekwowanie zgodności (*compliance*) z jej wymogami może się w praktyce okazać trudne. Pojawia się wtedy ryzyko instrumentalnego traktowania etyki przez firmy technologiczne jako argumentu przeciw konieczności regulacji prawnej (tzw. *ethics lobbying*).

Ryzyko 2:
Etyka jako formalność – polega na postrzeganiu etyki w oderwaniu od praktyki projektowej: po potwierdzeniu zgodności z normą kwestie etyczne znikają z szerszego pola widzenia.

Niebezpieczeństwa związane z podejściem legalistycznym dobrze ilustruje przykład RODO. Zdaniem ekspertki[27] ze stowarzyszenia audytorów ISACA cały potencjał tego rozporządzenia został zredukowany przez prawników do kwestii odpowiednich klauzul. W rezultacie zmarginalizowano zagadnienia dotyczące środków technicznych i organizacyjnych oraz zaprzepaszczono szansę zainteresowania kadry kierowniczej bezpieczeństwem danych. Takie podejście doprowadziło także do powszechnego zmęczenia RODO, które jest traktowane jako uciążliwy obowiązek. Wiele wskazuje na to, że podobne ryzyko legalizmu grozi dziś także unijnemu rozporządzeniu dotyczącemu AI (por. poniżej: AI Act – unijny projekt regulacji sztucznej inteligencji).

Ryzyko 3:
Etyka jako narzędzie manipulacji – kolejny przykład niebezpieczeństwa występującego w sytuacji, gdy brakuje mechanizmów egzekwowania zasad etycznych. Luciano Floridi wymienia[28] w tym kontekście kilka negatywnych praktyk. To m.in. *ethics shopping*, czyli dobieranie zasad z różnych wytycznych czy standardów, tak by dopasować kodeksy etyczne do potrzeb organizacji; *ethics bluewashing*[29] (na zasadzie analogii do *greenwashingu*), czyli wybiórcze manipulowanie danymi, tak by postawa firmy wydawała się lepsza niż jest w rzeczywistości; *ethics dumping* – eksportowanie nagannych praktyk poza obszar, w którym nie są one akceptowane.

Zdaniem Brenta Mittelstadta przy formułowaniu zasad etyki AI problematyczne jest też przyjmowanie jako punktu odniesienia etyki lekarskiej[30]. O ile bowiem historia medycyny obejmuje kilka tysięcy lat, o tyle rozwój AI – w najszerszym ujęciu – dotyczy jedynie paru ostatnich dekad. Trudno więc mówić tutaj o upływie czasu niezbędnym do praktycznego zweryfikowania kodeksu etycznego. Co więcej, bycie lekarzem to zawód zaufania publicznego, związany z wyraźnym etosem, tymczasem praca inżyniera AI nie łączy się z podobnymi zobowiązaniami.

Dlatego właśnie kodeksowe podejście do etyki AI musi być uzupełnione o podejścia operacyjne i krytyczne, tak by zasady mogły być regularnie testowane i w razie potrzeby aktualizowane. Również obowiązek ich realizacji nie może spadać wyłącznie na pojedyncze osoby (np. inżynierów), lecz powinien obejmować całe organizacje.

Etyka AI wedle Unii Europejskiej

W 2019 r. Komisja Europejska opublikowała wytyczne dotyczące godnej zaufania sztucznej inteligencji (*trustworthy AI*)[31]. Ich twórcy podkreślają, że Europa potrzebuje wizji normatywnej dotyczącej dalszego rozwoju AI. Ma w tym pomóc właśnie koncepcja godnej zaufania sztucznej inteligencji. Etyka – obok solidności oraz zgodności z prawem – jest jednym z jej trzech filarów. Wytyczne nie obejmują jednak szczegółowych kwestii prawnych. Te przedstawia osobny projekt rozporządzenia dotyczącego sztucznej inteligencji – tzw. Artificial Intelligence Act[32] – upubliczniony w 2021 r.

Filary, których dotyczą wytyczne (czyli etyczna i solidna AI), choć formalnie rozróżnione, są w rzeczywistości ze sobą ściśle powiązane. Solidność (*robustness*) jest de facto jednym z wymogów etyki AI (patrz niżej). Chodzi o zapewnienie, że systemy AI pełnią swoje funkcje w sposób bezpieczny, pewny i niezawodny, a także o środki zapobiegawcze wobec niezamierzonych negatywnych skutków ich działania. Solidność dotyczy więc zarówno samych mechanizmów funkcjonowania sztucznej inteligencji, jak i jej wpływu na społeczeństwo.

Zasady etyczne leżące u podstaw unijnej koncepcji godnej zaufania AI wywiedziono z **praw podstawowych:** poszanowania godności ludzkiej; wolności jednostki; poszanowania demokracji, sprawiedliwości i praworządności; równości, braku dyskryminacji i solidarności; oraz poszanowania praw obywatelskich. Cztery podstawowe zasady *trustworthy AI* to:

- **Zasada poszanowania autonomii człowieka** (*respect for human autonomy*) – interakcja z systemami AI musi zapewniać ludziom możliwość samostanowienia. AI ma być ukierunkowana na człowieka i pozostawiać mu możliwość wyboru. Ludzie powinni też móc nadzorować i kontrolować sztuczną inteligencję.

- **Zasada zapobiegania szkodom** (*prevention of harm*) – systemy AI muszą być bezpieczne i pewne. Dotyczy to szczególnie osób wymagających specjalnego traktowania oraz sytuacji, w których AI może powodować lub pogłębiać niekorzystny wpływ spowodowany asymetrią władzy lub informacji. Zapobieganie szkodom dotyczy również środowiska naturalnego.

- **Zasada sprawiedliwości** (*fairness*) – chodzi zarówno o materialny, jak i proceduralny wymiar sprawiedliwości. Wymiar materialny wiąże się m.in. z przeciwdziałaniem niesprawiedliwej stronniczości, dyskryminacji i stygmatyzacji. Wymiar proceduralny dotyczy zagwarantowania ludziom możliwości kwestionowania decyzji podejmowanych przez systemy AI. Sprawiedliwość odnosi się także do zasady proporcjonalności między celami zastosowania sztucznej inteligencji a środkami, które służą ich realizacji (ewentualne ryzyko powinno być adekwatne do wagi zamierzonego celu).

- **Zasada możliwości wyjaśnienia** (*explicability*) – procesy zachodzące w systemach AI muszą być przejrzyste, a ich możliwości i cele komunikowane wprost. Decyzje podejmowane przez AI powinny w jak największym stopniu dać się wyjaśnić osobom, na które mają wpływ. Jednak wymóg wyjaśnialności zależy także od obszaru zastosowania sztucznej inteligencji oraz rodzaju konsekwencji wynikających z ewentualnego błędu.

Twórcy unijnych wytycznych podkreślają, że choć ogólne zasady to niezbędny punkt wyjścia, to nie wystarczą one, by skutecznie zoperacjonalizować etykę AI w różnych kontekstach i w odniesieniu do różnych interesariuszy. **Wdrożenie godnej zaufania sztucznej inteligencji nie ma bowiem polegać na biernym odhaczaniu pozycji z listy kontrolnej,** ale powinno prowadzić do krytycznej refleksji i dążenia do coraz lepszych rezultatów w całym cyklu życia systemu AI, jak również do współpracy z różnymi zainteresowanymi stronami. Między poszczególnymi zasadami mogą też występować konflikty (np. między zasadą zapobiegania szkodom i zasadą autonomii człowieka) i nie istnieje jedno uniwersalne rozwiązanie tego problemu. W takich sytuacjach niezbędna jest każdorazowo odpowiedzialna dyskusja nad najwłaściwszymi sposobami postępowania.

