Godna zaufania AI

Jak używać sztucznej inteligencji zgodnie z wytycznymi Komisji Europejskiej w zakresie etyki?

Opracowanie przygotowane przez Grupę Roboczą ds. Sztucznej Inteligencji

Podgrupa ds. etyki i prawa

Warszawa, styczeń 2025

Ministerstwo Cyfryzacji | Tu tworzymy przyszłość

GRAI GRUPA ROBOCZA DS. SZTUCZNEJ INTELIGENCJI

Spis treści

Praca zbiorowa pod redakcją Macieja Chojnowskiego.

Partnerzy społeczni raportu

Wsparcie organizacyjne ze strony Ministerstwa Cyfryzacji Zespół pracowników Departamentu Badań i Innowacji

Korekta językowa, skład i opracowanie graficzne: Zespół kreacji Centralnego Ośrodka Informatyki oraz Zespół Dostępności Cyfrowej

Audyt WCAG: Zespół dostępności cyfrowej Centralnego Ośrodka Informatyki

Autorzy

Ewelina Bogiel – prawnik, praktyk z kilkunastoletnim doświadczeniem w sektorze bankowym. Posiada międzynarodowy Certyfikat CAMS w zakresie AML/CFT. Na co dzień zarządza departamentem zgodności w instytucji finansowej. Ma bogate i wieloletnie doświadczenie w obszarze compliance. Specjalizuje się w prawie bankowym, prawie karnym, przepisach z obszaru usług płatniczych. Ma doświadczenie w zakresie tworzenia i automatyzacji procesów zarządzania zgodnością oraz ryzykiem, ich implementacji przy wykorzystaniu nowych technologii w tym SI jak również w obszarze wykonywania analiz prawnych dopuszczających do tego rodzaju wdrożenia i technologie dla sektora finansowego. Zajmuje się prawem nowych technologii oraz sztuczną inteligencją. Jest członkiem Grupy Roboczej ds. Sztucznej Inteligencji przy Ministerstwie Cyfryzacji. Prelegent na polskich oraz zagranicznych konferencjach poświęconych sektorowi finansowemu.

Maciej Chojnowski (redaktor niniejszego opracowania) – współtwórca i dyrektor programowy Centrum Etyki Technologii Instytutu Humanites. Autor pierwszej w języku polskim popularnonaukowej syntezy na temat odpowiedzialnego rozwoju AI pt. „Etyka sztucznej inteligencji. Wprowadzenie". Członek Grupy Roboczej ds. Sztucznej Inteligencji (GRAI) przy Ministerstwie Cyfryzacji, gdzie prowadzi projekt poświęcony praktycznej realizacji wymogów etyki AI zgodnie z podejściem Ethics by Design. Członek Komitetu Technicznego 338 ds. Sztucznej Inteligencji w PKN – komitetu wiodącego w zakresie współpracy z ISO/IEC JTC 1/SC 42 oraz CEN/CLC/JTC 21. Był redaktorem pierwszego niekomercyjnego portalu poświęconego AI i cyfrowym technologiom: sztucznainteligencja.org.pl, w którym zajmował się tematyką etyki i prawa AI. W centrum superkomputerowym ICM UW promował rozwój High Performance Computingu w Polsce i współtworzył serwis Otwarta Nauka popularyzujący idee otwartego dostępu do danych i badań naukowych.

Iwona Karkliniewska – ekspert w obszarze rozwoju innowacji i nowych technologii. Kierowała pracami międzynarodowego interdyscyplinarnego zespołu badawczego w Centre for AI and Digital Policy w obszarze analizy i ewaluacji wdrożenia wiarygodnych, przejrzystych polityk systemów sztucznej inteligencji na świecie. Jest współautorem raportu „AI Policy Index Ranks 2023", który co roku dostarcza kompleksową analizę etycznego i demokratycznego wdrażania AI. Doktorantka na Wydziale Zarządzania Uniwersytetu Warszawskiego. Jej obszarem zainteresowania są modele współpracy gospodarczo-

społecznej w Regionie Morza Bałtyckiego, tworzenie rynków cyfrowych i ekosystemów innowacji oraz gospodarki opartej o dane.

Alicja Kaszuba – adwokat, specjalista w zakresie compliance oraz społecznej odpowiedzialności biznesu. Specjalizuje się we wdrażaniu zasad etyki biznesu, CSR, a także narzędzi ESG oraz tworzeniu polityk i procedur compliance. Współpracuje z międzynarodowymi firmami, wspierając ich działania z zakresu etyki, praw człowieka, przeciwdziałania korupcji oraz oceny ryzyka w kontekście zagadnień etycznych. W obszarze sztucznej inteligencji zajmuje się kompleksowym doradztwem, obejmującym opracowywanie strategii oraz wdrażanie rozwiązań zgodnych z regulacjami unijnymi, a także wsparciem w procesie certyfikacji ISO 42001:2023. Posiada również doświadczenie w obszarach prawa związanych z ochroną danych osobowych, ochroną konkurencji, e-commerce oraz prawem umów. Brała udział w inicjatywach promujących odpowiedzialne, etyczne oraz zgodne z prawem wdrażanie systemów AI.

Dr hab. Robert Sroka – partner w funduszu Abris Capital, a także członek Katedry Przedsiębiorczości i Etyki w Biznesie Akademii Leona Koźmińskiego. Członek zarządu Sustainable Investment Forum Poland (POLSIF). Zasiada w Radzie CET – Centrum Etyki Technologii oraz ESG Committe Invest Europe. Zrealizował kilkadziesiąt projektów doradczych wspierając rady nadzorcze i zarządy we wdrażaniu programów etycznych, strategii zrównoważonego rozwoju, ESG i ładu korporacyjnego. Był członkiem Grupy Roboczej ds. Etyki i Prawa Sztucznej Inteligencji w Ministerstwie Cyfryzacji oraz Grupy Roboczej ds. Sztucznej Inteligencji przy KPRM. Autor artykułów i raportów oraz książek „Nieodkryci przywódcy współczesnej etyki biznesu" oraz „Etyka i prawa człowieka w biznesie – w poszukiwaniu metody". W 2021 roku Forbes Polska uhonorował go nagrodą główną w kategorii „Odpowiedzialny Kapitał" a The Drawdown przyznał nagrodę „ESG Professional of the Year" 2022. Absolwent University Cambridge Institute for Sustainable Leadrship. Jego najważniejsze artykuły i wypowiedzi można znaleźć na jego stronie internetowej.

Marzena Tyl – radca prawny, Approved Compliance Expert (ACE), Approved Whistleblowing Officer (AWBO), manager, certyfikowany mediator, trener, Chief Compliance Officer w Narodowym Centrum Badań i Rozwoju, ekspert w zakresie prawa nowych technologii. Koordynuje procesy związane z zarządzaniem zgodnością (compliance), ochroną danych osobowych, informacją publiczną. Tworzy procedury służące eliminowaniu niewłaściwych praktyk w organizacji. Założycielka kancelarii prawnej wspierającej klientów między innymi w zakresie wdrażania rozwiązań technologicznych opartych w szczególności o sztuczną inteligencję. Doradza w zakresie prawidłowego wykorzystywania innowacyjnych technologii mitygując ryzyka prawne i etyczne. Jako

pasjonatka nowych technologii aktywnie uczestniczy w pracach Grupy Roboczej ds. Sztucznej Inteligencji (GRAI) działającej przy Ministerstwie Cyfryzacji.

Od redaktora

Publikacja ta jest owocem prac zespołu ekspertek i ekspertów nad rekomendacjami w zakresie realizacji wymogów godnej zaufania sztucznej inteligencji (z ang. trustworthy AI) przez organizacje używające AI. Miałem przyjemność przewodniczyć tym pracom w podgrupie ds. etyki i prawa, która działa w ramach Grupy Roboczej ds. Sztucznej Inteligencji przy Ministerstwie Cyfryzacji.

Czytelnicy zainteresowani odpowiedzialnym rozwojem technologii znajdą tu zarówno rekonstrukcję kontekstu normatywnego potrzebną do właściwego zrozumienia koncepcji godnej zaufania AI, jak i charakterystykę podejścia pryncypalistycznego (kodeksowego) w etyce (do którego wypada zaliczyć unijne wytyczne w tym zakresie), a także omówienie poszczególnych wymogów trustworthy AI oraz sposobów operacjonalizacji tych wymogów w praktyce – zarówno w odniesieniu do organizacji rozwijających AI (podejście Ethics by Design), jak i tych, które jej używają (podejście Ethically aligned Implementation). Całość podsumowują zwięzłe rekomendacje dla administracji centralnej dotyczące wdrażania godnej zaufania sztucznej inteligencji.

Autorzy poszczególnych rozdziałów często odwołują się bezpośrednio do materiałów źródłowych. Były one przedmiotem naszych dyskusji podczas spotkań, które odbyły się w trakcie przygotowań tego opracowania. To przejaw zarówno uczciwości intelektualnej, jak i przekonania, że rzetelna synteza ma kluczowe znaczenie, by sprawnie pozyskiwać wiarygodne informacje z danej dziedziny. Jest to tym bardziej istotne w sytuacji, gdy chodzi o dziedzinę skomplikowaną, która ma konkretny wymiar praktyczny i wymaga dużej odpowiedzialności. A tak właśnie jest w przypadku sztucznej inteligencji.

Mamy zatem nadzieję, że opracowanie, które oddajemy w Państwa ręce, będzie pomocnym drogowskazem przy poszukiwaniu rozwiązań dopasowanych do potrzeb poszczególnych organizacji. Zawarte w tym zbiorze artykuły najlepiej czytać, mając w pobliżu materiały, do których się odnoszą. Dotyczy to szczególnie części praktycznej, gdzie mowa o wytycznych UE dotyczących godnej zaufania sztucznej inteligencji, jak i o podejściach, które umożliwiają ich wejście w życie (stworzonych w unijnych projektach SIENNA i SHERPA). Opracowanie to nie miało bowiem zastąpić owych materiałów źródłowych, ale pomóc w efektywnym wykorzystaniu zawartych w nich rozwiązań.

Należy także zaznaczyć, że nie jest ono wyczerpujące, choć staraliśmy się możliwie szeroko uwzględnić kluczowe konteksty koncepcji trustworthy AI. Przedstawienie całości dyskusji toczonych wokół odpowiedzialnego rozwoju i stosowania sztucznej inteligencji byłoby jednak niemożliwe ze względu na dużą liczbę prac stale publikowanych w tej dziedzinie. Ponadto mijałoby się z praktycznymi celami przyświecającymi tej publikacji. Jej głównym przedmiotem nie są również ani niedawno przyjęte unijne rozporządzenie dotyczące AI, czyli Akt w sprawie sztucznej inteligencji, ani standardy publikowane przez organizacje normalizacyjne w obszarze AI governance. To rozwiązanie nieprzypadkowe. Z jednej strony bowiem postanowiliśmy skupić się przede wszystkim na wymiarze

etycznym koncepcji trustworthy AI, z drugiej zaś staraliśmy się unikać powierzchowności, która byłaby nieunikniona, gdybyśmy zdecydowali się odnosić do treści norm chronionych prawem autorskim.

Wierzymy, że publikacja ta okaże się dla Państwa przydatna podczas urzeczywistniania wartości, zasad i wymogów dotyczących godnej zaufania AI. Zgodnie z intencjami twórców tej koncepcji ma ona sprawić, że europejskie innowacje, które wykorzystują sztuczną inteligencję, będą przyjazne człowiekowi i środowisku. Skorzystają na tym nie tylko obywatele, ale również przedsiębiorcy z UE, którzy dzięki godnym zaufania produktom będą mogli zdobywać przewagę konkurencyjną na światowym rynku. I tego właśnie Państwu oraz nam wszystkim życzymy!

Maciej Chojnowski

Fundamenty etyki człowieka wobec sztucznej inteligencji. Karta praw podstawowych jako źródło etycznego podejścia do AI w UE

Robert Sroka

W dyskusji o etyce sztucznej inteligencji ważna jest optyka. Warto pamiętać, że „etyka sztucznej inteligencji" oznacza „etykę człowieka wobec sztucznej inteligencji". Na obecnym etapie rozwoju AI to bowiem wciąż człowiek określa warunki (w tym możliwość zakończenia działania) i zasady funkcjonowania algorytmów. Wciąż mamy sprawczość i narzędzia do tego, aby sztuczna inteligencja wspierała dobro, opierając się na etyce, a nie zwiększała ilość zła, przyczyniając się do wzrostu cierpienia fizycznego i psychicznego.

Wielu przeraża wizja przyszłości ze sztuczną inteligencją. Są i tacy, którzy obserwując szybkość, z jaką rozwijana jest AI, twierdzą, że jesteśmy w „momencie Oppenheimera", który nawiązuje do doświadczeń twórcy bomby atomowej. Jest to moment, w którym zdajemy sobie sprawę, że nasze najlepsze intencje przyniosły tragiczne skutki. Moment, kiedy nasz produkt zostaje użyty w nieoczekiwany i niechciany sposób. Kiedy uświadamiamy sobie, że nawet jeśli nie jesteśmy rasistami, to nasze algorytmy się nimi stają, zaś oprogramowanie, które miało łączyć ludzi, doprowadza do ich plemiennej izolacji. Moment, w którym dostrzegamy, że społeczne i ekonomiczne mechanizmy kontroli i równowagi już nie działają[1].

Scenariuszy przyszłości jest wiele i nie wiemy, jak będzie ona wyglądała. Natomiast to, w jakim świecie i w oparciu o jakie wartości chcemy żyć, już zdefiniowaliśmy (przynajmniej w zachodnim kręgu kulturowym) i zapisaliśmy – między innymi w Karcie Praw Podstawowych Unii Europejskiej. I to właśnie zdefiniowane tam fundamenty są obecnie wdrażane do dokumentów szczegółowych wyznaczających etyczne ramy działania AI. Jeśli uda się je z żelazną dyscypliną i skutecznie egzekwować, to możemy być spokojni o przyszłość świata z AI. Na razie jednak, niestety, zawodzimy. Nie zawodzi AI, tylko ludzie decydujący o jej regulowaniu: twórcy, naukowcy, użytkownicy, którzy szukają różnych wymówek, takich jak: prawa kapitalizmu, konkurencyjność, brak odwagi, ego, brak wyobraźni, chęć zysku, przedkładanie wartości nad interes itd.

[1] Por. A. Cooper. The Oppenheimer Moment [dostęp: 07.11.2024]

Przyszłość zależy więc od tego, czy i jak ludzie wykorzystają etykę. Zgodnie z definicją francuskiego filozofa Paula Ricoeura etyka to zamiar osiągnięcia dobrego życia wraz z drugim człowiekiem i dla drugiego w sprawiedliwych instytucjach[2]. Fundamentalne pytanie brzmi dziś: Czy w ogóle chcemy do tego dążyć?

Jeśli jednak dążenie do życia dobrego ma być czymś więcej niż tylko indywidualnym spontanicznym poszukiwaniem szczęścia, to powinno być ono powiązane z szacunkiem dla uniwersalnych norm, takich jak te zapisane w Karcie Praw Podstawowych UE. Niestety, włączenie ich do bardziej szczegółowych dokumentów, takich jak projekt rozporządzenia AI Act, często napotyka na opór.

Mówią co etycznym rozwoju AI w kontekście filozofii Ricoeura, warto zwrócić uwagę jeszcze na dwa elementy. Po pierwsze, żeby tworzenie etycznie działających systemów sztucznej inteligencji mogło być skuteczne, potrzebne są odpowiednie instytucje. Z pewnym uproszczeniem można powiedzieć, że etyczną rolą instytucji jest dla Ricoeura dążenie do sprawiedliwości rozumianej jako zapewnienie każdemu tego, co mu się słusznie należy[3]. Efektywnie działające instytucje umożliwiają skuteczną realizację wartości, również tych zawartych w Karcie.

Po drugie, ważna jest edukacja etyczna. Pewnym ograniczeniem zasad ogólnych jest to, że nie uwzględniają rozmaitości przypadków szczegółowych, również tych, w których dochodzi do konfliktu obowiązków. Edukacja etyczna pozwala na rozwój mądrości praktycznej urzeczywistniającej się w sądzie moralnym wydawanym w konkretnej sytuacji, który jest wynikiem wciąż na nowo podejmowanej mediacji między uniwersalnością normy a szczegółowością konkretnego przypadku, oraz mediacji między normami ewentualnie konkurującymi[4]. Mądrość praktyczna, która pozwala nam na etyczne dążenie do dobrego życia i tworzenia AI, która będzie mu sprzyjać, wymaga znajomości wizji dobrego życia, jak i zasad uniwersalnych. Przejdźmy zatem do ich przedstawienia.