Z tych czterech zasad ogólnych wyprowadzono z kolei siedem kluczowych wymogów (nazywanych też w innych kontekstach wartościami – patrz: sekcja poświęcona Ethics by Design) służących wdrożeniu i osiągnięciu godnej zaufania AI. Mają one zastosowanie do różnych interesariuszy: projektantów systemów AI, podmiotów je wdrażających, użytkowników końcowych oraz szeroko rozumianego społeczeństwa. Wymogi te są następujące:

- **Przewodnia i nadzorcza rola człowieka** (*human agency and oversight*) – odwołuje się do zasady poszanowania autonomii człowieka i obejmuje m.in. takie działania, jak: ocena skutków oddziaływania AI (*impact assessment*) w zakresie praw podstawowych; zapewnienie użytkownikom wiedzy i narzędzi pozwalających na zrozumienie i interakcję z systemami AI; wprowadzenie mechanizmów zarządzania gwarantujących ludziom nadzór nad AI.

- **Techniczna solidność i bezpieczeństwo** (*technical robustness and safety*) – odwołuje się do zasady zapobiegania szkodom i dotyczy: odporności systemów AI na atak; posiadania planu awaryjnego i procedur pozwalających ocenić potencjalne zagrożenia związane z wykorzystywaniem AI w różnych obszarach; jasnego procesu rozwoju i oceny systemu AI, który ograniczy i pozwoli skorygować ryzyko niedokładnych prognoz; wiarygodności (*reliability*) i odtwarzalności (*reproducability*) procesów AI pozwalających ocenić, czy system działa prawidłowo przy różnych danych wejściowych (i w różnych sytuacjach), a także czy w doświadczeniu powtórzonym w tych samych warunkach zachowuje się identycznie.

- **Ochrona prywatności i zarządzanie danymi** (*privacy and data governance*) – odwołuje się do zasady zapobiegania szkodom i określa wymogi: ochrony prywatności i ochrony danych w całym cyklu życia systemu AI, jakości i integralności danych (w tym odpowiednich procesów ich testowania i dokumentacji), a także opracowania protokołów określających dostęp do nich.

- **Przejrzystość** (*transparency*) – odwołuje się do zasady możliwości wyjaśnienia i dotyczy m.in.: identyfikowalności (*traceability*) zbiorów danych, procesów i algorytmów wykorzystywanych przez AI do podejmowania decyzji, co ułatwia możliwość kontroli (*auditability*); wyjaśnialności (*explainability*) dostosowanej do poziomu wiedzy zainteresowanych stron w sytuacjach, gdy system AI ma znaczący wpływ na życie ludzi; jawnego komunikowania informacji, że użytkownicy mają do czynienia z systemem AI.

- **Różnorodność, niedyskryminacja i sprawiedliwość** (*diversity, non-discrimination and fairness*) – odwołuje się do zasady sprawiedliwości i obejmuje m.in.: unikanie niesprawiedliwej stronniczości (*bias*) podczas szkolenia i użytkowania systemów AI; zapewnianie różnorodności opinii poprzez włączenie do projektowania AI osób reprezentujących różne środowiska, kultury i dziedziny; udział interesariuszy, na których dany system miałby bezpośredni lub pośredni wpływ; zabieganie o regularne informacje zwrotne od użytkowników po uruchomieniu systemu.

- **Dobrostan społeczny i środowiskowy** (*societal and environmental wellbeing*) – odwołuje się do zasady sprawiedliwości oraz zasady zapobiegania szkodom i wskazuje m.in. na potrzebę: tworzenia zrównoważonej i przyjaznej dla środowiska sztucznej inteligencji poprzez kontrolę rozwoju, wdrażania i wykorzystywania systemów AI (np. kwestia zużycia energii podczas szkolenia systemu), a także całego łańcucha dostaw; monitorowania i uwzględniania przy projektowaniu, wdrażaniu i użytkowaniu systemów AI ich oddziaływania na społeczeństwo w różnych obszarach, w tym na instytucje demokratycznego państwa.

- **Odpowiedzialność** (*accountability*) – odwołuje się do zasady sprawiedliwości i jest uzupełnieniem wszystkich wcześniejszych wymogów. Dotyczy m.in. konieczności: zagwarantowania możliwości kontroli (*auditability*) algorytmów, danych i procesów projektowych; przeprowadzania oceny oddziaływania (*impact assessment*) systemów AI zarówno przed ich opracowaniem, wdrożeniem i wykorzystywaniem, jak i w trakcie; zapewnienia mechanizmu zgłaszania błędnych decyzji AI i reagowania na negatywne konsekwencje; wypracowania kompromisów między powyższymi zasadami w sytuacji ewentualnych konfliktów między nimi i oceny tych kompromisów ze względu na ryzyko dla zasad etycznych (gdy nie da się określić etycznie dopuszczalnych kompromisów, rozwijanie, wdrażanie i użytkowanie systemu AI powinno zostać odpowiednio zmodyfikowane).

Jak w praktyce stosować etykę AI

Ogólnie sformułowane zbiory zasad i wartości to punkt wyjścia do realizacji etyki sztucznej inteligencji. Oprócz nich potrzebne jest jednak także spełnienie dodatkowych wymogów: metodycznych oraz organizacyjnych.

Wymogi metodyczne to narzędzia, dzięki którym można zoperacjonalizować ogólne wytyczne dotyczące godnej zaufania AI. Chodzi przede wszystkim o podejście **Ethics by Design**, które umożliwia urzeczywistnianie poszczególnych wartości przy projektowaniu sztucznej inteligencji.

Oprócz metod służących projektowaniu AI zgodnie z wartościami istnieją rozwiązania pozwalające wdrażać i stosować sztuczną inteligencję w sposób zgodny z wytycznymi *trustworthy AI*. To ważne rozróżnienie, bo w obu przypadkach chodzi o odmienny typ odbiorców: w pierwszym o inżynierów czy organizacje rozwijające sztuczną inteligencję, w drugim – o podmioty ją stosujące[33].

Z kolei **wymogi organizacyjne** dotyczą włączania wartości leżących u podstaw koncepcji godnej zaufania AI w funkcjonowanie firm czy instytucji jako całości. Chodzi o zmianę perspektywy z wąsko pojmowanej etyki

zawodowej na etykę organizacji. Realizacja godnej zaufania AI nie ogranicza się bowiem do pojedynczych osób czy zespołów, ale ma ścisły związek z modelem działania całego biznesu czy administracji publicznej[34].

Ważne jest tu także uzupełnienie ogólnych zasad i wymogów etycznych o bardziej szczegółowe wskazania dla poszczególnych sektorów. W działania te powinny być zaangażowane organizacje zrzeszające przedsiębiorców. Pozwoli to na stworzenie instytucjonalnych mechanizmów bezpieczeństwa niezbędnych dla skutecznej samoregulacji (np. certyfikacje czy sądy branżowe)[35].

Ponadto istotne jest dostrzeganie oddziaływania etycznego w jego szerokim kontekście. Porządkującym punktem odniesienia może być tutaj Model „Wioski"™ Instytutu Humanites, który zwraca uwagę na potrzebę integracji wielu perspektyw w praktycznym myśleniu o etyce. Model ten wskazuje na znaczenie takich składowych społecznych, jak: rodzina, edukacja, biznes, świat kultury i mediów, a także na rolę instytucji samorządowych i państwowych.

Rys. 1: Model „Wioski" Rozwoju Ekosystemu Społecznego™, ©Zofia Dzik

Rys.2: Model Spójnego Przywództwa™, ©Zofia Dzik

Rozwijany w Instytucie Humanites model Spójnego Przywództwa™[36] podkreśla złożoność człowieka oraz wielowymiarowość bycia liderem – funkcji nazbyt często sprowadzanej do wąsko pojętej skuteczności oraz zawodowego autorytetu. Tymczasem najbardziej inspirującą cechą przywódcy jest autentyczność i integralność, czyli wierność deklarowanym wartościom. Kodeksy etyczne oraz narzędzia do realizacji wytycznych pozostaną bowiem bezużyteczne, jeśli decydenci nie wpiszą zawartych w nich wartości w DNA swoich organizacji i sami nie będą im wierni.

Podstawą etycznej transformacji organizacji i sektorów są jednak ludzie. Chodzi przede wszystkim o liderów, którzy nie tylko wskazują kierunki zmian, ale też odpowiadają za ich uwiarygodnienie w oczach pracowników i konsumentów. Dlatego istotną rolę w realizacji etyki AI odgrywa spójne przywództwo, czyli taki model działania, zgodnie z którym konsekwencja w realizacji wartości jest warunkiem budowania etycznej tożsamości liderów i organizacji.

Pierwsza i najbardziej znana metoda projektowania zgodnie z wartościami – Value Sensitive Design – powstała w latach 90. XX w. na Uniwersytecie Waszyngtońskim. Inne podejścia tego typu to Design for Values rozwijane na Politechnice w Delft czy właśnie Ethics by Design opracowane w ramach unijnych projektów SHERPA i SIENNA m.in. przez badaczy z Uniwersytetu Twente.

Ethics by Design, czyli projektowanie AI zgodnie z wartościami

Projektowanie sztucznej inteligencji zgodnie z wartościami polega na traktowaniu wytycznych określonych w kodeksach etyki AI na równi z innymi ogólnymi wymogami systemowymi, jak np. poprawność działania (*reliability*) czy bezpieczeństwo (*security*), i uwzględnianiu ich na każdym etapie projektowania systemu AI: począwszy od specyfikacji, przez szczegółowy rozwój oprogramowania, a skończywszy na testach i ewaluacji.

Opracowane w unijnych projektach badawczych SHERPA i SIENNA podejście Ethics by Design bazuje m.in. na zasadach i wymogach godnej zaufania AI[37]. To, co wytyczne UE nazywają wymogami, zostało tu określane jako wartości – jest ich sześć, a nie siedem (jak w unijnym dokumencie), ponieważ odpowiedzialność połączono z nadzorczą rolą człowieka.

Każdej z tych sześciu wartości przypisano określone warunki (*requisites*), służące ich urzeczywistnieniu. Mogą się one przejawiać przez określone funkcjonalności, struktury danych czy procesy budowy systemu itp. Na przykład taka wartość jak sprawiedliwość wymaga spełnienia następujących warunków: unikania stronniczości algorytmicznej, zapewnienia ogólnej dostępności różnym użytkownikom oraz sprawiedliwego oddziaływania.

Aby móc spełnić te warunki, należy je zmapować z poszczególnymi etapami realizacji projektu AI, a następnie włączyć do konkretnych metod zarządzania projektami informatycznymi (np. Agile).

Autorzy podejścia Ethics by Design zaznaczają, że choć zostało ono opracowane z myślą o inżynierach niedysponujących specjalistyczną wiedzą z zakresu etyki, to jego poprawne zastosowanie może wymagać włączenia etyków do zespołów projektujących AI[38].

AI Act – unijny projekt regulacji sztucznej inteligencji

Jak pamiętamy, trzecim filarem godnej zaufania sztucznej inteligencji – obok etyki i solidności – jest zgodność z prawem. W unijnym projekcie rozporządzenia dotyczącym AI (AI Act) za główne kryterium oceny systemów sztucznej inteligencji przyjęto stopień ryzyka.

W rezultacie powstała następująca klasyfikacja:

- **systemy o niedopuszczalnym ryzyku** – zakazane wykorzystanie AI
- **systemy o wysokim ryzyku** – regulowane stosowanie AI
- **systemy o innym stopniu ryzyka**[39] – wymóg przejrzystości AI (w określonych przypadkach) oraz rekomendacja dobrowolnego przyjmowania kodeksów postępowania (*codes of conduct*) przez dostawców AI

Projekt AI Act zawiera również opis instytucjonalnego ekosystemu kontroli sztucznej inteligencji. Na jego czele znajdzie się Europejska Rada ds. AI, która będzie koordynować prace krajowych organów nadzorczych. Te z kolei mają za zadanie kontrolować i notyfikować podmioty przeprowadzające audyty wśród dostawców i operatorów AI oraz wydające certyfikaty zgodności.

Kategoria niedopuszczalnego ryzyka obejmuje m.in. systemy AI, które:

- stosują techniki podprogowe w celu manipulacji ludzkim zachowaniem,
- wykorzystują dowolne słabości określonej grupy osób ze względu na wiek czy niepełnosprawność,
- służą do przeprowadzania klasyfikacji społecznej (*social scoring*) na użytek władz,
- są używane do rozpoznawania twarzy w miejscach publicznych w czasie rzeczywistym na potrzeby egzekwowania prawa.

Z kolei systemy AI o wysokim ryzyku podzielono na dwie grupy:

- AI stanowiąca produkt (bądź jego istotną część składową) objęty oddzielną regulacją dotyczącą zdrowia bądź bezpieczeństwa, jak np. samoloty, samochody, urządzenia medyczne, windy czy zabawki;
- Osobne (*stand-alone*) systemy AI stosowane w ośmiu wrażliwych obszarach: identyfikacji biometrycznej (np. na granicy); zarządzania infrastrukturą krytyczną; edukacji; zatrudnienia i zarządzania pracą; usług publicznych i prywatnych wymagających zapewnienia prywatności; egzekucji prawa; migracji, uchodźstwa, ruchu granicznego; administracji w obszarze sektora sprawiedliwości i procesów demokratycznych.

Systemy wysokiego ryzyka należące do obu tych grup muszą spełniać określone w projekcie rozporządzenia zasadnicze wymogi:

1. Być wyposażone w system zarządzania ryzykiem utrzymywany przez cały cykl życia systemu
2. Zapewniać właściwe mechanizmy zarządzania danymi
3. Posiadać dokumentację techniczną potwierdzającą zgodność systemu z wymogami AI Act
4. Być wyposażone w rejestry zdarzeń (logi) umożliwiające monitorowanie ich funkcjonowania
5. Być zaprojektowane w sposób transparentny i pozwalający użytkownikom na poprawną interpretację ich działania
6. Zapewnić możliwość nadzoru przez człowieka
7. Gwarantować odpowiedni poziom dokładności, solidności i cyberbezpieczeństwa

Co istotne, wymogi zawarte w AI Act pokrywają się w dużym stopniu[40] z wymogami (wartościami) przyjętymi w unijnych wytycznych dotyczących godnej zaufania sztucznej inteligencji oraz w podejściu Ethics by Design. Weryfikacji, czy systemy AI je spełniają, mają służyć wewnętrzne bądź zewnętrzne audyty o różnym poziomie szczegółowości (w zależności od kontekstu działania systemu AI wysokiego ryzyka).

Projekt rozporządzenia przewiduje także wydawanie certyfikatów zgodności dla firm czy instytucji, które pozytywnie przejdą audyt. Takie kontrole muszą być jednak cyklicznie ponawiane ze względu na dynamikę działania samouczących się systemów AI.