[2] P. Ricoeur, *O sobie samym jako o innym,* tł. B. Chełstowski, Wydawnictwo Naukowe PWN, Warszawa 2003, s. 284 - 285.

[3] P. Ricoeur, *O sobie samym jako o innym,* tł. B. Chełstowski, Wydawnictwo Naukowe PWN, Warszawa 2003, s. 322.

[4] M. Kowalska, *Wstęp. Dialektyka bycia sobą,* w: P. Ricoeur, *O sobie samym jako o innym,* tł. B. Chełstowski, Wydawnictwo Naukowe PWN, Warszawa 2003, s. XXIX.

Karta Praw Podstawowych Unii Europejskiej

Karta praw podstawowych Unii Europejskiej[5] to zbiór fundamentalnych praw człowieka i obowiązków obywatelskich podpisany w 2000 roku w Nicei, któremu moc wiążącą nadał Traktat Lizboński w 2009 roku. Traktat o Unii Europejskiej stanowi obecnie, że „Unia uznaje prawa, wolności i zasady określone w Karcie praw podstawowych Unii Europejskiej (…), która ma taką samą moc prawną jak Traktaty". W związku z tym Karta stanowi akt prawa pierwotnego Unii i jako taki akt służy za kryterium oceny ważności aktów prawa wtórnego UE oraz przepisów krajowych. Właśnie na tej podstawie znajduje swoje odzwierciedlenie w takich dokumentach, jak AI Act czy w Wytycznych w zakresie etyki dotyczących godnej zaufania sztucznej inteligencji.

W Preambule Karty czytamy, że świadoma swego duchowo-religijnego i moralnego dziedzictwa Unia jest zbudowana na niepodzielnych, powszechnych wartościach godności osoby ludzkiej, wolności, równości i solidarności, a także opiera się na zasadach demokracji i państwa prawa. Zatem już na samym początku został podkreślony fundament systemu wartości UE, jakim jest godność osoby ludzkiej, która może być realizowana w pełni w warunkach wolności, równości i solidarności. Dlatego nie jest zaskoczeniem, że najważniejsze dokumenty poświęcone etyce w AI podkreślają centralną rolę człowieka i jego dobro, a nie interes gospodarczy.

Karta jest podzielona na części. W pierwszej zatytułowanej „Godność" zapisane zostały takie wartości, zasady i prawa jak: godność człowieka; prawo do życia; prawo człowieka do integralności; zakaz tortur i nieludzkiego lub poniżającego traktowania lub karania; zakaz niewolnictwa i pracy przymusowej.

Część o tytule „Wolność" zawiera: prawo do wolności i bezpieczeństwa osobistego; poszanowanie życia prywatnego i rodzinnego; ochronę danych osobowych; prawo do zawarcia małżeństwa i prawo do założenia rodziny; wolość myśli, sumienia i religii; wolność wypowiedzi i informacji; wolność zgromadzania i stowarzyszania się; wolność nauki i sztuki; prawo do nauki; wolność wyboru zawodu i prawo do podejmowania pracy; wolność prowadzenia działalności gospodarczej; prawo własności; prawo do azylu; ochrona w przypadku usunięcia z terytorium państwa, wydalenia lub ekstradycji.

[5] *Karta praw podstawowych Unii Europejskiej*, (2016/C 202/02), Bruksela 2016.

Część poświęcona „Równości” obejmuje: równość wobec prawa; zakaz dyskryminacji; różnorodność kulturową, religijną i językową; równość kobiet i mężczyzn; prawa dziecka; prawa osób w podeszłym wieku; integrację osób niepełnosprawnych.

Dział zatytułowany „Solidarność” dotyczy następujących zagadnień: prawa pracowników do informacji i konsultacji w ramach przedsiębiorstwa; prawa do rokowań i działań zbiorowych; prawa dostępu do pośrednictwa pracy; ochrony w przypadku nieuzasadnionego zwolnienia; należytych i sprawiedliwych warunków pracy; zakazu pracy dzieci i ochrony młodocianych w pracy; życia rodzinnego i zawodowego; zabezpieczeń społecznych i pomocy społecznej; ochrony zdrowia; dostępu do usług świadczonych w ogólnym interesie gospodarczym; ochrony środowiska; ochrony konsumentów.

Karta zawiera również dział zatytułowany „Prawa obywatelskie”, w którym znajdują się: prawo głosowania i kandydowania w wyborach do Parlamentu Europejskiego; prawo głosowania i kandydowania w wyborach lokalnych; prawo do dobrej administracji; prawo dostępu do dokumentów; określenie funkcji Europejskiego Rzecznika Praw Obywatelskich; prawo petycji; swoboda przemieszczania się i pobytu; opieka dyplomatyczna i konsularna.

W części poświęconej „Wymiarowi sprawiedliwości” wskazano: prawo do skutecznego środka prawnego i dostępu do bezstronnego sądu; domniemanie niewinności i prawo do obrony; zasady legalności oraz proporcjonalności kar do czynów zabronionych pod groźbą kar; zakaz ponownego sądzenia lub karania w postępowaniu karnym za ten sam czyn zabroniony pod groźbą kary.

Zanim przejdę do wskazania relacji między prawami podstawowymi a dokumentami poświęconymi etyce AI, zwrócę uwagę na jeden z najnowszych dokumentów UE dotyczących wyzwań cyfrowych bazujący na prawach podstawowych, czyli opublikowaną w styczniu 2023 roku Europejską deklarację praw i zasad cyfrowych w cyfrowej dekadzie[6].

[6] Europejska deklaracja praw i zasad cyfrowych w cyfrowej dekadzie, (2023/C 23/01), Bruksela 2023 [dostęp 05.12.2024].

Europejska deklaracja praw i zasad cyfrowych w cyfrowej dekadzie

Deklaracja zwraca uwagę, że wraz z przyspieszeniem transformacji cyfrowej nadszedł czas, aby Unia Europejska określiła, w jaki sposób jej wartości i prawa podstawowe powinny być wdrażane w świecie internetowym. Autorzy dokumentu podkreślili bardzo wyraźnie, że transformacja cyfrowa nie powinna pociągać za sobą ograniczania praw. To, co jest nielegalne poza Internetem, pozostaje nielegalne online.

Deklaracja jest próbą zastosowania i uszczegółowienia zasad etycznych zawartych w Karcie praw podstawowych do zagadnień dotyczących transformacji cyfrowej. W dokumencie zwrócono uwagę głównie na takie obszary, jak: ochrona danych, prawo do prywatności, niedyskryminacja i równouprawnienie płci, a także zasady dotyczące ochrony konsumentów, neutralności technologicznej i neutralności sieci, wiarygodności i inkluzywności. Warto zwrócić uwagę, że w Deklaracji podkreślono konieczność wzmocnienia ochrony praw użytkowników w środowisku cyfrowym, a także ochrony praw pracowniczych i prawa do bycia offline.

Deklaracja jasno wskazuje, że nie ma sprzeczności między stosowaniem zasad etycznych a rozwojem gospodarczym. Zdecydowanie się z tym zgadzam. Dobrze rozumiana etyka, w tym etyki szczegółowe, takie jak etyka biznesu czy etyka AI, nie stanowi ograniczenia dla rozwoju gospodarczego. Jedyne, co etyka ogranicza, to cierpienie. Nawet jeśli powstaje wrażenie, że krótkoterminowo ogranicza ona rozwój gospodarczy, to przedłożenie wartości nad doraźne interesy wyzwala kreatywność i prowadzi do zrównoważonego rozwoju człowieka i społeczeństwa. Poświęcenie etyki na rzecz doraźnych interesów uderza w nas rykoszetem. Dla przykładu: za nieograniczoną wolność mediów społecznościowych, w zamian dostajemy podziały społeczne, uzależnienia i algorytmy prowadzące do chorób psychicznych, a nawet samobójstw dzieci i nastolatków. Za brak ograniczeń w działalności platform e-handlu, otrzymujemy w zamian wykorzystywanie osób starszych, łamanie prawa własności intelektualnej i wzrost handlu nielegalnymi i niebezpiecznymi towarami. Natomiast unijna wizja transformacji cyfrowej wyrażona w Deklaracji jest ukierunkowana na człowieka, wzmacnia pozycję jednostek i wspiera innowacyjne przedsiębiorstwa.

Autorzy Deklaracji podkreślają, że powinna ona służyć jako punkt odniesienia dla przedsiębiorstw i innych odpowiednich podmiotów przy opracowywaniu i wdrażaniu nowych technologii. W tym względzie ważne jest propagowanie badań naukowych i innowacji. Szczególną uwagę zwracają na małe i średnie przedsiębiorstwa oraz start-upy.

Jednym z celów Deklaracji jest wsparcie decydentów w refleksji nad wizją transformacji cyfrowej, w której najważniejszym elementem są ludzie. Etyczny świat cyfrowy ma wspierać solidarność i włączenie społeczne poprzez łączność, edukację cyfrową, szkolenia i umiejętności, uczciwe i sprawiedliwe warunki pracy, a także dostęp do cyfrowych usług publicznych online; podkreślać, jak ważna jest wolność wyboru w interakcji z algorytmami i systemami sztucznej inteligencji oraz w sprawiedliwym otoczeniu cyfrowym. Ponadto wspierać udział w cyfrowej przestrzeni publicznej; zwiększać bezpieczeństwo, ochronę i upodmiotowienie w środowisku cyfrowym, w szczególności w odniesieniu do dzieci i młodzieży, przy jednoczesnym zapewnieniu prywatności i indywidualnej kontroli nad danymi, a także promować zrównoważony rozwój.

Deklaracja odnosi się do szerokiego spektrum zagadnień etycznych w kontekście cyfrowego świata i podkreśla, że powinny one stanowić całościowe ramy, a nie być traktowane oddzielnie.

W Rozdziele II zatytułowanym „Solidarność i włączenie społeczne" autorzy Deklaracji zwracają uwagę, że transformacja cyfrowa powinna sprzyjać sprawiedliwemu i integracyjnemu społeczeństwu oraz sprawiedliwej i inkluzywnej gospodarce w UE. Zauważono w ten sposób istotny problem, jaki niosą ze sobą platformy społecznościowe, czyli szerzenie mowy nienawiści.

Autorzy Deklaracji zwracają także uwagę na przestrzeń gospodarczą i fakt, że związki zawodowe i organizacje pracodawców odgrywają ważną rolę w transformacji cyfrowej, w szczególności jeżeli chodzi o określanie uczciwych i sprawiedliwych warunków pracy, w tym w odniesieniu do korzystania w pracy z narzędzi cyfrowych. Od organizacji tych oczekuje się, że podejmą działania, które zapewnią:

- wszystkim pracownikom możliwość bycia offline i korzystania z gwarancji równowagi między życiem zawodowym a prywatnym w środowisku cyfrowym,
- aby w środowisku pracy narzędzia cyfrowe w żaden sposób nie zagrażały zdrowiu fizycznemu i psychicznemu pracowników,
- poszanowanie praw podstawowych pracowników w środowisku cyfrowym, w tym ich prawa do prywatności i prawa do zrzeszania się, prawa do rokowań i działań zbiorowych, a także ochrony przed bezprawnym i nieuzasadnionym nadzorem,

- aby wykorzystanie sztucznej inteligencji w miejscu pracy było przejrzyste i zgodne z podejściem opartym na ocenie ryzyka oraz aby przyjęto odpowiednie środki zapobiegawcze w celu utrzymania bezpiecznego i zdrowego środowiska pracy,
- aby w ważnych decyzjach mających wpływ na pracowników zagwarantowany był nadzór ze strony człowieka oraz aby pracownicy byli ogólnie informowani, że wchodzą w interakcję z systemami sztucznej inteligencji.

Oprócz wspomnianych wątków Deklaracja porusza zagadnienie swobody wyboru, w szczególności w interakcji z algorytmami i systemami sztucznej inteligencji. W dokumencie czytamy wprost, że sztuczna inteligencja powinna służyć jako narzędzie dla ludzi, a jej ostatecznym celem powinno być zwiększenie dobrostanu człowieka.

Dlatego Unia Europejska i kraje członkowskie zobowiązują się do:

- propagowania ukierunkowanych na człowieka, wiarygodnych i etycznych systemów sztucznej inteligencji na wszystkich etapach ich opracowywania, wdrażania i wykorzystywania, zgodnie z wartościami UE,
- zapewnienia odpowiedniego poziomu przejrzystości w zakresie korzystania z algorytmów i sztucznej inteligencji oraz wzmocnienia pozycji użytkowników przy korzystaniu z nich, a także zwiększenia poziomu ich poinformowania, kiedy wchodzą w interakcję z algorytmami i sztuczną inteligencją,
- zapewnienia, aby systemy algorytmiczne opierały się na odpowiednich zbiorach danych w celu uniknięcia dyskryminacji i umożliwienia sprawowania przez człowieka nadzoru nad wszystkimi wydarzeniami mającymi wpływ na bezpieczeństwo ludzi i ich prawa podstawowe,
- zapewnienia, aby technologie, takie jak sztuczna inteligencja, nie były wykorzystywane do narzucania ludziom wyborów, na przykład w odniesieniu do zdrowia, kształcenia, zatrudnienia i życia prywatnego,
- zapewnienia zabezpieczeń i podejmowania odpowiednich działań, w tym poprzez propagowanie wiarygodnych norm, w celu dopilnowania, aby sztuczna inteligencja i systemy cyfrowe były przez cały czas bezpieczne i wykorzystywane z pełnym poszanowaniem praw podstawowych,
- przyjęcia środków w celu zapewnienia, aby badania nad sztuczną inteligencją były zgodne z najwyższymi standardami etycznymi i odpowiednimi przepisami UE.

Warto zapoznać się z całą treścią Deklaracji, która prezentuje prawa podstawowe w specyficznym kontekście cyfryzacji, zwłaszcza że jest ona spójna z bardziej szczegółowymi wytycznymi w zakresie etyki sztucznej inteligencji.

AI Act oraz Wytyczne w zakresie etyki dotyczące godnej zaufania sztucznej inteligencji

W 2021 roku Komisja Europejska zaproponowała pierwsze unijne ramy legislacyjne dotyczące sztucznej inteligencji, tzw. AI Act. Projekt analizuje i klasyfikuje systemy AI według ryzyka, jakie stwarzają one dla użytkowników. W regulacji zakłada się, że różne poziomy ryzyka będą niosły ze sobą mniej lub więcej obowiązków. Parlament Europejski uzgodnił swoje stanowisko negocjacyjne 14 czerwca 2023 roku[7], po czym nastąpiła faza negocjacji z państwami UE w Radzie na temat ostatecznego kształtu ustawy.

W kontekście tego opracowania istotny jest fakt, że podstawą do oceny ryzyka w AI Act są właśnie prawa podstawowe zawarte w Karcie praw podstawowych.

W wersji AI Act z 2021 roku możemy przeczytać, że wykorzystywanie sztucznej inteligencji wraz z jej szczególnymi cechami (np. efekt czarnej skrzynki, złożoność, zależność od danych, autonomiczne zachowanie) może mieć negatywny wpływ na szereg praw zapisanych w Karcie praw podstawowych Unii Europejskiej. Dlatego regulacja ma na celu zapewnienie wysokiego poziomu ich ochrony i zmierza do uwzględnienia różnych źródeł ryzyka poprzez jasno określone podejście oparte na analizie ryzyka.

Dzięki zestawowi wymogów dotyczących godnej zaufania sztucznej inteligencji oraz proporcjonalnym obowiązkom nałożonym na wszystkich uczestników łańcucha wartości regulacja ma wzmocnić i będzie promowała ochronę praw zapisanych w Karcie, do których zaliczają się chociażby: prawo do godności człowieka (art. 1), poszanowanie życia prywatnego i ochrona danych osobowych (art. 7 i 8), niedyskryminacja (art. 21) oraz równość kobiet i mężczyzn (art. 23).