Prawo nie zastąpi etyki

Zbieżność wymogów zapisanych w AI Act z tymi, które zawarto w rekomendacjach dotyczących *trustworthy AI*, budzi u niektórych osób wątpliwości, czy zajmowanie się etyką sztucznej inteligencji ma w tej sytuacji sens. Skoro bowiem podobne do etycznych wymogi będą ujęte w prawie regulującym użycie systemów AI wysokiego ryzyka, to zastosowanie etyki wydaje się sprowadzać wyłącznie

Uczestnicząca w unijnym projekcie SIENNA filozofka i etyczka Anaïs Rességuier zauważa, że **różne sposoby rozumienia etyki są dziś często ze sobą mylone**[41]. Mieszanie pojęć w tej dziedzinie jest niebezpieczne, bo może skutkować manipulacją lub niewłaściwym rozumieniem zadań, jakie stoją przed etyką AI. Z kolei Komisja Europejska przestrzega przed mechanicznym podejściem do wymogów etycznych. Tę przestrogę można śmiało odnieść także do realizacji zapisów zawartych w AI Act. Istnieje bowiem ryzyko, że dominacja legalistycznego podejścia doprowadzi tutaj do sformalizowanego traktowania zagadnień, które każdorazowo wymagają osobnego namysłu.

do systemów o niskim ryzyku, co do których AI Act rekomenduje dobrowolną samoregulację. Takie rozumowanie nie jest jednak słuszne.

W przypadku systemów o wysokim ryzyku zapewnienie zgodności z wymogami będzie się przecież wiązało z koniecznością przeprowadzania okresowych audytów. Chęć sprostania takim kontrolom (czyli innymi słowy: chęć tworzenia i wykorzystywania AI zgodnie z wartościami) wymusi więc posługiwanie się podczas projektowania, wdrażania i użytkowania systemów AI narzędziami umożliwiającymi rzetelną operacjonalizację tych wartości. A takie narzędzia powstają właśnie na gruncie etyki AI (np. omawiane wyżej podejście Ethics by Design).

Stworzenie regulacji jest zatem dopełnieniem systemu normatywnego i nie polega na zastąpieniu etyki czymś lepszym. Powstałe ramy prawne wzmocnią po prostu konieczność przestrzegania wymogów, których realizacja będzie nadal wymagała etycznych kompetencji.

Prawo może zatem „zastąpić" etykę wyłącznie, gdy rozumiemy ją w wąskim legalistycznym (kodeksowym) sensie – czyli wtedy, kiedy odgórna regulacja w pewnych obszarach zajmie miejsce dobrowolnie spełnianych wymogów godnej zaufania AI. W innym znaczeniu – operacyjnym (pragmatycznym) czy kontekstowym (krytycznym) – prawo tylko przyczyni się do wzrostu zapotrzebowania na fachową wiedzę etyczną.

Korzyści z wdrożenia etyki AI

Skoro prawo nie tylko nie anuluje wymogów zawartych w wytycznych dotyczących godnej zaufania AI, ale jeszcze je wzmocni, to z wdrażaniem etyki AI nie warto zwlekać do chwili, aż zaczną obowiązywać nowe przepisy. Do efektywnej realizacji koncepcji *trustworthy AI* potrzeba zmian na wielu szczeblach organizacji, a do tego niezbędny jest czas i dobrze przemyślane działania.

Na etykę sztucznej inteligencji warto przy tym patrzeć jak na szansę, a nie uciążliwy obowiązek. Już dziś spełnianie wymogów godnej zaufania AI jest rekomendowane składającym wnioski o granty w unijnym programie Horyzont Europa[42]. W przyszłości będzie to zapewne standardowe wymaganie. A zatem im szybciej znajdziemy się w awangardzie firm i instytucji podchodzących etycznie do AI,

Jednym z celów unijnego systemu normatywnego dotyczącego sztucznej inteligencji jest **stworzenie nowych globalnych standardów.** Firmy spoza obszaru europejskiego chcące oferować swoje usługi na Starym Kontynencie będą się musiały dostosować do obowiązujących w UE przepisów. Dlatego europejscy przedsiębiorcy, którzy zawczasu zajmą się wdrażaniem nowych wymogów, mogą zdobyć przewagę nad podmiotami z innych części świata. To również jeden z elementów unijnej polityki AI.

tym większą mamy szansę na stworzenie innowacyjnych rozwiązań, którymi rynek nie jest jeszcze nasycony.

Trzeba też pamiętać, że projektowanie i używanie AI zgodnie z etyką to także korzyści wizerunkowe. Badania dowodzą, że godnej zaufania AI chce dziś nie tylko coraz więcej użytkowników, ale także rosnąca liczba kadry zarządzającej wysokiego szczebla[43]. Jeśli tak myślą konsumenci i członkowie zarządu organizacji w krajach zaliczanych do największych gospodarek świata, to ignorowanie tych trendów osłabia nasz potencjał konkurencyjny w ogólnoświatowej grze rynkowej.

Etyka AI jest wreszcie coraz częściej łączona z ONZowskimi celami zrównoważonego rozwoju i z inwestowaniem ESG (Environmental, Social, and corporate Governance – odpowiedzialność środowisk wa i społeczna oraz ład korporacyjny)[44]. Ponieważ kwestie społecznej i środowiskowej odpowiedzialności przy tworzeniu i stosowaniu systemów AI są wpisane w unijne wytyczne, należy się spodziewać, że niebawem powstaną konkretne wskaźniki umożliwiające ustandaryzowaną ocenę w tym obszarze.

Etyka sztucznej inteligencji w Polsce

W polskiej polityce rozwoju sztucznej inteligencji[45] koncepcja godnej zaufania AI pojawia się wielokrotnie. Ważna rola przypada w tym kontekście administracji, która ma wyznaczać standardy wdrożeń (ze szczególnym uwzględnieniem zasad etyki sztucznej inteligencji) i w najbliższych latach powinna wypracować reguły przejrzystości, audytowania i odpowiedzialności za zastosowanie algorytmów w sektorze publicznym.

Dokument zakłada również, że do 2027 r. Polska dołączy do grona krajów najbardziej aktywnych

Polska nie należy obecnie do grona państw, w których prowadzone są intensywne badania nad odpowiedzialnym rozwojem technologii, zaś w programie szkolnym etyka wciąż zajmuje raczej marginalne miejsce. Trudno zatem spodziewać się, żeby w krótkim czasie uczniowie i studenci mogli zacząć zdobywać rzetelną wiedzę w zakresie etyki AI na każdym poziomie kształcenia. Aby pokonać te przeszkody, **niezbędne jest zapewnienie finansowania projektom badawczym, edukacyjnym i popularyzującym ten obszar wiedzy**. Z kolei dla rozwoju badań w dziedzinie etyki AI bardzo ważne jest też zaangażowanie przedsiębiorców do pilotaży testujących etyczne tworzenie i wdrażanie innowacji.

w wykorzystywaniu danych zgodnie z koncepcją *trustworthy AI*. Nad Wisłą powinien też powstać międzynarodowy Digital Innovation Hub specjalizujący się w godnej zaufania sztucznej inteligencji – odpowiedzialny rozwój AI ma stać się jednym z czynników sprzyjających inwestycjom zagranicznym w innowacyjne projekty w Polsce. Ponadto etyka AI powinna zostać włączona do strategii cyfrowej edukacji dla wszystkich poziomów nauczania.

Nadrzędnym celem tych działań jest stworzenie w Polsce takich rozwiązań z dziedziny AI, które z jednej strony zapewniałyby bezpieczeństwo ich użytkownikom (w tym osobom zależnym od decyzji podejmowanych przez algorytmy), a z drugiej przyczyniały się do rozwoju nowych umiejętności w społeczeństwie. Dlatego potrzeba jak najszybciej wypracować niezależne metody audytowania systemów AI i rozpocząć analizę etycznych skutków ich wdrożeń – szczególnie w obszarach etycznie wrażliwych, np. dotyczących sfery publicznej czy praw człowieka.