Warto zauważyć, że autorzy regulacji piszą wprost o nakładaniu pewnych ograniczeń na niektóre prawa podstawowe, takie jak wolność prowadzenia działalności gospodarczej (art. 16) oraz wolność sztuki i nauki (art. 13) w celu zapewnienia zgodności z nadrzędnym

[7] Bruksela 2023. [dostęp: 07.11.2024]

interesem publicznym, przejawiającym się w takich dziedzinach, jak zdrowie, bezpieczeństwo, ochrona konsumentów i ochrona innych praw podstawowych w przypadku opracowywania i stosowania technologii sztucznej inteligencji wysokiego ryzyka[8].

AI Act jest pierwszym, przełomowym prawem regulującym sztuczną inteligencję. Nie byłoby go jednak, gdyby nie wcześniejsza szczegółowa wykładnia praw podstawowych w kontekście sztucznej inteligencji, która znalazła się w Wytycznych w zakresie etyki dotyczących godnej zaufania sztucznej inteligencji[9] przygotowanych przez Grupę ekspertów wysokiego szczebla ds. sztucznej inteligencji.

Autorzy Wytycznych podkreślają, że prawa podstawowe stanowią najważniejszy element prawa międzynarodowego i unijnego w zakresie praw człowieka oraz leżą u podstaw możliwych do wyegzekwowania na drodze prawnej praw, które gwarantują traktaty i Karta praw podstawowych Unii Europejskiej. Ze względu na prawnie wiążący charakter praw podstawowych ich przestrzeganie wchodzi w zakres pierwszej cechy godnej zaufania sztucznej inteligencji – „zgodnej z prawem AI". Prawa podstawowe mogą być jednak rozumiane jako wyrażające również szczególne uprawnienia moralne wszystkich jednostek, które wynikają z faktu bycia człowiekiem, niezależnie od ich prawnie wiążącego statusu. W tym sensie wpisują się one również w drugą cechę godnej zaufania sztucznej inteligencji – „etyczną AI"[10] (por. artykuł Koncepcja godnej zaufania sztucznej inteligencji jako wzorzec normatywny dla rozwoju AI w Unii Europejskiej – założenia, cele, wymogi w tym zbiorze).

Podsumowanie

Immanuel Kant, jeden z ojców współczesnej etyki, zaproponował imperatyw, wedle którego mamy traktować człowieka jako cel sam w sobie, a nie tylko środek do celu. Podejmując decyzje, również w zakresie AI, zbyt często zapominamy o tej zasadzie. Wybierając inny cel godzimy w godność osoby ludzkiej.

Nie przez przypadek w Wytycznych w zakresie etyki dotyczących godnej zaufania sztucznej inteligencji mowa o takiej etyce sztucznej inteligencji, która ma być antropocentryczna.

[8] Dz. cyt., s. 14.

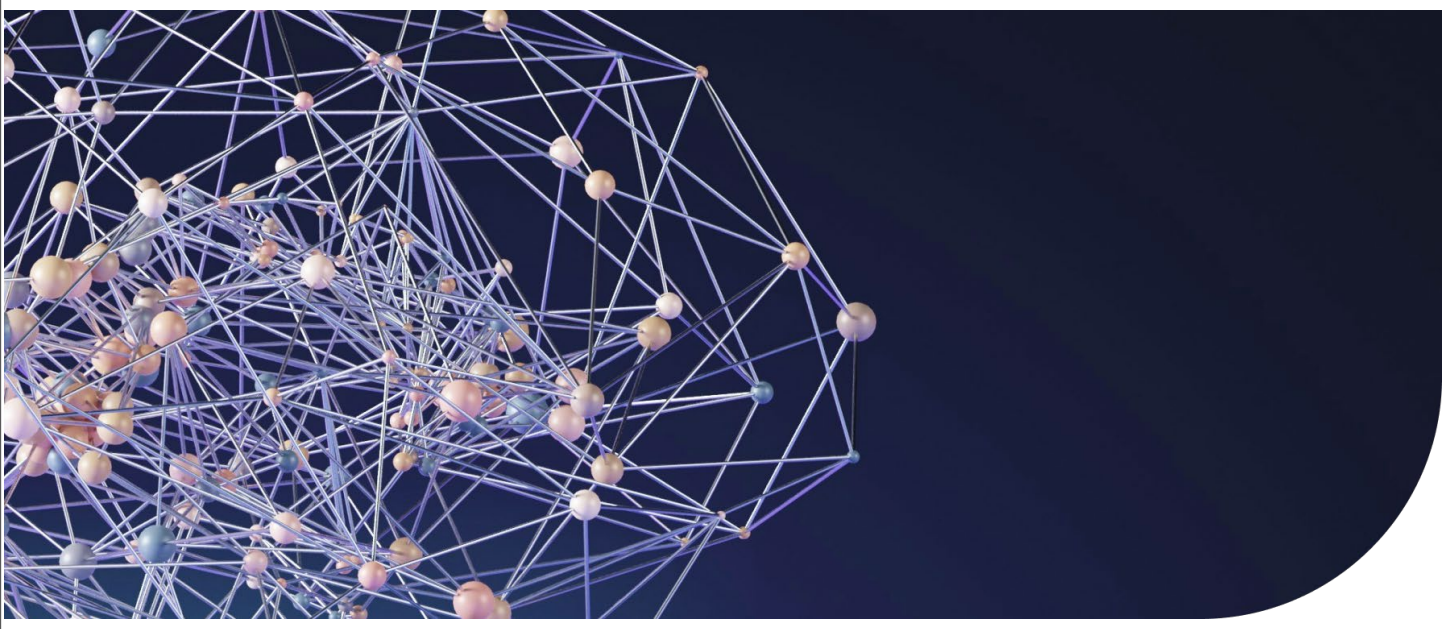
[9] Wytyczne w Zakresie Etyki Dotyczące Godnej Zaufania Sztucznej Inteligencji, Grupa ekspertów wysokiego szczebla ds. sztucznej inteligencji, Bruksela 2019.

[10] Dz. cyt., s. 9.

W Wytycznych możemy dopatrzeć się pewnej wykładni Kanta, gdy czytamy, że godność ludzka zakłada, że każdy człowiek ma „wrodzoną wartość", która w żadnym przypadku nie powinna być ograniczana, naruszana lub tłumiona przez inne osoby ani przez nowe technologie, np. modele czy systemy sztucznej inteligencji.

W kontekście AI poszanowanie godności ludzkiej oznacza, że wszyscy ludzie są traktowani z szacunkiem, jaki im się należy jako podmiotom moralnym, a nie jak przedmioty, które mogą być przesiewane, sortowane, oceniane, gromadzone lub manipulowane. Systemy sztucznej inteligencji należy zatem opracowywać w sposób, który wspiera oraz zapewnia poszanowanie i ochronę integralności cielesnej i psychicznej człowieka oraz tożsamości osobowej i kulturowej, a także gwarantuje zaspokojenie jego podstawowych potrzeb[11].

W etyce AI oraz w prawach podstawowych nie chodzi o sztuczną inteligencję, lecz o człowieka. W dyskusjach o etyce AI musimy więc bardziej skupić się na nas i naszej etyce oraz wyborach moralnych niż na samej sztucznej inteligencji. AI nie wyręczy nas w tym zadaniu. A z pewnością nie tak, jakbyśmy sobie tego życzyli. Etyka to przecież dążenie do realizacji naszej wizji dobrego życia.



[11] Dz. cyt., s. 13.

Etyka, regulacja i governance w systemie normatywnym dotyczącym sfery cyfrowej

Alicja Kaszuba

Nowoczesne technologie, w tym sztuczna inteligencja, wywierają wpływ na środowisko, w którym żyjemy i zmieniają je na dużą skalę. Zmiany przez nie wywołane są niezwykle złożone i dotyczą niemal każdego aspektu życia społecznego, ekonomicznego i kulturowego.

Jak zauważa Luciano Floridi[12], społeczeństwo funkcjonuje dziś w tzw. infosferze. Chodzi o przestrzeń, w której środowisko cyfrowe przenika się z rzeczywistością analogową. Floridi trafnie wskazuje, że moment wejścia w obszar infosfery jest niepowtarzalny – zdarza się tylko raz w historii. Charakter i jakość życia w infosferze zależą zatem od tego, w jaki sposób ją skonstruujemy. Z tego powodu na obecnym etapie rozwoju nowoczesnych technologii warto skierować uwagę na zaproponowane przez Floridiego trzy koncepcje: digital governance, digital regulation oraz digital ethics.

Digital governance obejmuje planowanie i opracowywanie stosownych polityk, wytycznych oraz strategii odpowiedzialnego rozwoju. Chodzi tu o szeroko pojęte zarządzanie zmianami zachodzącymi w ramach systemu normatywnego, który odnosi się do obszaru cyfrowego. Jego celem jest wyznaczenie kierunku rozwoju, w tym określenie obszarów, gdzie społeczność nie chce wykorzystywania nowoczesnych technologii.

W przeciwieństwie do postawy reaktywnej koncepcja digital governance ma umożliwić ludziom przywództwo (leadership) w obszarze nowoczesnych technologii. Jeśli sami wybierzemy kierunek rozwoju, będziemy mieli nad nim większą kontrolę. Narzędziami normatywnymi, które pomogą nam właściwie kształtować infosferę, są etyka cyfrowa (digital ethics) oraz regulacje dotyczące technologii cyfrowych (digital regulation).

Digital regulation to prawo stanowione, które precyzuje, co jest dopuszczalne, a co zakazane w sferze cyfrowej. Rządy muszą przyjąć właściwe podejście, aby regulacje nie stanowiły przeszkody dla rozwoju technologii, a raczej umożliwiały ich odpowiedzialny i bezpieczny rozwój. Z drugiej strony regulacje prawne powinny koncentrować się na

[12] L. Floridi, Soft Ethics and the Governance of the Digital, "Philosophy & Technology" vol. 31, s. 1–8 (2018).

właściwej identyfikacji praw jednostek w infosferze, a następnie – jeżeli zachodzi taka potrzeba – na wdrożeniu środków zapewniających ich odpowiednie stosowanie i zabezpieczenie.

Przykłady takich regulacji lub ich projektów obejmują: RODO, Dyrektywę o prywatności i komunikacji elektronicznej (2002/58/WE), AI Act, Dyrektywę w sprawie prawa autorskiego i pokrewnych praw na jednolitym rynku cyfrowym (2019/790), a także Europejską Deklarację Praw i Zasad Cyfrowych w dekadzie cyfrowej (por. artykuł Fundamenty etyki człowieka wobec sztucznej inteligencji. Karta praw podstawowych jako źródło etycznego podejścia do AI w UE w niniejszym zbiorze). Projekt dyrektywy dotyczącej odpowiedzialności za sztuczną inteligencję, Dyrektywa Cyfrowa (2019/771) oraz Akt w zakresie zarządzania danymi (2022/868) to kolejne przykłady regulacji wprowadzanych w obszarze technologii cyfrowych.

Wyzwaniem związanym ze stanowieniem prawa jest unikanie przeregulowania, które może utrudniać zrozumienie i przestrzeganie przepisów.

Digital ethics to z kolei według Floridiego dziedzina etyki, która bada i ocenia problemy moralne związane z nowymi technologiami, danymi i informacjami oraz sztuczną inteligencją. Jej celem jest formułowanie i wspieranie rozwiązań moralnie właściwych.

Etyka cyfrowa wpływa na tworzenie regulacji (digital regulation) i zarządzanie w sferze cyfrowej (digital governance). Regulacje określają, co jest legalne lub nielegalne, jednak nie dostarczają wskazówek (lub są one niewystarczające) na temat tego, jakie postępowanie jest najwłaściwsze oraz w jakim kierunku powinien zmierzać rozwój.

Etyka związana ze sztuczną inteligencją (AI Ethics), która jest istotną częścią digital ethics, znajdzie moim zdaniem zastosowanie w czterech głównych obszarach.

1. Będzie pomocna w podejmowaniu decyzji podczas projektowania lub aktualizacji systemów, analizując ich rozwiązania z perspektywy etycznej.
2. W fazie użytkowania umożliwi bieżącą ocena tych systemów pod kątem etycznym. Obejmuje to analizę potencjalnych skutków i ryzyk na podstawie danych, które nie były dostępne w fazie projektowania.
3. Będzie wskazywać wzorce etycznych zachowań, postępowania i podejmowania decyzji samym systemom AI.
4. Będzie wspomagać kształtowanie regulacji oraz kierunku rozwoju w dziedzinie technologii cyfrowych.

Digital ethics ma więc za zadanie wpływać na kształtowanie kierunku przyjmowanych regulacji, a zarazem uzupełniać te obszary regulacyjne, które wymagają dookreślenia. AI Ethics powinna skupić się na wielu istotnych obszarach, aby zagwarantować odpowiednie korzystanie z tej technologii. Poniżej przedstawię kilka z nich, w mojej ocenie najistotniejszych.

Przede wszystkim, należy zwrócić uwagę na kwestie związane z danymi, a szczególnie na dwa zagadnienia. Po pierwsze, na ochronę prywatności i danych osobowych osób fizycznych. Po drugie, na jakość danych.

Zapewnienie odpowiedniej ochrony prywatności i danych osobowych stanowi fundamentalny i niezwykle istotny wymóg. Zgodnie z tekstem art. 7 Karty praw podstawowych Unii Europejskiej, „każdy ma prawo do poszanowania swojego życia prywatnego i rodzinnego, domu oraz komunikacji". W dobie dynamicznego rozwoju technologii zapewnienie prywatności stało się bardziej skomplikowane niż kiedykolwiek wcześniej. W związku z tym istnieje realne zagrożenie dla praw człowieka. Twórcy systemów AI muszą zatem uwzględniać wartości godnej zaufania AI (trustworthy AI) (por. artykuł Koncepcja godnej zaufania sztucznej inteligencji jako wzorzec normatywny dla rozwoju AI w Unii Europejskiej – założenia, cele, wymogi w niniejszym zbiorze) i przestrzegać prawa oraz odpowiednich standardów, aby zapewnić poszanowanie prywatności każdego człowieka.

Z kolei jakość danych w kontekście uczenia i trenowania modeli sztucznej inteligencji odnosi się do stopnia ich dokładności, reprezentatywności i odpowiedniości dla określonego celu. Właściwa jakość danych jest jednym z kluczowych warunków osiągnięcia skuteczności, rzetelności i uczciwości przez systemy oparte na sztucznej inteligencji. Źródła danych muszą być zgodne z obowiązującymi regulacjami prawnymi tak, by zapewnić ich transparentne i legalne przetwarzanie. Ponadto dane muszą być adekwatne i reprezentatywne dla celu, jakiemu ma służyć konkretny system sztucznej inteligencji. Ważne jest unikanie błędów i zniekształceń w danych tak, aby zapobiec potencjalnej dyskryminacji lub błędnym wnioskom generowanym przez systemy oparte na sztucznej inteligencji.

Poza kwestiami związanymi z ochroną prywatności i jakością danych współcześnie napotykamy też na nowe wyzwania związane z przetwarzaniem danych o osobach fizycznych. Powstał problem z pozyskiwaniem kolejnych warstw informacji z już zebranych danych. Przykładem tego jest tworzenie tzw. profili jednostek, które na bazie analizy

istniejących danych stają się swoistym odzwierciedleniem ich zachowań, upodobań, antypatii i wielu innych, często niedostrzegalnych gołym okiem aspektów[13].

To zagadnienie prowadzi nas do głębszego problemu, który Shoshana Zuboff nazwała **kapitalizmem inwigilacji**. Wykorzystanie analizy behawioralnej może potencjalnie zagrozić autonomii człowieka: istnieje ryzyko, że poprzez takie działania można manipulować zachowaniami oraz przekonaniami ludzi. Przetwarzając ogromne ilości danych, firmy i instytucje są w stanie precyzyjnie kształtować reklamy, treści oraz interakcje, co może prowadzić do zmiany zachowań jednostek bez ich świadomej zgody.

W tym miejscu pragnę również zwrócić uwagę na istotny aspekt związany z pojęciem **architektury wyboru** w kontekście technologii cyfrowych. Jest to koncepcja z dziedziny ekonomii behawioralnej i nauk społecznych, która dotyczy sposobu, w jaki ludziom prezentowane są opcje wyboru oraz jak te prezentacje wpływają na ich decyzje.