W 2021 r. powstało też Centrum Etyki Technologii[46]. Utworzono je w Instytucie Humanites z myślą o wdrażaniu etyki AI tak w wymiarze metodologicznym, jak i systemowym, oraz umożliwieniu współpracy interesariuszy z różnych środowisk – naukowych, biznesowych, rządowych i obywatelskich. CET to zarazem pierwszy w Polsce ośrodek orędujący za odpowiedzialnym tworzeniem i stosowaniem innowacji.

Dalsze wyzwania związane z etyką AI

Etyka sztucznej inteligencji to złożona dziedzina obejmująca m.in. kodyfikację zasad, przekładanie wartości na praktyczne działanie i krytyczną refleksję nad różnymi kontekstami AI (ekonomicznym, środowiskowym czy społeczno-politycznym). Jej względnie krótka historia oraz dynamika rozwoju samej sztucznej inteligencji sprawiają, że nie sposób traktować dotychczasowych dokonań na tym polu jako ostatecznych.

Przeciwnie: jesteśmy dopiero u początku drogi, zaś specyfika myślenia etycznego polega na tym, że nigdy nie dobiega ono końca. Nie oznacza to oczywiście, że tu i teraz nie możemy wypracować żadnych skutecznych rozwiązań. Chodzi po prostu o to, żebyśmy zawsze byli gotowi podjąć od nowa krytyczny namysł nad regułami kształtującymi nasze funkcjonowanie w cyfrowym świecie[47].

Taka refleksja jest dziś prowadzona przez wielu badaczy i dotyczy np. wpływu infrastruktury AI na globalne ekosystemy albo mechanizmów politycznych koniecznych do rozstrzygnięcia spornych kwestii związanych z funkcjonowaniem tej technologii w naszym wspólnym świecie[48]. Oba te kierunki badań pokazują, że myślenie w obszarze etyki AI zmierza ku problemom zbiorowości, także w wymiarze globalnym. Słychać tu echa filozofii Hansa Jonasa, który przed ponad 40 laty podkreślał,

że w zaawansowanej cywilizacji technicznej odpowiedzialność nie może być dłużej postrzegana przez pryzmat tradycyjnej, jednostkowej etyki, ale wymaga zbiorowego wysiłku, także na poziomie państw czy organizacji międzynarodowych.

Zdaniem filozofa Marka Coeckelbergha etyka AI stoi dziś przed trzema dużymi wyzwaniami[49]. Pierwsze dotyczy tempa zachodzących zmian (prędki rozwój technologii oraz postępujący kryzys klimatyczny) i pytania, czy będziemy mieli wystarczająco dużo czasu, aby skutecznie poradzić sobie z tymi problemami. Drugie odnosi się do różnorodności kultur i tradycji oraz kwestii dostosowania do nich rozwiązań z obszaru etyki AI. Trzecie zaś wiąże się z problemem władzy, czyli tego, kto powinien podejmować decyzje dotyczące stosowania AI w naszym wspólnym świecie (przedsiębiorcy? obywatele? politycy? gremia międzynarodowe?).

Każda z tych kwestii jest sama w sobie skomplikowana, jednak w zestawieniu z pozostałymi tworzy splot zagadnień domagających się maksymalnie odpowiedzialnych decyzji przy ich rozwiązywaniu. Tym bardziej powinno nam więc zależeć, by wiedza na temat odpowiedzialnego rozwoju i stosowania sztucznej inteligencji była możliwie obiektywna i wielowymiarowa.

Wyzwania związane z etyką AI można też podzielić wedle typu problemów, których one dotyczą.

Pierwsza grupa odnosi się do **operacjonalizacji** i **kontekstualizacji**, czyli wymiaru **praktycznego**: jak przekładać wartości etyki AI na konkretne działania adekwatne do okoliczności. Obejmuje takie zagadnienia, jak: metody, narzędzia i dobre praktyki wdrażania wymogów etyki sztucznej inteligencji, sposoby testowania ich realizacji, analiza kontekstowa, analiza ryzyka itp.

Druga grupa dotyczy **instytucjonalizacji**, czyli kwestii **formalnych**: w jaki sposób zapewnić zgodność z wymogami etyki AI. Uwzględnia się tutaj m.in. samoregulację, rozwiązania legislacyjne (np. projekt rozporządzenia AI Act), ale też raportowanie ESG.

AI wymienia się dziś często jako **przydatne narzędzie do walki z kryzysem klimatycznym**. Należy przy tym pamiętać, że sama AI także powinna być zrównoważona (*sustainable AI*), czyli opierać się na takich komponentach czy rozwiązaniach, które będą jak najbardziej zbieżne z celami zrównoważonego rozwoju. Coraz więcej badaczy dowodzi dziś, że przy ocenie oddziaływania tej technologii ważne jest uwzględnianie różnych kosztów zewnętrznych, jak np. wyzysk ludzi i degradacja środowiska przy wydobywaniu metali rzadkich niezbędnych do produkcji podzespołów AI czy ślad węglowy związany z dużą energochłonnością uczenia maszynowego[50].

Trzecia grupa wiąże się z **integracją** i ma charakter **systemowy**: dotyczy uwzględnienia różnych rodzajów oddziaływania AI w jednej całościowej analizie. Chodzi o to, by taka analiza wpływu AI na różne obszary nie była rozproszona między odmiennymi systemami ewaluacji (z których jeden byłby oparty np. na wytycznych *trustworthy AI*, a drugi na celach zrównoważonego rozwoju ONZ), ale dążyła do syntetycznego, inkluzywnego ujęcia.

Rys.3: Obszary problemowe związane z etyką AI

Jak zwiększać popyt na etykę AI

Skoro etyka AI wiąże się z tyloma globalnymi wyzwaniami, łatwo dojść do wniosku, że jako jednostki nie mamy tu wiele do zrobienia i powinniśmy raczej zdać się na decyzje ekspertów. Tym bardziej, że w kontekście współczesnych wyzwań cywilizacyjnych coraz głośniej piętnuje się dziś tzw. prywatyzowanie odpowiedzialności, czyli skupianie się na indywidualnych wyborach konsumenckich. Jeśli więc nasze prywatne decyzje dotyczące troski o środowisko czy korzystania z technologii mają niewielkie znaczenie w porównaniu z działaniami podejmowanymi przez rządy czy międzynarodowe korporacje, to po co w ogóle się wysilać, a na dodatek jeszcze brać na siebie winę, którą ponosimy w znikomym stopniu?

Taka fatalistyczna postawa jest jednak błędna. Niezgoda na zrzucanie na obywateli odpowiedzialności za globalne problemy nie oznacza, że mamy automatycznie przestać się nimi zajmować. Wręcz przeciwnie – powinno to skłaniać do jeszcze większego zainteresowania sprawami dotyczącymi nas wszystkich. Filozof Tomasz Markiewka przekonuje[51], że potrzebujemy dziś odnowy myślenia politycznego, które

Cyfrowe uzależnienia to jeden z symptomów kryzysu więzi międzyludzkich, do którego przyczyniają się też algorytmy AI każdego dnia kuszące nas coraz bardziej spersonalizowaną rozrywką. Między innymi w odpowiedzi na ten niepokojący trend Instytut Humanites rozwija **globalny ruch Dwie godziny dla Rodziny / Człowieka na rzecz bliskości oraz zmiany kultury pracy i życia**[52], który promuje twórcze spędzanie czasu z bliskimi z różnych pokoleń i podpowiada, jak dbać o relacje rodzinne, by wzbogacały one starszych i młodszych. Ruch oddziałuje na pracodawców w zakresie kształtowania kultury pracy służącej wielowymiarowemu rozwojowi, tak by ludzie mogli dobrze wykorzystać szanse niesione przez nowe technologie, a nie stawali się ich niewolnikami.

– wbrew obiegowym opiniom – nie sprowadza się do piętnowania poczynań tej czy innej partii, lecz przede wszystkim do większego zaangażowania w sprawy publiczne. I choć trudno tu stworzyć uniwersalne kompendium postępowania, to można wskazać kilka działań, które wszyscy możemy i powinniśmy podjąć.

Po pierwsze zatem, trzeba rozwijać w sobie krytyczne myślenie o technologii i nie sięgać bezrefleksyjnie po każdy dostępny gadżet czy usługę. Warto wpierw zastanowić się, czy rzeczywiście są nam one potrzebne, oraz dowiedzieć się więcej na temat mechanizmów ich działania (np. jakie dane udostępniamy, instalując konkretną aplikację, albo ile czasu poświęcamy na korzystanie z cyfrowych technologii kosztem innych aktywności). Taka postawa wzmacnia naszą świadomość obywatelską.

Po drugie, wspierać – a jeśli to możliwe – także angażować się w inicjatywy dążące do wywierania wpływu na decydentów, tak by ci zwracali większą uwagę na wymogi etycznego rozwoju i wykorzystywania technologii. Politycy są wrażliwi na sondaże, dlatego im więcej obywateli będzie oczekiwało od nich aktywności w tym zakresie, tym większa szansa, że się tym zajmą.

Po trzecie, podejmować decyzje wspierające odpowiedzialny rozwój technologii. Obejmuje to zarówno większe zainteresowanie samą tematyką etyki AI, jak i bardziej konkretne działania w sferze zawodowej i prywatnej. Osoba na stanowisku menedżerskim może np. zdecydować o przeprowadzeniu szkoleń dla pracowników na temat odpowiedzialnego rozwoju i użytkowania sztucznej inteligencji oraz rozpocząć pilotaż dotyczący włączenia Ethics by Design do procesu projektowego. Ktoś inny będzie po prostu dbał o tzw. cyfrową higienę, czyli dawał dobry przykład w rodzinie i w pracy, jak odpowiedzialnie korzystać z cyfrowych technologii (np. brak telefonów komórkowych przy wspólnych posiłkach). Takie pozornie błahe zachowania wpływają na kształtowanie naszych nawyków w wymiarze rodzinnym i towarzyskim.

Odpowiedzialność za etyczny rozwój technologii wiąże się też czasem z koniecznością podejmowania trudnych decyzji. Zofia Dzik wskazuje w tym kontekście na tzw. pozytywnych rebeliantów, czyli pracowników firm technologicznych, którzy nie wahali się przeciwstawić nieakceptowalnym praktykom nawet kosztem własnego komfortu czy ryzykując posadę. Takie oddolne działania są często niezbędne, by w organizacjach zaszły pozytywne zmiany. Dlatego potrzeba nam odwagi w obronie wartości, z którymi się utożsamiamy.

Przede wszystkim jednak nie powinniśmy przyjmować biernej postawę wobec technologii, które nas otaczają, i ulegać poczuciu niemocy. AI to wynalazek człowieka i możemy wpływać na to, żeby jej udział w naszym życiu był pozytywny. To właśnie główne zadanie etyki sztucznej inteligencji. Warto więc traktować ją nie jako abstrakcyjną wiedzę dla wtajemniczonych, ale dziedzinę, która dotyczy każdego z nas.

Polecane książki na temat etyki sztucznej inteligencji

Mark Coeckelbergh, *AI Ethics*, Cambridge and London 2020.

Mark Coeckelbergh, *Robot Ethics*, Cambridge and London 2022.

Kate Crawford, *Atlas of AI. Power, Politics, and the Planetary Costs of Artificial Intelligence*, New Haven and London 2021.

Peter Dauvergne, *AI in the Wild: Sustainability in the Age of Artificial Intelligence*, Cambridge and London 2020.

Virginia Dignum, *Responsible Artificial Intelligence. How to Develop and Use AI in a Responsible Way*, Cham 2019.

Ethics of Artificial Intelligence, ed. S.M. Liao, New York 2020.

Cathy O'Neil, *Weapons of Math Destruction. How Big Data Increases Inequality and Threatens Democracy,* [b.m.] 2016.

Henrik Skaug Sætra, *AI for the Sustainable Development Goals*, Boca Raton 2022.

Ben Shneiderman, *Human-Centered AI*, New York 2022.

Bernd Carsten Stahl, *Artificial Intelligence for a Better Future. An Ecosystem Perspective on the Ethics of AI and Emerging Digital Technologies*, Cham 2021.
https://link.springer.com/book/10.1007/978-3-030-69978-9

The Oxford Handbook of Ethics of AI, eds. M.D. Dubber, F. Pasquale, S. Das, New York 2020.

Shannon Vallor, *Technology and the Virtues. A Philosophical Guide to a Future Worth Wanting*, New York 2016.

Przypisy

[1] Por.: Z. Dzik, *Czy pandemia przyspieszy automatyzację i z jakim skutkiem?*, „Rzeczpospolita" 21.10.2020, https://www.rp.pl/gospodarka/art464341-czy-pandemia-przyspieszy-automatyzacje-i-z-jakim-skutkiem

[2] Por.: A. Rudnicka, D. Kaczorowska-Spychalska, M. Kulik, J. Reichel, 2020, *Digital ethics – polscy konsumenci wobec wyzwań etycznych związanych z rozwojem technologii. I Ogólnopolski Raport*, Łódź 2020 https://wydawnictwo.uni.lodz.pl/wp-content/uploads/2021/01/Digital_ethics_raport.pdf.
Omówienie raportu: M. Chojnowski, *Polacy i technologie: tylko przyszłość jest groźna* https://ethicstech.eu/polacy-i-technologie-tylko-przyszlosc-jest-grozna/

[3] Por.: https://inventory.algorithmwatch.org/

[4] Por. A. McAfee, E. Brynjolfsson, *The Second Machine Age. Work, Progress, and Prosperity in a Time of Brilliant Technologies,* New York and London 2016.

[5] Por.: L. Floridi, *Enveloping the world: the constraining success of smart technologies* [w:] *CEPE 2011*: *Crossing Boundaries. Ethics in Interdisciplinary and Intercultural Relation* https://coeckelbergh.files.wordpress.com/2015/03/48.pdf

[6] Por.: definicje zawarte w pracach: E. Alpaydin, *Machine Learning: The New AI*, Cambridge and London 2016; M. Coeckelbergh, *AI Ethics*, Cambridge and London 2020.

[7] Por.: A. Webb, *The Big Nine. How the Tech Titans & Their Thinking Machines Could Warp Humanity*, New York 2019.

[8] Por.: E. Kim, *Amazon Echo secretly recorded a family's conversation and sent it to a random person on their contact list*, CNBC May 24 2018 https://www.cnbc.com/2018/05/24/amazon-echo-recorded-conversation-sent-to-random-person-report.html

[9] Por.: Z. Dzik, J. Rubin, *Człowiek i sztuczna inteligencjia. Wybory i odpowiedzialność*, „Newsweek Psychologia" nr 2, kwiecień 2019 https://www.humanites.pl/wp-content/uploads/2019/05/newsweek_psychologia_2019_04_01_czlowiek_i_sztuczna_inteligencja__wybory_odpowiedzialnosci_pdf_k-1.pdf

[10] Por. ramka w sekcji Moda na etykę AI?

[11] Por.: K. Crawford, *Atlas of AI. Power, Politics, and the Planetary Costs of Artificial Intelligence*, New Haven and London 2021.