Chodzi o swoiste projektowanie środowiska wyboru w taki sposób, aby skłonić ludzi do podejmowania określonych decyzji. Rozwiązania takie są stosowane na przykład przy projektowaniu wyglądu okienek zgody lub ustawień opcji domyślnej na stronach internetowych. Ich stosowanie powinno być przede wszystkim uczciwe, przejrzyste i zgodne z prawem oraz z wartościami etycznymi. Odpowiedzialne wykorzystanie architektury wyboru może przyczynić się do lepszego dopasowania usług do potrzeb ludzi, pod warunkiem że zostanie poddane nadzorowi etycznemu.

Kolejnym istotnym zagadnieniem dla AI ethics jest tzw. **emotion AI**, czyli sztuczna inteligencja emocji. Polega ona na próbie wyposażenia maszyn w zdolność rozpoznawania emocji wyrażanych zarówno w mowie, tekście, jak i obrazach oraz reagowania na nie. Emotion AI ma na celu tworzenie bardziej spersonalizowanych, ludzkich i efektywnych doświadczeń. Jednak pod tą fasadą obiecujących możliwości kryją się również zagrożenia.

Przede wszystkim istnieje problem prywatności, bezpieczeństwa oraz potencjalnej manipulacji i niekorzystnego wypływu na stan psychiczny człowieka. Systemy Emotion AI mogą dowiedzieć się znacznie więcej na temat naszych emocji, niż byśmy chcieli ujawnić. Refleksji powinno zostać również poddane wykorzystanie Emotion AI w procesach rekrutacyjnych, np. ocenianie kandydatów na podstawie „odczytanych" przez AI emocji, jeszcze zanim rekruter zdąży się zapoznać z nagranymi materiałami.

[13] Shoshana Zuboff, Wiek kapitalizmu inwigilacji, Zysk i S-k Wydawnictwo, Poznań 2019.

Problemem jest również sama interpretacja emocji przez AI. O ile naukowcy podejmują wiele wysiłków, aby algorytmy były skuteczne w rozpoznawaniu emocji, to systemy te uczą się na podstawie danych, które mogą zawierać uprzedzenia lub nierówności, wskutek czego algorytmy mogą powielać te same błędy. Problem braku uwzględnienia kontekstu kulturowego także rzuca cień na rozwijającą się dziedzinę Emotion AI. Różne kultury wyrażają bowiem emocje na różne sposoby, a ignorowanie tych różnic może prowadzić do niewłaściwych wyników.

W przypadku tzw. **generatywnej AI** ważne jest również rozważenie pod względem etycznym procesu tworzenia promptów, czyli instrukcji dla algorytmów, które mogą wpływać na ich działanie. Określenie, jakie cele ma osiągać dany algorytm, ma wpływ na to, jakie decyzje i rezultaty może generować.

Wszystkie te różnorodne aspekty muszą być dokładnie rozważane, analizowane i oceniane, aby zagwarantować właściwe i etyczne wykorzystanie sztucznej inteligencji we wszystkich dziedzinach jej zastosowań. Niemniej ograniczenie się do samej refleksji nie wystarcza – należy podjąć konkretne działania.

Jednym z praktycznych narzędzi jest **ocena wpływu** (ang. risk assessment) na etapie projektowania, dzięki czemu można odpowiednio wcześnie wykryć potencjalne zagrożenia systemu AI. Dodatkowo w fazie użytkowania ważne jest przeprowadzanie regularnych audytów, które oceniają działania systemów AI pod kątem zgodności z wartościami etycznymi i standardami.

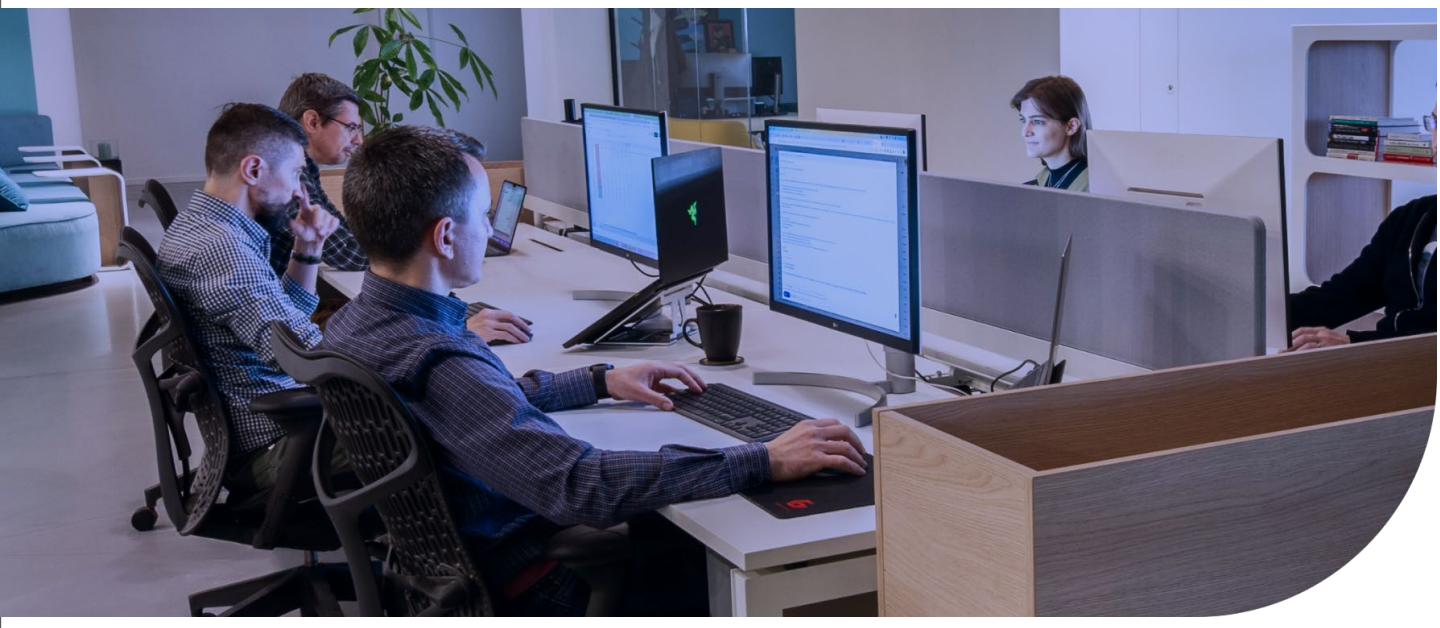
Ponadto niezmiernie istotne są kodeksy etyczne, które stanowią wytyczne i wskazówki dla podmiotów korzystających z technologii sztucznej inteligencji. Pełnią one również rolę deklaracji skierowanej do użytkowników produktów i informują ich o etyczności działań przedsiębiorstwa oraz wartościach, jakimi się ono kieruje.

Kolejnym kluczowym narzędziem będzie wprowadzenie procedury odwoławczej od decyzji podejmowanych przez systemy oparte na sztucznej inteligencji. Pozwoli to na weryfikację oraz na poprawienie decyzji zgodnie z ludzką perspektywą i wartościami wtedy, kiedy będzie to konieczne.

Podsumowując, trzeba podkreślić, że etyczne wykorzystanie nowoczesnych technologii to zadanie, które wymaga analizy, refleksji oraz działania. Szybkie tempo rozwoju technologicznego oraz wysoki poziom skomplikowania technologii sprawia, że same regulacje są niewystarczające. Należy zaprojektować infosferę w odpowiedzialny i humanistyczny sposób, możliwie sprawiedliwy dla całej społeczności globalnej.

Rozwój technologiczny powinien skłonić społeczeństwo do zastanowienia się nad priorytetowymi wartościami, a następnie wdrożenia ich w świecie technologii. Potrzeba nam efektywnego i odpowiedzialnego zarządzania i przewodnictwa. Ludzie nie powinni stracić swojego statusu ani zdolności do nadzorowania technologii. W centrum tego procesu powinno znajdować się dobro człowieka, a konkretnie ochrona jego autonomii, prywatności oraz prawa do danych.

Etyka technologii nie polega na uczeniu maszyn, co jest dobre, a co złe. Jest to niezbędny łącznik między władzą publiczną, przedsiębiorstwami, a konsumentami, zapewniający równowagę i uczciwe współistnienie. Przyjęcie etyki, jako konkretnej strategii, pomoże przedsiębiorstwom działać zgodnie z prawem i standardami etycznymi, jednocześnie przekonując konsumentów do swoich produktów, dzięki deklaracji etyczności.



Wartości, wymogi, kodeksy – podejście pryncypalistyczne do etyki AI w kontekście krytyki Brenta Mittelstadta

Ewelina Bogiel

W ostatnich latach etyka sztucznej inteligencji stała się ważnym tematem dyskusji w kręgach akademickich i politycznych na całym świecie. Zawarte w kodeksach etycznych wartości i wymogi mają służyć twórcom AI jako kompas moralny i pomagać im w procesie projektowania, tworzenia i wdrażania systemów sztucznej inteligencji. W wielu publicznych i prywatnych inicjatywach wypracowano ogólne wartości i zasady, którymi należy się kierować, by móc etycznie rozwijać i zarządzać AI.

Podejście to, określane mianem pryncypalizmu, opiera się na określonych zasadach i wymogach, które stanowią ramy do analizy i podejmowania decyzji etycznych. W kontekście AI zasady te mogą obejmować, np.:

- **autonomię:** poszanowanie praw użytkowników i ich zdolności do podejmowania decyzji,
- **nieczynienie krzywdy**: zapewnienie, że systemy AI nie szkodzą ludziom ani nie naruszają ich praw,
- **sprawiedliwość:** zapewnienie, że korzyści i obciążenia związane z AI są sprawiedliwie rozdzielane.

Amerykański filozof technologii Brent Mittelstadt podkreśla jednak, że same zasady nie są wystarczające do zapewnienia etycznego wykorzystania AI[14]. Zauważa on, że w wielu inicjatywach, dotyczących etyki sztucznej inteligencji wypracowano zestawy zasad, które przypominają cztery klasyczne zasady etyki medycznej. Jednak w porównaniu z medycyną w AI brakuje:

- wspólnych celów i obowiązków powierniczych,

[14] Por. B. Mittelstadt, Principles alone cannot guarantee ethical AI, „Nat Mach Intell" **1**, 501–507 (2019).

- historii i norm zawodowych,
- sprawdzonych metod przekładania zasad na praktykę,
- solidnych mechanizmów prawnych i odpowiedzialności zawodowej.

Przyjrzyjmy się po kolei tym kwestiom.

Cele i obowiązki

Medycynie przyświeca wspólny cel: promocja zdrowia i dobrego samopoczucia pacjenta, zaś lekarze mają moralny obowiązek bronić interesów pacjentów wbrew interesom instytucjonalnym.

Tymczasem w przypadku AI podstawowe cele deweloperów, użytkowników i innych zainteresowanych stron niekoniecznie są ze sobą zgodne. Trudno też wskazać tu jednoznacznie odpowiednik pacjenta, którego interesy miałyby pierwszeństwo przy podejmowaniu decyzji etycznych. Brak wspólnego celu utrudnia znalezienie równowagi między interesami publicznymi i prywatnymi.

Historia i normy zawodowe

Medycyna ugruntowała swoje normy w zbiorach zasad, które były tworzone i korygowane na przestrzeni stuleci, począwszy od przysięgi Hipokratesa, a skończywszy na deklaracjach genewskiej i helsińskiej czy kodeksie etyki lekarskiej Amerykańskiego Towarzystwa Medycznego, które stanowią podstawę podejmowania decyzji klinicznych i etyki badawczej. Istotne jest też licencjonowanie zawodu lekarza – dostęp do zawodu jest ograniczony przez uprawnienia wydawane osobom o wysokich kwalifikacjach, co ma na celu ochronę społeczeństwa.

Natomiast w przypadku AI trudno dziś mówić o porównywalnej historii dziedzinowej, jednorodnej kulturze czy tożsamości zawodowej oraz dobrze zdefiniowanych normach właściwego postępowania. Twórcy sztucznej inteligencji wywodzą się bowiem z różnych dyscyplin i środowisk, które mają odmienną historię, kulturę, struktury motywacyjne czy etos zawodowy.

Przekładanie zasad na praktykę

W przypadku medycyny skuteczne sposoby przekładania zobowiązań i zasad na praktyczne wymagania i normy dobrych praktyk są możliwe dzięki stowarzyszeniom, radom zawodowym, komisjom ds. oceny etyki, a także systemom akredytacji, licencjonowania i kodeksom postępowania. Tego typu mechanizmy wspierane instytucjonalnie pomagają w ocenie codziennej praktyki i trudnych przypadków pod względem etycznym.

Inaczej wygląda to w obszarze sztucznej inteligencji: tu metody przekładania zasad na praktykę w rzeczywistych kontekstach są dopiero testowane. Widać też trudności z zakorzenieniem wartości i zasad etycznych w projektowaniu technologii. Ponadto AI jest często rozwijana za zamkniętymi drzwiami – bez udziału interesariuszy. Zbieranie opinii przedstawicieli różnych środowisk, włączenie etyków do prac zespołów programistów oraz rozwiązywanie konfliktów między różnymi koncepcjami generuje dodatkową pracę i koszty.

Odpowiedzialność zawodowa i mechanizmy prawne

Mittelstadt wskazuje także, że medycyna podlega określonym ramom prawnym i zawodowym, w tym prawu dotyczącemu błędów w sztuce lekarskiej. Obejmują ją także wspomniane wyżej systemy licencjonowania i certyfikacji, jak również komisje etyczne oraz zawodowe. W rezultacie w sytuacji złamania zasad kodeksu bądź prawa lekarze są narażeni na sankcje.

Tymczasem w przypadku AI brakuje obecnie mechanizmów odpowiedzialności prawnej i zawodowej. Wykonywanie zawodu informatyka nie jest związane z ewentualnością nałożenia sankcji, które mogłyby mieć wpływ na utrzymanie zawodu. Z kolei kodeksy etyczne w tym obszarze bywają traktowane dość mechanicznie, na zasadzie listy kontrolnej, a nie jako część krytycznej, refleksyjnej praktyki.

Mając na względzie powyższe rozróżnienia, wypada podkreślić, że obecnie:

- tworzenie sztucznej inteligencji nie jest zawodem, którego cele muszą być w sformalizowany sposób zgodne z interesem publicznym,

- programiści nie podlegają historycznie potwierdzonym opisom tego, co to znaczy dobrze (etycznie) wykonywać swój zawód,
- w środowisku inżynierów AI metody przekładania zasad etyki na praktykę nie są dostatecznie rozpowszechnione.

Dlatego należy pamiętać, że w miarę jak technologia sztucznej inteligencji będzie się rozwijać, podejście do etyki AI będzie wymagało ciągłego dostosowywania i aktualizacji. Ważne jest, aby twórcy i użytkownicy AI byli świadomi zalet i ograniczeń każdego podejścia, by podejmować świadome i odpowiedzialne decyzje w dziedzinie AI.

Pryncypializm jest więc ważnym krokiem, aby stworzyć właściwe ramy dla etycznego rozwoju technologii, jednak musi zostać uzupełniony o dodatkowe elementy umożliwiające urzeczywistnienie zasad etycznych w praktyce (por. artykuł Wdrażanie etyki sztucznej inteligencji przy projektowaniu systemów AI z użyciem podejścia Ethics by Design – założenia i mechanizm działania w niniejszym zbiorze). W przeciwnym razie będzie mało użyteczny. W najgorszym razie może zostać wykorzystany do manipulacji wizerunkowych.

Organizacje wobec wyzwań etycznych związanych z AI

Iwona Karkliniewska

Rozwój sztucznej inteligencji rozpoczął się w latach 50. ubiegłego stulecia. Jednak dopiero w następnej dekadzie pojawiły się głosy, które rozpoczęły dyskusję związaną z etycznym aspektem i powszechnością zastosowania AI. W ciągu kolejnych dziesięcioleci postęp technologiczny i inwestycje w infrastrukturę ICT tylko przybierały na sile, wywierając ogromny wpływ na rozwój i wdrożenie systemów sztucznej inteligencji w gospodarce.

Obecnie systemy AI towarzyszą ludziom niemal w każdej sferze życia gospodarczego i społecznego. Wiele czynności, procesów i dziedzin tradycyjnej gospodarki przechodzi do domeny cyfrowej, ponieważ systemy AI stwarzają możliwość gromadzenia, analizowania i przetwarzania dużych zasobów danych. Dotyczy to również coraz bardziej danych osobowych, w tym danych wrażliwych i biometrycznych. Dlatego wśród decydentów i opinii publicznej powstała potrzeba szerszej dyskusji na temat obszarów zastosowania sztucznej inteligencji.