[12] Por.: E. Strubell, A. Ganesh, A. McCallum, *Energy and Policy Considerations for Deep Learning in NLP*.
https://arxiv.org/abs/1906.02243

[13] Przykładem takiej sytuacji jest niedawne stwierdzenie Blake Lemoine'a – inżyniera AI z Google – że zbudowany przezeń model sztucznej inteligencji LaMDA jest obdarzony świadomością. Zostało to zdementowane przez firmę jako nieuzasadnione.
Por.: https://www.theguardian.com/technology/2022/jul/23/google-fires-software-engineer-who-claims-ai-chatbot-is-sentient

[14] Bernd C. Stahl pisze o wzajemnych powiązaniach AI z innymi zaawansowanymi technologiami, jak nanotechnologia, biotechnologia, technologie informacyjne czy kognitywne. Traktuje też AI jako element systemów społeczno-technicznych (*socio-technical systems*), w których z etycznego punktu widzenia najważniejszy jest złożony splot interakcji technologii ze światem społecznym. Por.: B.C. Stahl, A*rtificial Intelligence for a Better Future. An Ecosystem Perspective on the Ethics of AI and Emerging Digital Technologies*, Cham 2021.
https://link.springer.com/book/10.1007/978-3-030-69978-9

[15] Por.: M. Coeckelbergh, *AI Ethics*, Cambridge and London 2020.

[16] Por.: B.C. Stahl et al., *Artificial intelligence for human flourishing – Beyond principles for machine learning*, „Journal of Business Research".
https://doi.org/10.1016/j.jbusres.2020.11.030

[17] Por.: M. Coeckelbergh, op. cit., s. 38.

[18] Por.: A. van Wynsberghe, S. Robbins, *Critiquing the Reasons for Making Artificial Moral Agents*, „Sci Eng Ethics" 25, 719–735 (2019).
https://doi.org/10.1007/s11948-018-0030-8

[19] Por. M. Chojnowski, *Maszyny przyszłości – moralne czy tylko bezpieczne?*
https://ethicstech.eu/maszyny-przyszlosci/
M. Chojnowski, *Moralne roboty? To zły pomysł.*
https://www.sztucznainteligencja.org.pl/moralne-roboty-to-zly-pomysl/

[20] Por.: B.C. Stahl, M. Chojnowski, *Bernd C. Stahl: W centrum etyki AI jest ludzka szczęśliwość*
https://ethicstech.eu/bernd-c-stahl-w-centrum-etyki-ai-jest-ludzka-szczesliwosc/

[21] Por.: K. Saja, *Etyka normatywna. Między konsekwencjalizmem a deontologią*, Kraków 2015.

[22] S. Lem, *Etyka technologii i technologia etyki* [w:] tegoż, *Dialogi* (*Dzieła*, t. XXXII), Warszawa 2010.

[23] Wyrażenie „etyka technologii" w niektórych kontekstach powinno być zastąpione „etyką techniki". Jednak w ostatnich latach w języku polskim – ku dezaprobacie niektórych ekspertów z obszaru nauk technicznych (por. np. Z. Łucki, *Proszę... nie mówmy „technologia" na technikę!* (https://www.cri.agh.edu.pl/bip/63/11_63.htm) – angielskie słowo technology jest najczęściej tłumaczone właśnie jako technologia (por. https://sjp.pwn.pl/poradnia/haslo/technika-i-technologia;6953.html). W tej postaci utrwala się również w polszczyźnie w powiązaniu z innymi wyrazami, stąd też etyka technologii.

[24] Por.: H. Jonas, *Zasada odpowiedzialności. Etyka dla cywilizacji technologicznej*, Kraków 1996.

[25] Por. np.: N. Bostrom, *Superintelligence. Paths, Dangers, Strategies*, Oxford 2016; M. Tegmark, *Life 3.0. Being human in the age of Artificial Intelligence*, [b.m.] 2017.

[26] Por.: A. Rességuier, M. Chojnowski, *Anaïs Rességuier: Podejście krytyczne do AI to ciągłe zaczynanie od nowa*
https://ethicstech.eu/anais-resseguier-podejscie-krytyczne-do-ai-to-ciagle-zaczynanie-od-nowa/

[27] Por.: M. Chojnowski, *Cybersmuta, czyli casus wachmistrza Flanderki*
https://www.sztucznainteligencja.org.pl/cybersmuta-casus-wachmistrza-flanderki/

[28] Por. L. Floridi, *Translating Principles into Practices of Digital Ethics: Five Risks of Being Unethical*

https://link.springer.com/article/10.1007/s13347-019-00354-x
M. Chojnowski, *Grzechy główne etycznej SI*
https://www.sztucznainteligencja.org.pl/grzechy-glowne-etycznej-si/

[29] Floridi mówi o zieleni i błękicie, mając na myśli odpowiednio kwestie ekologii i technologii: „W mojej ostatniej książce *The green and the blue* staram się opisać politykę w perspektywie gambitu technologicznego. Musimy połączyć zieleń, czyli przyjazną środowisku, zrównoważoną gospodarkę obiegu zamkniętego, z błękitem, czyli cyfrowymi technologiami, innowacjami czy silną sztuczną inteligencją, która może osiągnąć więcej mniejszym kosztem, na przykład w obszarze zasobów czy wydajnego i skutecznego rozwiązywania problemów". Cytat za: L. Floridi, M. Chojnowski, *Luciano Floridi: Mój plan to zieleń i błękit*.
https://www.sztucznainteligencja.org.pl/luciano-floridi-moj-plan-to-zielen-i-blekit/

[30] Por.: B. Mittelstadt, *Principles alone cannot guarantee ethical AI*.
https://arxiv.org/abs/1906.06668

[31] *Ethics guidelines for trustworthy AI*
https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai
– na stronie dostępna jest także polska wersja dokumentu.

[32] *Proposal for a regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain union legislative acts*
https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206
– na stronie dostępna jest także polska wersja dokumentu.

[33] Por.: P. Brey, M. Chojnowski, *Philip Brey: Ethics by Design to skuteczne podejście do AI*.
https://ethicstech.eu/philip-brey-ethics-by-design-to-skuteczne-podejscie-do-ai/

[34] Por.: B. Mittelstadt, op. cit.

[35] Por.: R. Sroka, M. Chojnowski, *Robert Sroka: Etyka AI wymaga konsensusu i partycypacji*
https://ethicstech.eu/robert-sroka-etyka-ai-wymaga-konsensusu-i-partycypacji/;
R. Sroka, *Nieodkryci przywódcy współczesnej etyki biznesu*, Warszawa 2021.

[36] Por.: Z. Dzik, *Spójne Przywództwo. Kompleksowy program rozwoju liderów i zespołów wygrywających w sferze prywatnej i zawodowej, gdzie biznes osiąga rezultaty a ludzie odnajdują poczucie sensu.*
https://www.humanites.pl/wp-content/uploads/2022/09/fundacja-humanites-broszura-2022update-web.pdf

[37] Por.: P. Brey, M. Chojnowski, op. cit.

[38] Ibid.

[39] W rozporządzeniu mowa jest z jednej strony o obowiązku przejrzystości w przypadku *pewnych systemów AI* („transparency obligations for certain AI systems"), z drugiej zaś o kodeksach postępowania dotyczących *systemów AI innych niż wysokiego ryzyka* („other than high-risk AI systems"). Por.: *Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS*.
https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206

[40] Por. V. Dignum, C. Muller, *Artificial Intelligence Act. Analysis & Recommendations*.
https://allai.nl/wp-content/uploads/2021/08/EU-Proposal-for-Artificial-Intelligence-Act-Analysis-and-Recommendations.pdf

[41] A. Rességuier, M. Chojnowski, *Anaïs Rességuier: Podejście krytyczne do AI to ciągłe zaczynanie od nowa*.
https://ethicstech.eu/anais-resseguier-podejscie-krytyczne-do-ai-to-ciagle-zaczynanie-od-nowa/

[42] Por.: B.C. Stahl, M. Chojnowski, op. cit.; P. Brey, M. Chojnowski, op. cit.