Istotnym jej elementem są kwestie etycznego podejścia do wykorzystywania systemów AI w różnych obszarach społeczno-gospodarczych. Konieczność uwzględnienia etycznych aspektów przy wykorzystywaniu technologii AI jest kluczowa i podyktowana w dużej mierze możliwościami, jakie stwarzają maszyny uczące się, które w przyszłości w wielu obszarach życia mogą stać się autonomiczne.

Dlatego w rozwoju i zastosowaniu technologii sztucznej inteligencji ważny jest element nadzoru ze strony człowieka. Powinno to zapewnić bezpieczeństwo, ochronę prywatności i poszanowania godności oraz podstawowych praw człowieka. Ma to szczególne znaczenie w przypadku wrażliwych grup społecznych oraz osób, które są wykluczone cyfrowo.

Kwestie etycznego wykorzystania systemów sztucznej inteligencji należy poruszać i wdrażać kompleksowo z uwzględnieniem roli i odpowiedzialności każdego uczestnika, w tym państwa oraz wszystkich podmiotów gospodarczych na rynku. Ponadto bardzo ważna jest też świadomość i wiedza społeczeństwa w tym zakresie

Etyka w rozwoju technologii

Etyczne podejście do rozwoju technologii sztucznej inteligencji jest procesem złożonym i wieloetapowym. W celu zrozumienia złożoności tego zagadnienia, w szczególności w przypadku technologii AI wprowadzonej do obrotu rynkowego, należy brać pod uwagę stopień oddziaływania danej technologii na człowieka i społeczeństwo. Bernd C. Stahl zauważa[15], że w dotychczas prowadzonych dyskusjach można wyróżnić trzy główne obszary problemowe, które dotyczą etycznego zastosowania AI.

Pierwszy dotyczy problemów związanych z konkretnymi technikami wykorzystywanymi w AI, jak na przykład uczenie maszynowe (ang. machine learning). Zwiększenie ilości przetwarzanych danych i powszechności wykorzystania uczenia maszynowego powoduje problem związany z wiarygodnością i przejrzystością tych rozwiązań. Ma to szczególne znaczenie w przypadku takich wrażliwych obszarów, jak np. diagnostyka kliniczna czy pomoc społeczna. Dane wykorzystywane do trenowania systemów uczenia maszynowego mogą posiadać tzw. bias, czyli skrzywienie polegające na niedostatecznej reprezentatywności, które może skutkować dyskryminacją różnych grup demograficznych (np. ze względu na płeć). Przykładowo algorytmy wytrenowane wyłącznie na podstawie danych wejściowych dotyczące jednej płci mogą znacząco zniekształcić obiektywizm danego zagadnienia. Dlatego modele AI oparte na uczeniu maszynowym wymagają szczegółowego audytu w obszarze stosowanych algorytmów oraz danych wejściowych i wyjściowych.

Drugi obszar problemowy związany z etyką sztucznej inteligencji skupia się wokół szeroko rozumianego funkcjonowania rynków cyfrowych. Należy tu uwzględnić czynniki wpływające na trendy i kierunek rozwoju społeczeństwa informacyjnego we współczesnym świecie. W tym kontekście należy brać pod uwagę m.in. następujące obszary:

- ekonomiczny,
- prawny,
- dotyczący wolności i praw człowieka.

[15] Por. B.C. Stahl, J. Antoniou, M. Ryan, et al. Organisational responses to the ethical issues of artificial intelligence, „AI & Society" 37, 23–37 (2022).

Na przykład automatyzacja procesów podejmowania decyzji może mieć w przyszłości wpływ na politykę zatrudnienia w poszczególnych sektorach, co wiąże się z koniecznością zmiany kompetencji zawodowych, a w efekcie przebranżowienia wiele grup społecznych na rynku pracy.

Z kolei w odniesieniu do aspektów prawnych kluczowe wydaje się zapewnienie odpowiedniej legislacji, która zagwarantuje skuteczne przestrzeganie praw podstawowych, a także praw w obszarach związanych z bezpieczeństwem, zdrowiem lub porządkiem publicznym. Zapewnienie skuteczności wykorzystywanych narzędzi i mechanizmów prawnych powinno zawierać wymóg przejrzystości, przede wszystkim w odniesieniu do grup szczególnie wrażliwych oraz osób wykluczonych społecznie. Szczególny nacisk należy również położyć na kwestie wolności człowieka w kontekście autonomiczności maszyn: może tu nastąpić zachwianie równowagi, które wpłynie negatywnie na jednostkę i jej zdrowie, w tym zdrowie psychiczne. Dobrym przykładem są tu np. pojazdy autonomiczne i zdarzające się wypadki drogowe z ich udziałem lub usługi opieki zdrowotnej świadczone bez udziału człowieka np. czatboty, wirtualni asystenci.

Trzeci obszar dyskusji wokół etyczności sztucznej inteligencji łączy elementy filozofii, religii, a niekiedy także koncepcji rodem z science fiction. Powszechne wykorzystanie systemów AI może stać się przyczyną zmian cywilizacyjnych i wpłynąć na sposób definiowania istoty samego człowieka. W dyskusji tej nierzadkie są podejścia skrajnie pesymistyczne, zakładające wysokie prawdopodobieństwo stopniowego wypierania inteligencji człowieka przez inteligencję maszyn w życiu społecznym. W tym obszarze możliwe są jednak także ujęcia bardziej pozytywne, dostrzegające możliwość większej zgodności miedzy rozwojem technologicznym a dobrostanem ludzi.

Przytoczone główne obszary problemowe etyki AI wraz z konkretnymi przykładami nie stanowią listy zamkniętej. Wraz ze wzrostem zaawansowania systemów AI mogą pojawić inne wątki i obszary wymagające analizy, na które należy zwrócić uwagę i włączyć do dyskusji w kwestii etyczności wykorzystywania sztucznej inteligencji.

Należy zatem stworzyć odpowiednie mechanizmy interakcji i oddziaływania technologii ze społeczeństwem, a także zasady wprowadzania systemów sztucznej inteligencji w obrót rynkowy oraz wypracować odpowiednie kodeksy postępowania i narzędzia monitorujące.

Strategie uwzględniania etyki w pracach nad AI

Wdrożenie wymogów etycznego wykorzystania AI wymaga uwzględnienia wszystkich potencjalnych obszarów i płaszczyzn, które zostały przytoczone powyżej lub mogą pojawić się w przyszłości w związku z wykorzystaniem systemów w obrocie rynkowym. Należy zatem wypracować odpowiednie zasady i mechanizmy pozwalających na włączenie etyki AI na każdym poziomie aktywności społeczno-gospodarczej, gdzie technologia ta ma zastosowanie. Spośród możliwych mechanizmów, które są obecnie przedmiotem dyskusji, można wymienić:

- polityki na poziomie krajowym, regionalnym i międzynarodowym,
- wytyczne, standardy, metody implementacji,
- mechanizmy z obszaru AI governance obejmujące np. CSR czy zarządzanie ryzykiem.

Dostosowywanie polityk i strategii krajowych powinno być oparte o rekomendacje organizacji międzynarodowych (np. KE, ONZ, UNESCO, OECD) (por. artykuł Koncepcja godnej zaufania sztucznej inteligencji jako wzorzec normatywny dla rozwoju AI w Unii Europejskiej – założenia, cele, wymogi w tym zbiorze), a także akty prawne wypracowane na poziomie unijnym, które obejmują między innymi takie obszary, jak ochrona danych osobowych, własność intelektualna, prawo konkurencyjności, odpowiedzialność cywilna za produkty i usługi czy ochrona praw podstawowych człowieka. Wypracowane narzędzia wymagają również stworzenia odpowiednich struktur instytucjonalnych na poziomie regionalnym i krajowym, które zapewnią prawidłowe funkcjonowanie mechanizmu monitoringu i ewaluacji, a także będą pełnić funkcje doradcze. Istotnym aspektem jest też włączanie interesariuszy na poziomie państwowym i unijnym w inicjatywy, których celem jest budowanie przyjaznego, przejrzystego i wiarygodnego ekosystemu cyfrowych technologii.

Dobrym przykładem takich działań jest tworzenie wyspecjalizowanych organizacji regulacyjnych, grup eksperckich składających się z przedstawicieli różnych branż, sektorów gospodarki, dziedzin naukowych oraz organizacji społecznych. Umożliwi to współpracę przedstawicieli różnych grup społecznych i wypracowanie kompleksowych wytycznych, rekomendacji, wymogów oraz kodeksów postępowania na rynku cyfrowym.

Szczególną uwagę należy też poświęcić kwestii wymagań technicznych dotyczących stosowanych rozwiązań AI. W tym przypadku należy wypracować jasne wytyczne i wymogi odnośnie do standaryzacji technologii włączanych w obrót rynkowy

Kwestie etyczne, dotyczące zastosowania technologii i algorytmów sztucznej inteligencji, powinny być uwzględniane już na etapie opracowania i testowania danego systemu. Można tu wykorzystać projektowanie zgodnie z wartościami (ang. ethics by design) (por. artykuł Wdrażanie etyki sztucznej inteligencji przy projektowaniu systemów AI z użyciem podejścia Ethics by Design – założenia i mechanizm działania w niniejszym zbiorze). **Dobrą praktyką powinno być prowadzenie audytów i oceny ryzyka na poziomie organizacji.**

Dobre praktyki etycznego wykorzystywania AI w organizacji

Zagadnienia związane z etyką AI są złożone i wymagają podejmowania działań dotyczących zarówno funkcjonowaniu samych systemów AI, jak i organizacji, w których te systemy są wytwarzane i używane.

Dobrą praktyką, którą wykorzystuje część organizacji, jest weryfikacja wdrożonych technologii i systemów AI pod względem takich aspektów, jak:

- przejrzystość,
- integralność,
- bezpieczeństwo,
- prywatność,
- minimalizacja przetwarzanych danych osobowych.

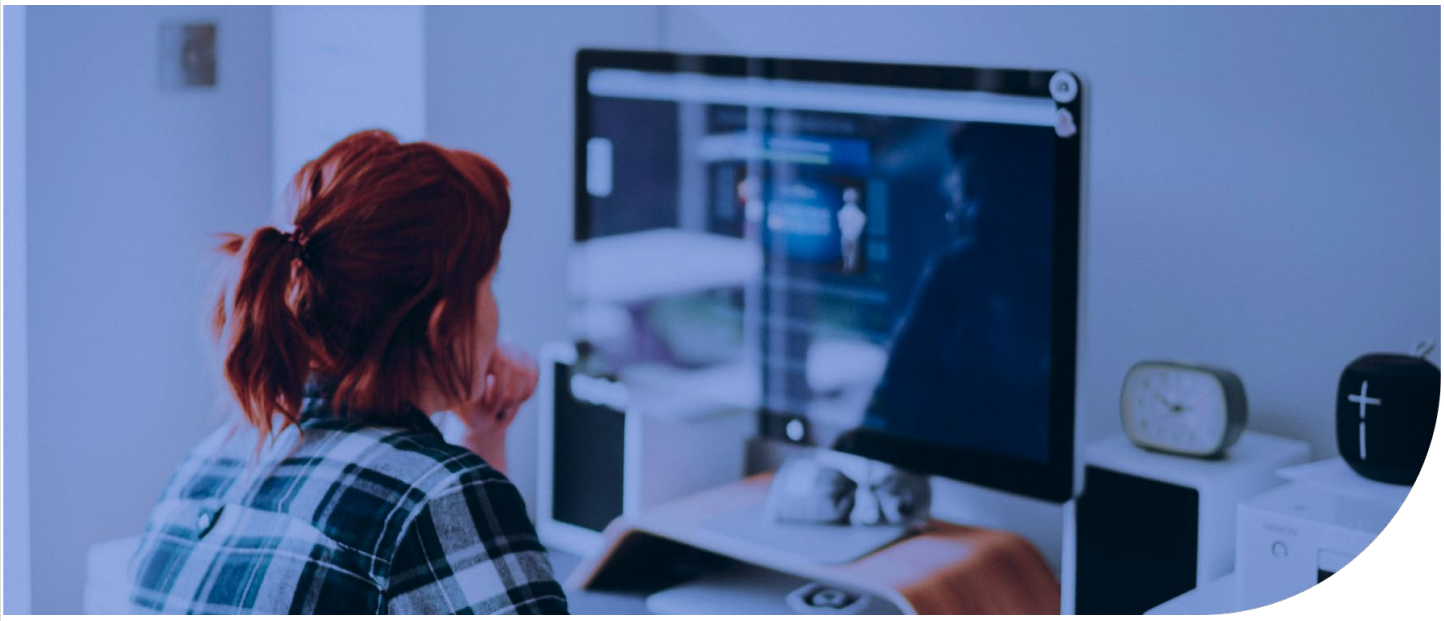
Organizacje przeprowadzają też analizę procesów biznesowych, które w sposób pośredni lub bezpośredni są włączone w realizację zadań i funkcjonowania technologii.
W szczególności chodzi o takie obszary, jak zarządzanie ryzykiem, ocenę wpływu systemu na otoczenie i produkty końcowe, a także wdrożone procedury i standardy ISO.

Kolejną dobrą praktyką jest udział programistów, etyków, prawników, właścicieli procesów biznesowych, a także kadry menedżerskiej średniego i wyższego szczebla w analizie ryzyka. Co więcej, jej wyniki powinny być monitorowane i kontrolowane na każdym poziomie zarządzania i stanowić element sprawozdawczości zarządczej organizacji.

Ważnym elementem budowania etycznego podejścia do wykorzystywania technologii AI jest także możliwość ludzkiego nadzoru na każdym etapie przetwarzania danych przez AI. Pozwala to na zapewnienie większej wiarygodności i przejrzystości technologii AI oraz unikaniu potencjalnych tzw. czarnych dziur w systemie.

Wreszcie przy tworzeniu etycznego wykorzystywania AI w organizacjach ważne jest, by uwzględnić i wdrożyć strategię szkoleń i stałego podnoszenia kompetencji pracowników. Dotyczy to zarówno ekspertów technicznych i programistów, jak i użytkowników danych systemów.

Opracowano na podstawie materiału z artykułu: B.C. Stahl, J. Antoniou, M. Ryan, K. Macnish, T. Jiya (2021), Organisational responses to the ethical issues of artificial intelligence, AI&SOCIETY.



Koncepcja godnej zaufania sztucznej inteligencji jako wzorzec normatywny dla rozwoju AI w Unii Europejskiej – założenia, cele, wymogi

Maciej Chojnowski

Opublikowane w kwietniu 2019 roku Wytyczne w zakresie etyki dotyczące godnej zaufania sztucznej inteligencji zawierają wizję normatywną, która określa pożądany kierunek rozwoju i zastosowania AI w Unii Europejskiej. Wizję tę syntetycznie oddaje tytułowe pojęcie godnej zaufania sztucznej inteligencji (ang. trustworthy AI). Chodzi o to, aby technologia AI była w Europie tworzona i użytkowana zgodnie z zasadami demokracji i praworządności oraz z poszanowaniem praw podstawowych, tak by obywatele mogli z niej korzystać bez obaw, że zostaną systemowo pokrzywdzeni.

Zawarty w pojęciu trustworthy AI dość idealistyczny postulat zaufania wobec sztucznej inteligencji wzbudził zastrzeżenia części ekspertów zajmujących się społecznym oddziaływaniem technologii, także wśród osób związanych z unijnymi instytucjami. Paul Nemitz z Komisji Europejskiej podczas konferencji „Co poza zbiorem zasad? Szersze spojrzenie na etykę AI. Konteksty, podejścia, ryzyka" zorganizowanej w czerwcu 2022 roku przez Centrum Etyki Technologii Instytutu Humanites podkreślał, że we współczesnej rzeczywistości niezbędne jest zachowanie krytycznego stosunku wobec wszelkich form władzy – także tych sprawowanych dzięki technologiom – i sugerował szczególną ostrożność wobec pomysłu zaufania AI.

Myśląc o koncepcji trustworthy AI, nie należy jednak skupiać się wyłącznie na samym pojęciu (i ewentualnych wątpliwościach co do trafności jego wyboru), ale przede wszystkim trzeba dostrzec szerszy kontekst, w którym jest ono zakorzenione, oraz poznać części składowe samej tej koncepcji i ich wzajemne powiązania.

Na godną zaufania sztuczną inteligencję składają się trzy główne komponenty. Musi być ona:

- zgodna z prawem,
- etyczna,
- solidna.

Omawiane Wytyczne obejmują dwa ostatnie obszary. Z kolei unijne prawo dotyczące AI zostało określone w osobnym rozporządzeniu – Akcie w sprawie sztucznej inteligencji przyjętym 13 marca 2024 roku (jak również w innych zharmonizowanych przepisach).

Autorzy Wytycznych podkreślają, że podejście do rozwoju sztucznej inteligencji oparte na koncepcji trustworthy AI ma na celu wspieranie odpowiedzialnych i zrównoważonych innowacji AI, a zarazem uczynienie z Europy lidera w budowaniu sztucznej inteligencji, która wspiera dobrostan indywidualny oraz dobro wspólne. Zwracają także uwagę, że omawiana koncepcja może być zrealizowana tylko wtedy, gdy urzeczywistniane są jej wszystkie trzy komponenty.

I tak zgodność z prawem jest niezbędna, by można było mówić o godnej zaufania AI. Podobnie jest z pozostałymi obszarami: solidnością i etyką. Aby były godne zaufania, systemy AI muszą być solidne, czyli bezpieczne zarówno pod względem użytych technologii, jak i dopasowania do społecznego otoczenia, w którym działają. Z kolei etyka pozwala wykroczyć w myśleniu o AI poza znaną z prawa opozycję dozwolone-niedozwolone, umożliwiając refleksję nad zagadnieniami, których przepisy (jeszcze lub w ogóle) nie obejmują, jak na przykład kwestia pożądanego kierunku rozwoju technologii czy – z drugiej strony – rozwiązywania konfliktów występujących niekiedy pomiędzy poszczególnymi wartościami. Każdy z tych komponentów musi być jednak zawsze ujmowany w kontekście pozostałych dwóch.

Kwestia współzależności prawa, etyki i solidności jest szczególnie ważna w sytuacji, gdy mamy już do czynienia z opublikowanym Aktem w sprawie sztucznej inteligencji. Istnieje bowiem poważne ryzyko, że dążenie do zapewnienia czysto formalnej zgodności z zawartymi w nim przepisami przesłoni szerszą wizję normatywną zawartą w koncepcji godnej zaufania sztucznej inteligencji. O podobnym ryzyku mówią wprost Wytyczne, gdy przestrzegają przed mechanicznym odhaczaniem pozycji z listy kontrolnej i postulują troskę o etyczne rozumowanie uwzględniające za każdym razem indywidualny kontekst. Preambuła samego Aktu również odnosi się wprost do Wytycznych, wymieniając wszystkie siedem zawartych w nich wymogów[16] i zachęcając do ich przestrzegania.

Dowodzi to, iż unijna wizja zawarta w komunikacie KE Budowanie zaufania do sztucznej inteligencji ukierunkowanej na człowieka z 8 kwietnia 2019 roku jest cały czas aktualna:

[16] W preambule Aktu wymogi te są określane mianem zasad etycznych. Może być to mylące, ponieważ w Wytycznych cztery główne zasady etyczne to coś innego – są one nadrzędne wobec siedmiu kluczowych wymogów, do których odnosi się Akt. Jednak jako że preambuła wymienia i zwięźle opisuje wszystkie te wymogi, owa nieścisłość terminologiczna nie powinna być źródłem merytorycznych nieporozumień.

„wymiar etyczny SI nie jest luksusem czy dodatkiem: musi stanowić integralną część rozwoju SI. Dążąc do stworzenia ukierunkowanej na człowieka SI opartej na zaufaniu, zabezpieczamy poszanowanie naszych podstawowych wartości społecznych i tworzymy charakterystyczną markę Europy i jej przemysłu jako lidera w zakresie najnowocześniejszej SI, której cały świat może zaufać”[17].

U źródła koncepcji godnej zaufania sztucznej inteligencji leżą prawa podstawowe określone w traktatach UE i Karcie UE (por. artykuł Fundamenty etyki człowieka wobec sztucznej inteligencji. Karta praw podstawowych jako źródło etycznego podejścia do AI w UE w niniejszym zbiorze). Ich wspólnym mianownikiem jest pojęcie godności ludzkiej wraz z postulatem jej poszanowania. Na przykładzie praw podstawowych widać wyraźnie wzajemne powiązanie prawa i etyki. Są one bowiem z jednej strony czymś, co można wyegzekwować na drodze prawnej, z drugiej jednak pozostają ściśle powiązane ze statusem moralnym człowieka, przez co wykraczają poza dziedzinę prawa, a stają się częścią etyki.

Kluczowe dla koncepcji trustworthy AI są następujące prawa podstawowe:

- poszanowanie godności ludzkiej,
- wolność jednostki,
- poszanowanie demokracji, sprawiedliwości i praworządności,
- równość, brak dyskryminacji i solidarność,
- prawa obywatelskie.

Z praw tych wywiedziono cztery zasady [18], które są podstawą godnej zaufania sztucznej inteligencji:

---

[17] Komunikat Komisji do Parlamentu Europejskiego, Rady, Europejskiego Komitetu Ekonomiczno-Społecznego i Komitetu Regionów: Budowanie zaufania do sztucznej inteligencji ukierunkowanej na człowieka, Bruksela, dnia 8.4.2019 r., COM(2019) 168 final, s. 10.

[18] Zasady te są w Wytycznych nazywane ‘zasadami etycznymi’, zarazem jednak określa się ja mianem ‘podstaw godnej zaufania sztucznej inteligencji’. Wypada więc uznać, że jako takie nie są one ograniczone wyłącznie do komponentu etycznego koncepcji *trustworthy AI*, ale mają charakter całościowy. Zarzut ewentualnej niespójności może tu wynikać z faktu, że na kolejnym etapie – nazwanym ‘osiągnięcie godnej zaufania sztucznej inteligencji’ – mowa jest o siedmiu kluczowych wymogach wyprowadzonych właśnie z owych czterech zasad etycznych, tymczasem koncepcja *trustworthy AI* obejmuje przecież także komponent prawny, bez którego jej urzeczywistnienie nie będzie możliwe. Wydaje się więc, że przedstawiony schemat relacji części do całości nie jest do końca konsekwentny czy przejrzysty. Oczywiście, można argumentować, że to pominięcie wynika po prostu stąd, iż Wytyczne nie obejmują kwestii prawnych. Nota bene, na rzecz ‘uniwersalnej’ interpretacji funkcji zasad etycznych i wymogów *trustworthy AI* wydaje się przemawiać przywołanie w Preambule Aktu o sztucznej inteligencji rzeczonych Wytycznych wraz z omówieniem zawartych w nich poszczególnych wymogów (por. przypis 16).

- **Zasada 1**: **Poszanowanie autonomii człowieka** – chodzi w niej o zagwarantowanie, by osoby wchodzące w interakcje z systemami AI zachowały zdolność samostanowienia oraz uczestnictwa w procesach demokratycznych. Należy zapewnić ludziom możliwość kontroli nad AI między innymi dzięki takiemu zaprojektowaniu sztucznej inteligencji, by człowiek nie był zdeterminowany przez jej decyzje. Ponadto systemy AI powinny wspierać ludzi, a nie powodować ich wykluczenie w sferze zawodowej.
- **Zasada 2**: **Zapobieganie szkodom** – systemy AI nie mogą wywierać negatywnego wpływu na człowieka ani na środowisko naturalne i żyjące w nim istoty. Szczególną uwagę należy zwrócić na osoby wymagające specjalnego traktowania, jak również na konteksty, w których zastosowanie AI mogłoby skutkować asymetrią w zakresie sprawowania kontroli społecznej i dostępu do informacji.
- **Zasada 3**: **Sprawiedliwość** – w wymiarze materialnym wymaga zapewnienia sprawiedliwego podziału korzyści i kosztów oraz przeciwdziałania stronniczości i dyskryminacji. W wymiarze proceduralnym dotyczy możliwości zakwestionowania decyzji systemów AI oraz zagwarantowania skuteczności ewentualnych roszczeń, gdyby decyzje te były krzywdzące. Wiąże się również z koniecznością przestrzegania zasady proporcjonalności, tak by działania podejmowane na potrzeby funkcjonowania danego systemu były adekwatne do jego celów.
- **Zasada 4**: **Możliwość wyjaśnienia** – sposób działania systemów AI powinien być jak najbardziej przejrzysty, ich możliwości i cele jawne, a decyzje przez nie podejmowane możliwe do wyjaśnienia osobom, na które będą miały wpływ. Postulat wyjaśnialności AI jest stopniowalny i zależy od kontekstu oraz oddziaływania danego systemu: inny będzie w przypadku AI wykorzystywanej w obszarach wrażliwych (jak służba zdrowia czy opieka społeczna), a inny w domenach bardziej neutralnych, jak na przykład rekomendacje sprzedażowe.

Autorzy Wytycznych podkreślają, że powyższe – dość abstrakcyjne – zasady mogą czasem się ze sobą kłócić. Jako przykład takiej sytuacji wskazują wykorzystanie AI do przewidywania przestępstw, gdy możliwy jest konflikt między zasadą zapobiegania szkodom a zasadą poszanowania autonomii człowieka. Samo działanie prewencyjne jest bowiem wtedy uzasadnione w świetle zasady zapobiegania szkodom, ale już związana z nim inwigilacja jest sprzeczna z zasadą poszanowania autonomii i prywatności[19]. W takich sytuacjach należy

[19] W kontekście przywołanego przykładu daje o sobie znać inna kwestia, a mianowicie: czy wskazania wynikające z wyżej wymienionych zasad mają charakter negatywny, czy pozytywny? Innymi słowy: czy mają one tylko przeciwdziałać ewentualnym nadużyciem, czy może jednak także niejako „projektować" korzystne zastosowania systemów AI? Wydaje się, że sposób, w jaki zostały one opisane w Wytycznych, sugerowałby pierwszą z wymienionych możliwości. Tymczasem przykład z predykcją przestępstw potraktowany jako urzeczywistnienie zasady przeciwdziałania szkodom

unikać arbitralności i szukać rozwiązań w sposób zgodny z regułami rozumowania etycznego. Co ważne, w niektórych wypadkach kompromis może w ogóle nie być możliwy (np. gdyby chodziło o ustępstwa kosztem godności ludzkiej).

Cztery naczelne zasady etyczne zostały w Wytycznych rozwinięte w siedmiu kluczowych wymogach, których spełnienie ma pozwolić na osiągnięcie godnej zaufania inteligencji[20]. Każdy z nich odpowiada co najmniej jednej zasadzie. Wymogi te są następujące[21]:

- **Wymóg 1: Przewodnia i nadzorcza rola człowieka** – odwołuje się do zasady poszanowania autonomii ludzkiej i obejmuje między innymi obszary:
  - praw podstawowych – należy przeprowadzić ocenę oddziaływania (impact assessment) systemów AI w tym zakresie oraz stworzyć mechanizmy przyjmowania zgłoszeń o ewentualnych naruszeniach tych praw przez dany system,
  - przewodniej roli człowieka – użytkownikom systemów AI trzeba zapewnić informacje i narzędzia umożliwiające świadomą interakcję z systemami AI oraz autonomię w ocenie ich działania (włącznie z możliwością zakwestionowania wyniku wygenerowanego przez dany system),
  - sprawowania nadzoru nad AI przez człowieka – nadzór ten może odbywać się z wykorzystaniem różnych mechanizmów, jak np. human-in-the-loop (umożliwia interwencję w każdym cyklu decyzyjnym systemu), human-on-the-loop (zakłada możliwość interwencji ludzkiej w trakcie cyklu projektowego systemu oraz monitorowanie jego działania) czy human-in-command (pozwala na nadzorowanie ogólnego funkcjonowania systemu AI i decyzję, kiedy i jak z niego korzystać w zależności od sytuacji).
- **Wymóg 2: Techniczna solidność i bezpieczeństwo**[22] – odwołuje się do zasady zapobiegania szkodom i obejmuje między innymi obszary:

sugeruje, że może chodzić także o drugą opcję. W przeciwnym razie takie jego przedstawienie nie byłoby możliwe. Ujawnia się tu zatem niejednoznaczność, która nie znajduje rozstrzygnięcia w treści Wytycznych.

[20] Por. przyp. 18.

[21] Poniższe zestawienie zawiera wyciąg najważniejszych elementów składających się na poszczególne wymogi. Ich bardziej szczegółowe omówienie znajduje się w Wytycznych, jak również w innych dokumentach dotyczących konkretyzacji wymogów trustworthy AI w praktyce projektowej, jak np. podejście Ethics by Design (por. np. zawarty w niniejszym opracowaniu artykuł Wdrażanie etyki AI z użyciem podejścia Ethics by Design – założenia i mechanizm działania).

[22] Również w tym przypadku rzuca się w oczy pewna wewnętrzna (relacyjna czy strukturalna) niekonsekwencja Wytycznych: otóż wymóg technicznej solidności i bezpieczeństwa stanowi jedyne miejsce, w którym rozwinięta jest kwestia trzeciego komponentu koncepcji trustworthy AI, czyli solidności. Tym samym staje się on elementem podrzędnym innego komponentu, jakim jest etyka.

- odporności na atak i zapewnienia bezpieczeństwa – systemy AI, podobnie jak inne systemy IT, powinny być możliwie jak najlepiej chronione przed lukami, które umożliwiają zhakowanie danego systemu,
- planowania awaryjnego – oddane do użytku systemy AI muszą umożliwiać wdrożenie planu awaryjnego w przypadku wystąpienia problemów,
- bezpieczeństwa ogólnego – należy zapewnić procedury umożliwiające zrozumiałą ocenę ewentualnych zagrożeń związanych z wykorzystaniem systemów AI w różnych obszarach, które pozwalałyby na podjęcie adekwatnych środków bezpieczeństwa,
- dokładności – systemy AI powinny zostać opracowane zgodnie z odpowiednio przygotowanym procesem rozwoju oraz ewaluacji, tak by w jak największym stopniu ograniczyć ryzyko niedokładności generowanych przez nie prognoz,
- wiarygodności (reliability) i odtwarzalności (reproducibility) – należy zapewnić, że systemy AI będą wiarygodne, to znaczy, że będą działały prawidłowo na podstawie różnych danych wejściowych oraz w różnych kontekstach, oraz że ich działanie będzie odtwarzalne, tzn. że w doświadczeniu powtórzonym w takich samych warunkach dany system AI zachowa się w taki sam sposób.

- **Wymóg 3: Ochrona prywatności i zarządzanie danymi** – odwołuje się do zasady zapobiegania szkodom i obejmuje między innymi obszary:
  - ochrony prywatności i danych – wymóg ten dotyczy całego cyklu życia systemu AI, przy czym chodzi nie tylko o dane początkowo doń wprowadzone, ale również o te wygenerowane w wyniku interakcji użytkownika z systemem,
  - jakości i integralności danych – jeszcze przed rozpoczęciem trenowania systemu AI należy zadbać o jakość używanych danych, zaś w zapewnieniu ich integralności powinny pomóc odpowiednie procedury testowania oraz dokumentowania wszelkich czynności podejmowanych na każdym etapie prac nad systemem AI,
  - dostępu do danych – w przypadku przetwarzania przez system AI danych osobowych niezbędne jest stosowanie protokołów określających, kto z personelu i w jakich okolicznościach może mieć do nich dostęp.
- **Wymóg 4: Przejrzystość** – odwołuje się do zasady możliwości wyjaśnienia i obejmuje między innymi obszary:
  - identyfikowalności (traceability) – należy zadbać, by dane używane przez system AI oraz (na tyle, na ile to możliwe) procesy wiodące do podjęcia określonej decyzji były dokumentowane zgodnie z obowiązującymi standardami, tak by maksymalnie ułatwić

Można to oczywiście tłumaczyć w taki sposób, że kwestie bezpieczeństwa dalece wykraczają poza zakres tematyczny Wytycznych, a zarazem stanowią fundamentalny komponent podejścia trustworthy AI. Nie zmienia to jednak faktu, że wskazane tu pomieszanie poziomów może utrudniać orientację w logice dokumentu i prowokować zarzuty pewnej redundancji.

przeprowadzania audytów systemu (auditability) i jego wytłumaczalność (explainability),

- wytłumaczalności – wymóg odnosi się zarówno do wewnętrznych procesów zachodzących w systemie AI, jak i związanych z jego rekomendacjami decyzji człowieka, przy czym dokładność oczekiwanych wyjaśnień zależy od kontekstu, w jakim działa system, zaś samo wyjaśnienie powinno być dostosowane do poziomu wiedzy interesariuszy (zwiększenie przejrzystości danego systemu AI może skutkować jego mniejszą dokładnością, co należy zawsze rozważyć w odniesieniu do zakresu działania danego systemu AI),
- komunikacji – użytkowników systemów AI należy informować, że mają do czynienia ze sztuczną inteligencją, a w sytuacjach, gdy jest to niezbędne dla zapewnienia zgodności z prawami podstawowymi – zapewnić możliwość interakcji z człowiekiem w miejsce systemu AI.