[43] Por. raport Capgemini *Why addressing ethical questions in AI will benefit organizations*.
https://www.capgemini.com/wp-content/uploads/2021/02/AI-in-Ethics_Web-1.pdf

[44] Por. H. Skaug Sætra, *AI for the Sustainable Development Goals*, Boca Raton 2022; M. Minkkinen et al., *What about investors? ESG analyses as tools for ethics-based AI auditing*, „AI & Soc" 2022.
https://doi.org/10.1007/s00146-022-01415-0

[45] Por.: *Polityka dla rozwoju sztucznej inteligencji w Polsce od roku 2020*.

https://isap.sejm.gov.pl/isap.nsf/download.xsp/WMP20210000023/O/M20210023.pdf

[46] https://ethicstech.eu/

[47] Por.: A. Rességuier, M. Chojnowski, op. cit.

[48] Por.: M. Coeckelbergh, *The Political Philosophy of AI. An Introduction*, Polity Press, Cambridge 2022; M. Coeckelbergh, *Green Leviathan or the Poetics of Political Liberty. Navigation Freedom in the Age of Climate Change and Artificial Intelligence*, Routledge, New York 2021; M. Chojnowski, *Wolność w erze SI*.
https://www.sztucznainteligencja.org.pl/wolnosc-w-erze-si/

[49] Por.: M. Coeckelbergh, *AI Ethics*, op. cit., ss. 142–144.

[50] Por.: K. Crawford, V. Joler, *Anatomy of an AI System: The Amazon Echo As An Anatomical Map of Human Labor, Data and Planetary Resources*, AI Now Institute and Share Lab.
https://anatomyof.ai
K. Crawford, *Atlas of AI*, op. cit.; P. Dauvergne, *AI in the wild. Sustainability in the age of artificial intelligence*; M. Chojnowski, *Szersze spojrzenie na etykę sztucznej inteligencji [Kate Crawford „Atlas of AI"]*.
https://ethicstech.eu/szersze-spojrzenie-na-etyke-sztucznej-inteligencji-kate-crawford-atlas-of-ai/
M. Chojnowski, *Trzecia fala etyki AI – w stronę zrównoważonej sztucznej inteligencji*
https://ethicstech.eu/trzecia-fala-etyki-ai-w-strone-zrownowazonej-sztucznej-inteligencji/

[51] Por.: M. Chojnowski, *Między naturą a polityką. Dobro wspólne w dobie kryzysu klimatycznego i cyfrowej transformacji.*
https://ethicstech.eu/miedzy-natura-a-polityka-dobro-wspolne-w-dobie-kryzysu-klimatycznego-i-cyfrowej-transformacji/

[52] https://2godzinydlarodziny.pl/idea/

O autorze

Maciej Chojnowski

Współtwórca i dyrektor programowy Centrum Etyki Technologii Instytutu Humanites – pierwszego w Polsce ośrodka działającego na rzecz odpowiedzialnego rozwoju innowacji. Członek Grupy Roboczej ds. AI w KPRM, w której prowadzi projekt poświęcony operacjonalizacji zasad etyki AI wg modelu Ethics by Design.

Autor wielu artykułów na temat etyki sztucznej inteligencji i technologii cyfrowych. Był redaktorem pierwszego polskiego niekomercyjnego portalu poświęconego AI – sztucznainteligencja.org.pl. Przeprowadził kilkadziesiąt wywiadów z krajowymi i zagranicznymi ekspertami w dziedzinie inżynierii IT, filozofii technologii i prawa.

Zajmował się komunikacją w obszarze nauki i ICT oraz otwartym dostępem do danych i wyników badań naukowych. Ukończył filologię polską na UW. W przeszłości wykładowca i nauczyciel.

Centrum Etyki Technologii

CENTRUM ETYKI TECHNOLOGII
INSTYTUTU HUMANITES

Centrum Etyki Technologii to, powołany w Instytucie Humanites, ośrodek orędujący za innowacjami przyjaznymi człowiekowi i światu.

Popularyzujemy standardy i dobre praktyki dotyczące odpowiedzialnego rozwoju technologii. Podkreślamy znaczenie właściwej edukacji, tak by przyszli inżynierowie i przedsiębiorcy rozumieli wyzwania etyczne towarzyszące ich pracy, a użytkownicy chcieli współtworzyć bezpieczne rozwiązania.

KU ODPOWIEDZIALNYM INNOWACJOM

- Odpowiedzialne projektowanie technologii
- Przestrzeń do debaty wokół etyki technologii
- Krytyczny namysł nad narracjami technologicznymi
- Działanie na rzecz odpowiedniej legislacji

Chcemy, aby innowacje w Polsce były korzystne dla całego społeczeństwa, dlatego promujemy integrację i współpracę różnych środowisk przy tworzeniu, wdrażaniu i wykorzystywaniu cyfrowych narzędzi.

Oceniamy wpływ technologii na życie indywidualne i społeczne w ich różnych obszarach. Myślimy krytycznie, ale kierujemy się zdrowym rozsądkiem. Przekonujemy, że ani hurraoptymistyczne, ani apokaliptyczne wizje przyszłości nie oddają trafnie problemów, z którymi w najbliższym czasie musimy się zmierzyć.

Jesteśmy przekonani, że postęp techniczny nie jest zdeterminowany przez siły znajdujące się poza naszą kontrolą i że możemy wpływać na jego kierunek. Sądzimy również, że o technologii trzeba myśleć w szerszym kontekście: ekonomicznym, społecznym i politycznym.

ethicstech.eu

Instytut Humanites

humanites
CZŁOWIEK I TECHNOLOGIA

Think&DO tank systemowego rozwoju kapitału społecznego.

Misją instytutu jest świadomy, wewnątrzsterowny, szczęśliwy, otwarty poznawczo, wytrwały i społecznie wrażliwy człowiek. Instytut łączy tematykę Człowieczeństwa i Technologii i systemowo wspiera transformację społeczną, w szczególności w dobie rewolucji technologicznej, tak aby każdy człowiek miał przestrzeń do rozwoju swojego potencjału w oparciu o zdrowe poczucie własnej wartości.

Działamy na rzecz budowy wartości, postaw i kompetencji społecznych, takich jak: branie odpowiedzialności, miłość, wewnętrzny kompas, ciekawość, otwartość na ciągłe uczenie się, zdolność adaptacji, krytyczne myślenie, wytrwałość, inteligencja emocjonalna, współpraca, przedsiębiorczość, wrażliwość społeczna.

Swoją misję realizujemy w duchu afrykańskiego powiedzenia, że *„Potrzeba całej wioski, aby wychować jednego człowieka”* i każdy z nas ma w tym swój udział, odpowiadając za to, aby każdego dnia być lepszą wersją siebie z wczoraj.

Nasze działania opieramy na Modelu „Wioski” Humanites™ Rozwoju Ekosystemu Społecznego, oraz na Spójnym Przywództwie™, które zakłada życie w zgodzie z podstawowymi wartościami, integrację życia zawodowego i prywatnego oraz wielowymiarowy rozwój człowieka w sferze: fizycznej, umysłowej, emocjonalnej i duchowej.

Instytut jako jeden z pierwszych sygnalizował wpływ metazjawisk, takich jak: kryzys więzi rodzinnych, samotność, infodemia czy algorytmizacja życia, na rozwój gospodarczy, motywację oraz zdrowie fizyczne i psychiczne ludzi. Od dekady działa systemowo na rzecz zrównoważonego rozwoju będąc pionierem Human Economy, wspierając rozwój człowieka wobec wyzwań, które niesie technologiczna i społeczna rewolucja.

humanites.pl

CENTRUM ETYKI
TECHNOLOGII
INSTYTUTU HUMANITES