- **Wymóg 5: Różnorodność, niedyskryminacja i sprawiedliwość** – odwołuje się do zasady sprawiedliwości i obejmuje między innymi obszary:
  - unikania niesprawiedliwej stronniczości (bias) – możliwe do zidentyfikowania skrzywienia zawarte w danych trzeba eliminować jak najwcześniej (najlepiej już na etapie gromadzenia danych), a samo działanie istniejących systemów AI powinno być objęte procedurami nadzoru pozwalającymi analizować decyzje systemu w kontekście jego celów i ograniczeń,
  - dostępności i uniwersalności projektu – przy projektowaniu systemów AI (szczególnie tych skierowanych do nieinstytucjonalnych użytkowników końcowych) należy uwzględnić jak najszersze grono potencjalnych odbiorców oraz odpowiednie normy dostępności,
  - partycypacji interesariuszy – systemy AI powinny być tworzone w sposób uwzględniający konsultacje z osobami czy organizacjami, na które mógłby mieć on bezpośredni lub pośredni wpływ, a ich opinie powinny być przyjmowane i analizowane przez cały cykl życia danego systemu.
- **Wymóg 6: Dobrostan społeczny i środowiskowy** – odwołuje się do zasad sprawiedliwości oraz zapobiegania szkodom i obejmuje między innymi obszary:
  - zrównoważonej i przyjaznej dla środowiska AI – systemy AI powinny być poddawane ocenie pod względem swojego oddziaływania na społeczności i środowisko na każdym etapie ich rozwoju, wdrażania i użytkowania, z uwzględnieniem całego łańcucha dostaw, zaś podejmowane decyzje w tym zakresie powinny być możliwie jak najbardziej zrównoważone (na przykład w kwestii zużycia energii czy zasobów naturalnych),

- skutków społecznych – należy dokładnie analizować oddziaływanie społeczne systemów AI we wszystkich obszarach ich zastosowania, by móc rzetelnie ocenić ich wpływ na relacje społeczne,
- społeczeństwa i demokracji – ocena wpływu systemów AI na społeczeństwo powinna obejmować nie tylko jednostki, ale również instytucje, ze szczególnym uwzględnieniem ich oddziaływania na demokrację.

- **Wymóg 7: Odpowiedzialność** – odwołuje się do zasady sprawiedliwości, a zarazem jest uzupełnieniem wszystkich wcześniejszych wymogów, i obejmuje między innymi obszary:
  - możliwości kontroli (auditability) – systemy AI powinny być projektowane, wdrażane i użytkowane w sposób umożliwiający przeprowadzanie ich audytów – wewnętrznych lub zewnętrznych (w zależności od obszaru zastosowań danego systemu i jego wpływu na prawa podstawowe),
  - minimalizacji i zgłaszania negatywnych skutków – w przypadku stosowania systemów AI należy zapewnić możliwość zgłaszania zachowań lub decyzji uznanych za błędne lub krzywdzące, a ocena ich oddziaływania (impact assessment) powinna być przeprowadzana na etapach projektowania, wdrażania i użytkowania,
  - kompromisów – pomiędzy poszczególnymi wcześniej wymienionymi wymogami mogą występować konflikty, które powinny być rozwiązywane w sposób kompromisowy z uwzględnieniem oceny etycznej, o ile kompromisy te nie wiążą się z naruszeniem praw podstawowych (gdy etycznie dopuszczalny kompromis nie jest możliwy do osiągnięcia, należy wstrzymać dalsze rozwijanie bądź użytkowanie danego systemu w kształcie budzącym zastrzeżenia);
  - dochodzenia roszczeń – w sytuacji stosowania systemów AI należy zapewnić możliwość dochodzenia roszczeń na wypadek, gdyby działanie danego systemu okazało się niesprawiedliwe i krzywdzące.

Należy podkreślić, że realizacja powyższych wymogów nie jest zadaniem jednorazowym ani nie powinna sprowadzać się do powierzchownej i czysto formalnej oceny zgodności. Przeciwnie: urzeczywistnienie trustworthy AI to proces ciągły – wymagający regularnie ponawianych iteracji, także po wdrożeniu danego systemu. W związku z tym twórcy i operatorzy AI muszą liczyć się z tym, że systemy przez nich opracowane lub użytkowane będą z czasem wymagały zmian, tak by dostosować je do wymogów godnej zaufania sztucznej inteligencji.

Wymogi godnej zaufania AI w postaci zaprezentowanej powyżej (czy też w samych Wytycznych) są zbyt abstrakcyjne, by mogły zostać skuteczne zastosowane w praktyce

projektowej. Ich autorzy zauważają, że do konkretyzacji i operacjonalizacji wymogów służą różne metody techniczne i pozatechniczne.

Do tych pierwszych zaliczają oni między innymi projektowanie zgodne z podejściem Ethics by Design (por. artykuł Wdrażanie etyki sztucznej inteligencji przy projektowaniu systemów AI z użyciem podejścia Ethics by Design – założenia i mechanizm działania w niniejszym zbiorze), narzędzia z obszaru wyjaśnialnej AI (explainable AI) czy odpowiednie wskaźniki (benchmarks) pozwalające określić stopień zgodności danego systemu z wymogami bezpieczeństwa i etyki AI. Z kolei jako metody pozatechniczne wskazano między innymi normalizację zapewniającą jednorodność w zakresie różnych wymogów etycznych dotyczących systemów AI oraz certyfikację gwarantującą wiarygodność ich spełnienia przez podmioty używające AI.

Część ze wskazanych powyżej rozwiązań już istnieje, inne zaś są dopiero opracowywane. Można się spodziewać, że z czasem będą powstawały kolejne narzędzia pozwalające na coraz lepsze spełnianie wymogów z obszaru trustworthy AI. Również same wymogi, na co zwracają uwagę autorzy Wytycznych, nie stanowią listy zamkniętej i w przyszłości mogą być modyfikowane lub uzupełnianie o kolejne elementy. Wreszcie trzeba też zauważyć zbieżność znacznej części wymogów zawartych w Wytycznych z tymi, które wynikają z Aktu o sztucznej inteligencji. Są to bowiem dokumenty komplementarne, o czym mówi wprost zarówno sama koncepcja godnej zaufania sztucznej inteligencji, jak i przywoływana wcześniej preambuła europejskiego rozporządzenia.

Na koniec wypada przypomnieć, że w intencji Komisji Europejskiej koncepcja trustworthy AI ma na celu nie tylko eliminować potencjalne zagrożenia czy nadużycia związane w użytkowaniem systemów AI w Europie, ale także być siłą napędową innowacji przyjaznych człowiekowi i środowisku. Intencja ta znalazła rozwinięcie w drugim po Wytycznych dokumencie opracowanym przez grupę ekspertów wysokiego szczebla ds. sztucznej inteligencji, czyli Policy and investment recommendations for Trustworthy AI[23], który opublikowano w czerwcu 2019 roku.

Opracowano na podstawie: High-Level Expert Group on AI (AI HLEG), Ethics Guidelines for Trustworthy AI, April 2019 (oraz polskiej wersji tegoż dokumentu: Wytyczne w zakresie etyki dotyczące godnej zaufania sztucznej inteligencji).

[23] [dostęp: 17.04.2024].

Przegląd narzędzi służących realizacji wytycznych z zakresu etyki AI

Marzena Tyl

W erze dynamicznego rozwoju technologicznego i implementacji sztucznej inteligencji (AI) nieunikniona jest konfrontacja z etycznymi i prawnymi wyzwaniami, jakie niosą za sobą te nowatorskie rozwiązania. W tym artykule postaram się z perspektywy prawniczego eksperta przybliżyć treść oraz główne tezy zaprezentowane w pracy Putting AI ethics to work: are the tools fit for purpose? autorstwa Jacqui Ayling i Adriane Chapman.

Wprowadzenie AI do różnych obszarów życia niesie ze sobą wyzwania etyczno-prawne, którym obecne narzędzia etyczne nie zawsze mogą sprostać. Prawo ochrony danych osobowych, odpowiedzialność za szkody wyrządzone przez algorytmy czy kwestie związane z przejrzystością procesów decyzyjnych – to tylko część problemów, z którymi prawnicy i eksperci od AI muszą się mierzyć. Pomimo starań w zakresie etyki AI istnieje potrzeba lepszego dopasowania narzędzi służących realizacji wytycznych z zakresu etyki i prawa sztucznej inteligencji.

Omawiany artykuł powstał we wrześniu 2021 roku, co w kontekście zarówno rozwoju samej technologii sztucznej inteligencji, jak i wyzwań etycznych z nią związanych czyni go materiałem wymagającym uzupełnienia. Szczególnie po 30 listopada 2022 roku (kiedy to uruchomiono prototyp ChatGPT3) stało się jasne, że rozważania i dylematy etyczne dotyczące sztucznej inteligencji nie są wyłącznie domeną wąskiego grona naukowców i ekspertów AI, ale dotyczą (i zajmują) również zwykłych ludzi.

Badania Jacqui Ayling i Adriane Chapman

W analizowanej pracy autorki podejmują krytyczną ocenę dostępnych narzędzi etyki AI, koncentrując się na ich przydatności w praktyce. Wskazują na wyzwania związane z różnorodnością regulacji prawnych w różnych jurysdykcjach, co może prowadzić do niejednoznaczności w ocenie, czy narzędzia te są adekwatne. Ponadto Ayling i Chapman podkreślają konieczność tworzenia bardziej konkretnych i praktycznych wytycznych, które uwzględniałyby zarówno aspekty etyczne, jak i prawne. Autorki zidentyfikowały 169 ujęć

(frameworków) w szeroko rozumianej kategorii etyki sztucznej inteligencji, jednak na finalnej liście, po zastosowaniu kryteriów wykluczenia, pozostało 39, które zostały objęte analizą w omawianym artykule.

Kryteria i metodyka

Zgodnie z prezentowanym przez autorki stanowiskiem, aby prawidłowo zastosować proponowane narzędzia etyki AI, należy w pierwszej kolejności zrozumieć, co oferują, czym się różnią, a także dokonać identyfikacji wszelkich luk. To umożliwi w przyszłości udoskonalanie i rozwijanie narzędzi.

W ramach prowadzonej analizy autorki wyodrębniły dziewięć istniejących podejść do oceny oddziaływania technologii między innymi na środowisko, prawa człowieka, prywatność czy poziom ryzyka. Z perspektywy radcy prawnego i compliance officera szczególne znaczenie mają trzy kluczowe kryteria analizy narzędzi wspomagających implementację wytycznych dotyczących etyki sztucznej inteligencji.

Pierwszym z nich jest ocena wpływu na prawa człowieka, co wynika z konieczności uwzględnienia norm prawnych i etycznych wdrażanych technologii, mających potencjalny wpływ na społeczeństwo oraz jednostki.

Drugim istotnym kryterium jest ocena wpływu na prywatność oraz ochronę danych, co wydaje się niezbędne w kontekście rosnących obaw związanych z prywatnością w świecie cyfrowym.

Trzecim kluczowym aspektem jest ocena ryzyka, co jest istotne z perspektywy funkcji compliance officera, którego zadaniem jest identyfikacja, monitorowanie i zarządzanie ryzykiem związanym z zastosowaniem sztucznej inteligencji w organizacji.

Powyższe kryteria stanowią priorytetowe obszary analizy dla skutecznej implementacji zasad etyki w kontekście sztucznej inteligencji z uwagi na to, że łączą w sobie zarówno wymiar prawny, jak i etyczny.

Kryteria zostały wyczerpująco opisane w ramach przyjętej metodyki prowadzonej analizy narzędzi służących realizacji wytycznych etyki AI. Zainteresowanych szczegółami odsyłam bezpośrednio do artykułu Jacqui Ayling i Adriane Chapman, gdzie można znaleźć opisy, tabele oraz infografiki umożliwiające przeprowadzenie oceny odnośnie do każdego z wymienionych kryteriów.

Wyniki pracy

Z przeprowadzonej przez Ayling i Chapman analizy wynika, że narzędzia służące do realizacji wytycznych w zakresie etyki są opracowywane w trzech kluczowych obszarach:

- oceny wpływu (impact assessment),
- audytu,
- narzędzi technicznych/projektowych.

Podejście autorek koncentruje się na różnych etapach rozwoju systemu sztucznej inteligencji. Obejmuje problem biznesowy, projekt, dane treningowe, opracowanie modelu, budowę narzędzia, testowanie, wdrożenie, aż po etap monitorowania. Ocena wpływu ex ante jest stosowana we wczesnych fazach opracowywania problemu biznesowego i w procesach zamówień systemu AI. Ich celem jest dostarczenie narzędzia do podejmowania decyzji dotyczących przejścia do kolejnych etapów rozwoju systemu AI lub zakupu proponowanego systemu sztucznej inteligencji, a także identyfikacja potencjalnych konsekwencji jego zastosowania. Ocenę skutków ex post wykorzystuje się po wdrożeniu, umożliwiając analizę wpływu systemu, często w kontekście określonych interesariuszy lub kwestii takich jak na przykład przestrzeganie praw człowieka. Realizowane w ten sposób działania audytowe są wykonywane systematycznie według ściśle zdefiniowanych procesów i wymagają weryfikacji przez trzecią stronę. Przeprowadzone badania wykazały porównywalne wykorzystanie analizy wpływu na etapie budowy systemu AI jak i jego użytkowania.

W swoim badaniu Ayling i Chapman wyróżniły narzędzia techniczne do oceny określonych aspektów zestawów danych treningowych lub modeli. Według nich narzędzia te są bardzo istotnym elementem w rozwiązywaniu problemów etycznych w systemach sztucznej inteligencji i jako takie powinny być częścią szerszego systemu zarządzania. Efektem wprowadzonych mechanizmów jest skrupulatna dokumentacja, która stanowi kluczową część oceny skutków i ma na celu uchwyci wszystkie aspekty etyczne wytworzonego produktu.

Autorki zwracają uwagę, że wybór adekwatnego narzędzia jest skomplikowany i stanowi barierę zarówno dla podmiotów rozwijających rozwiązania sztucznej inteligencji, jak i dla kupujących (odbiorców). Ilość dostępnych narzędzi służących realizacji wytycznych etyki AI oraz różne podejścia do oceny i analizy powodują, że zainteresowani wejściem w technologię AI stają przed trudnym wyborem, tymczasem zagadnienia etyczne wymagają

konkretnych działań. Podjęcie właściwej decyzji w zakresie etyki w konkretnej sytuacji na podstawie danych otrzymanych z narzędzi audytowych wymaga czasu i zasobów.

Ayling i Chapman zauważają również, że nawet korzystając z dobrze zdefiniowanych procedur i procesów w zakresie oceny ryzyka w kontekście wymogów etyki AI, nie unikniemy konieczności rozstrzygania bardzo skomplikowanych problemów w realnym świecie. Pogodzenie roszczeń i oczekiwań różnych interesariuszy oraz różnych punktów widzenia na zagadnienia etyczne i próba balansowania pomiędzy stronami wymaga kompleksowego ludzkiego osądu. Ayling i Chapman przywołują stwierdzenia Luciano Floridiego, że „nie ma etyki bez wyborów, odpowiedzialności i ocen moralnych, z których wszystkie wymagają wielu istotnych i wiarygodnych informacji oraz dobrego zarządzania. (…) Narzędzia etyczne mogą zapewnić wiarygodną podstawę decyzyjną, jednak bez solidnego nadzoru może to skutkować procedurami generującymi działania na zasadzie listy kontrolnej i pozornych gestów, które stanowią ethics washing".

Autorki wyciągają również inny ważny wniosek z przeprowadzonych badań. Analizowane przez nie narzędzia służące realizacji wytycznych etyki AI nie bazują na konkretnych regulacjach prawnych ani przepisach. Oznacza to, że instrumenty te pełnią funkcję dobrowolnej autoregulacji. Tymczasem procedury oceny skutków i praktyki audytowe w innych obszarach są zgodne z regulacjami krajowymi i międzynarodowymi, co zapewnia obiektywną, zewnętrzną weryfikację oraz stanowi gwarancję jakości.

Naprzeciw temu wychodzi Unia Europejska, proponując regulację w zakresie AI opartą na ryzyku, która wprowadza nowe wymogi dla dokumentacji i przejrzystości. Parlament Europejski 13 marca 2024 roku przyjął Akt o sztucznej inteligencji, co oznacza, że UE jako pierwsza na świecie wprowadzi kompleksowe przepisy dotyczące sztucznej inteligencji. Działania w tym zakresie podejmują jednak również Stany Zjednoczone i Chiny.

Ayling i Chapman zwracają ponadto uwagę na „niejasności językowe oraz różnice w podejściu do tego, co jest rozumiane jako kluczowe cechy oceny skutków i kontroli" widoczne w badanych przez nie praktykach różnych organizacji. Wynika z tego silna potrzeba opracowania dobrze określonych wymogów compliance dla systemów AI do zastosowania przez producentów systemów sztucznej inteligencji, a w kolejnych krokach dążenie do skutecznej operacjonalizacji zasad etycznych, które będą motywacją do tworzenia praktycznych narzędzi związanych z etyką AI.

Autorki badania znalazły również luki w sposobach włączania szerokiego grona interesariuszy przy spełnianiu wymogów etyki AI. Narzędzia służące realizacji wytycznych etyki AI są przeznaczone dla osób uczestniczących w procesie produkcyjnym systemów

sztucznej inteligencji i kluczowych decydentów z nim związanych. Obejmowanie szerszego grona interesariuszy stawia przed firmami, które projektują systemy sztucznej inteligencji, wyzwanie wiążące się z koniecznością alokacji czasu i zasobów oraz posiadaniem określonych umiejętności, nie zawsze obecnych w zespołach programistycznych. Taki udział oznacza posiadanie władzy, w tym prawa do podejmowania decyzji dotyczących priorytetów odzwierciedlających przekonania i cele różnych interesariuszy.

Perspektywy na przyszłość

W świetle omawianej pracy oraz na podstawie własnej analizy dostrzegam, że obecnie dostępne narzędzia, które służą realizacji wymogów etyki AI mogą nie być dostatecznie dopasowane do wyzwań etyczno-prawnych. Konieczne jest ich dalsze doskonalenie, uwzględniające bardziej precyzyjne wytyczne dotyczące zgodności z regulacjami prawnymi.

Prawo ochrony danych, odpowiedzialność za decyzje podejmowane przez algorytmy, kwestie związane z przejrzystością procesów decyzyjnych, prawa podstawowe, podmiotowość ludzka i nadzór ze strony człowieka, różnorodność i niedyskryminacja, bezpieczeństwo i dobrostan społeczny powinny stanowić kluczowe obszary, na które koncentrują się narzędzia etyczne. W tym kontekście na uwagę audytową zasługuje także AI, która stwarza szczególne ryzyko manipulacji. W unijnym projekcie AI Act do tej kategorii zaliczone zostały: systemy przeznaczone do wchodzenia w interakcję z osobami fizycznymi, systemy rozpoznawania emocji i kategoryzacji biometrycznej, systemy generujące obrazy, treści dźwiękowe albo treści wideo, które łudząco przypominają istniejące osoby, obiekty, miejsca, podmioty lub zdarzenia albo, które tymi obrazami oraz treściami manipulują (deepfake).

Adekwatne dostosowanie narzędzi etycznych do regulacji prawnych nie oznacza jedynie narzucenia ograniczeń czy restrykcji. Wręcz przeciwnie, prawo i etyka mogą współgrać i uzupełniać się, tworząc ramy, w których AI może działać w sposób innowacyjny i zgodny z wartościami społecznymi. Konieczność współpracy pomiędzy ekspertami ds. etyki, technologii i prawa jest kluczowym elementem w budowaniu efektywnego ekosystemu, który promuje zrównoważony rozwój AI.

Podsumowanie

Współczesne wyzwania prawne związane z AI stawiają przed prawnikami nowe i wymagające zadania. Analiza pracy Ayling i Chapman wskazuje, że istnieje potrzeba lepszego dopasowania narzędzi etycznych do prawnego kontekstu. Prawo i etyka nie powinny działać osobno, ale stanowić integralną część ekosystemu AI.

Podczas analizy dokumentu Unii Europejskiej Ethics guidelines for trustworthy AI (por. artykuł Koncepcja godnej zaufania sztucznej inteligencji jako wzorzec normatywny dla rozwoju AI w Unii Europejskiej – założenia, cele, wymogi w niniejszym zbiorze) można zauważyć, że wymogi etyki i prawa są zbieżne i współpraca w zakresie etyki i prawa jest niejako wpisana w europejski paradygmat. Opracowywanie, wdrażanie i wykorzystanie sztucznej inteligencji, która zapewnia zgodność z normami etycznymi, w tym z prawami podstawowymi, jako szczególnymi uprawnieniami moralnymi, a także zasadami etycznymi i powiązanymi podstawowymi wartościami, stanowi bowiem jedną z podstawowych cech niezbędnych do osiągniecia godnej zaufania sztucznej inteligencji. Współpraca między naukowcami, prawnikami i inżynierami jest kluczowa i niezbędna do osiągnięcia harmonii między innowacją a zgodnością z regulacjami prawymi. Bezpieczna przyszłość z AI to tworzenie narzędzi etycznych, które będą zarówno efektywne z technicznego punktu widzenia, jak i kompatybilne z wartościami i wymogami etyczno-prawnymi.

Wdrażanie etyki sztucznej inteligencji przy projektowaniu systemów AI z użyciem podejścia Ethics by Design – założenia i mechanizm działania

Anna Kowalska

Czym jest Ethics by Design?

Ethics by design to podejście do tworzenia technologii, które umożliwia włączenie zasad etycznych do rozwiązań zawierających elementy AI już na etapie ich projektowania. Chodzi o to, żeby w przypadku sztucznej inteligencji i robotyki aspekty etyczne stały się integralnymi wymaganiami tych systemów, równorzędnymi z niezawodnością czy bezpieczeństwem, a nie tylko późniejszym dodatkiem. Dzięki temu problemy etyczne są podejmowane już na wstępie oraz na każdym późniejszym etapie rozwoju i wdrażania projektu.

Omawiane tutaj podejście Ethics by Design zostało wypracowane w unijnym projekcie badawczym SIENNA (z ang. Stakeholder-Informed Ethics for New technologies with high socio-ecoNomic and human rights impAct), który zajmował się opracowaniem ram etycznych oraz zaleceń dotyczących lepszej regulacji oraz operacjonalizacji rozwiązań związanych z genomiką człowieka, robotyką oraz sztuczną inteligencją.

W poniższym opracowaniu zaprezentuję pięciopoziomowy model Ethics by Design, wartości leżące u jego podstaw oraz sposoby, w jakie AI Ethics by Design można włączyć już na etapie rozwoju systemu, ale także później w czasie jego użycia i utrzymania.

Pięciopoziomowy model Ethics by Design

Podejście Ethics by Design może zostać opisane za pomocą pięciopoziomowego modelu. Jest on podobny do wielu innych modeli informatycznych, w których warstwa abstrakcji maleje wraz z kolejnymi poziomami, a warstwa szczegółowości – rośnie.

Poziomy te są określone następująco:

1. **Wartości** – podstawowe wartości etyczne, którymi kierujemy się przy projektowaniu systemów AI, takie jak: sprawiedliwość czy prywatność.
2. **Warunki (requisites)** – wymogi, które rozwiązanie lub aplikacja muszą spełnić, aby osiągnąć swoje cele w sposób zgodny z określonymi wartościami. Warunki etyczne są konkretyzacją wartości. Może to być osiągnięte poprzez funkcjonalności systemu, odpowiednie budowanie struktur danych, procesy tworzenia systemu itp. Na przykład wartość „sprawiedliwość" może być skonkretyzowana poprzez wymaganie, aby system nie wykazywał uprzedzeń rasowych.
3. **Wytyczne** – podczas gdy warunki określają wymagania dotyczące samego systemu, wytyczne dotyczą kroków, w których jest on tworzony. Ethics by Design opiera się na założeniu, że w procesie rozwoju AI istnieją elementy wspólne dla wszystkich metodyk projektowania takich systemów. Dla każdego takiego elementu Ethics by Design określa związane z nim wymagania etyczne. Wynikiem tego są konkretne wytyczne (zwykle formułowane jako zadania) dla każdej fazy projektu, które zapewniają, że końcowy system realizuje określone wymagania etyczne i tym samym nie narusza żadnych wartości. Dla przykładu, wytyczne określają, że na etapie gromadzenia danych powinny być one testowane pod kątem sprawiedliwości, a ewentualne zawarte w nich uprzedzenia (bias) mogące skutkować dyskryminacją należy korygować jeszcze przed ich wykorzystaniem.
4. **Metodyki** – istnieje wiele metodyk stosowanych w projektach z zakresu sztucznej inteligencji. Różnią się one, przynajmniej częściowo, sposobem organizacji procesu projektowania i zawartymi w nim etapami. Podejście Ethics by Design zostało tak pomyślane, aby umożliwić dopasowanie wytycznych do komponentów każdej indywidualnej metodyki. W tym dokumencie przedstawiamy przykład zastosowania Ethics by Design w ogólnym modelu procesu projektowania. Dla każdej konkretnej metodyki programiści powinni dostosować poszczególne wytyczne do odpowiednich kroków.
5. **Narzędzia i metody** – ta warstwa obejmuje szczegółowe rozwiązania wykorzystywane podczas konkretyzacji określonych wymogów metody Ethics by Design w praktyce projektowej. Niektóre z nich mogą być właściwe dla jednej metodyki, a nieodpowiednie dla innych. Inne są uniwersalne i mogą być stosowne na różnych etapach procesu rozwoju, bez względu na przyjętą metodykę.

Poniżej zostaną opisane dwa pierwsze poziomy modelu Ethics by Design, czyli wartości i warunki. Dopiero po ich zdefiniowaniu możemy przejść dalej do metodyk i przypisywania konkretnych zadań z obszaru etyki AI do faz procesu projektowania.

Wartości wyróżnione w podejściu Ethics by Design są w dużym stopniu zbieżne z wymogami godnej zaufania sztucznej inteligencji w ujęciu KE (por. artykuł Koncepcja godnej zaufania sztucznej inteligencji jako wzorzec normatywny dla rozwoju AI w Unii Europejskiej – założenia, cele, wymogi w niniejszym zbiorze). Dla każdej wartości określono warunki pozwalające ją skonkretyzować:

- **Przewodnia rola człowieka** – ze względu na niezbywalną wartość każdej osoby systemy AI nie powinny negatywnie wpływać na autonomię, wolność czy godność człowieka, ani też ograniczać jej udziału w procesach demokratycznych.
  - **Warunki**: Ludzie powinni być informowani, gdy mają do czynienia z systemami AI. Powinni także otrzymać informacje o możliwościach i ograniczeniach tych systemów oraz o tym, jak oceniać generowane przez nie informacje. Ludzie nie mogą być traktowani przedmiotowo, a systemy AI nie mogą ich sobie podporządkowywać, ani do niczego przymuszać czy też oszukiwać bądź nimi manipulować.
- **Prywatność i zarządzanie danymi** – systemy sztucznej inteligencji nie mogą naruszać niczyjego prawa do prywatności, a dane wykorzystywane do budowania systemów AI muszą być reprezentatywne i poprawne.
  - **Warunki**: systemy AI powinny umożliwiać wycofanie zgody na wykorzystanie danych osobowych; należy zapewnić skuteczne środki bezpieczeństwa zapobiegające nieautoryzowanemu dostępowi do danych, naruszeniom danych i wyciekom informacji; przetwarzanie danych osobowych przez system AI powinno się odbywać zgodnie z prawem; dane powinny być pozyskiwane, przechowywane i używane w taki sposób, żeby mogły być poddane audytowi przez ludzi.
- **Sprawiedliwość** – rozwój systemów AI powinien odbywać się w duchu sprawiedliwości i zapewniać przeciwdziałanie dyskryminacji.
  - **Warunki**: systemy sztucznej inteligencji muszą unikać dyskryminacji na poziomie danych wejściowych, a także wykorzystywanych algorytmów; system AI powinny być dostępne dla użytkowników z różnym poziomem sprawności; zespoły tworzące AI powinien oceniać potencjalny wpływ społeczny danego systemu.
- **Dobrostan** – systemy AI powinny wspierać wolność i autonomię ludzi i pomagać im (a przynajmniej nie przeszkadzać) w zaspokajaniu własnych potrzeb.
  - **Warunki**: systemy AI powinny brać pod uwagę dobrostan użytkowników końcowych i nie obniżać jego poziomu; tworzenie AI powinno uwzględniać zasady zrównoważonego rozwoju w wymiarze środowiskowym, zarówno w odniesieniu do samego systemu, jak i związanego z nim łańcucha dostaw; systemy AI nie mogą obniżać bezpieczeństwa w miejscu pracy; systemy robotyczne, które będą

współdzielić środowisko pracy z ludźmi lub zwierzętami, powinny być wyposażone w odpowiednie funkcje bezpieczeństwa.

- **Odpowiedzialność i nadzór** – aby umożliwić rozumienie działania systemu AI oraz zapewnić kontrolę nad jego tworzeniem i użyciem, niezbędny jest nadzór człowieka oraz świadomość odpowiedzialności, tak aby zapewnić przestrzeganie opisanych powyżej zasad i reagowanie na przypadki ich nieprzestrzegania.
  - **Warunki**: należy umożliwić ludziom nadzór nad decyzjami i działaniem (ang. human in the loop) systemów AI; konieczna jest możliwość wykrywania niepożądanego oddziaływania systemu na użytkowników końcowych i jego modyfikacji; konieczne jest przeprowadzanie oceny ryzyka etycznego systemów AI oraz zapewnienie możliwości zgłaszania potencjalnych obawy etycznych dotyczących systemu AI oraz skarg przez osoby uważające się za poszkodowane przez system; należy również umożliwić rzetelne audyty systemów AI przez podmioty trzecie.
- **Transparentność** – systemy AI powinny być jak najbardziej przejrzyste, ponieważ tylko wtedy możliwe jest zapewnienie efektywnego nadzoru ludzkiego i utrzymanie autonomii człowieka.
  - **Warunki:** dostępność środków umożliwiających monitorowanie systemu AI przez cały jego cykl życia; jasne komunikowanie użytkownikom celu, zdolności, ograniczeń, korzyści i ryzyk danego systemu AI oraz podejmowanych przez niego decyzji; dostępność instrukcji użytkowania systemu; zapewnienie, że procesy projektowania i rozwoju AI obejmują kwestie etyczne (np. problem dyskryminacji w danych) i ich dokumentację; decyzje podejmowane przez system sztucznej inteligencji powinny być wyjaśnialne dla użytkowników.

Jak wykorzystywać Ethics by Design w procesie budowania systemów AI?

Ethics by Design bazuje na ogólnym modelu procesu projektowania systemów AI. Dzięki takiemu podejściu kwestie etyczne, które należy uwzględnić, traktowane są jak ogólne wymagania systemowe, podobnie jak niezawodność czy bezpieczeństwo. Wytyczne określone w poprzedniej części są więc dołączane do odpowiedniego elementu procesu projektu jako konkretne zadania do wykonania. Poprzez odniesienie używanej przez siebie metodyki rozwoju oprogramowania do ogólnego prezentowanego tutaj modelu, deweloper/ka może określić odpowiednie wymagania etyczne dla każdego elementu swojej metodyki.

W poniższej części zostaną przedstawione etapy składające się na ogólny model procesu projektowania, z którym zmapowano opisane wcześniej wartości i warunki (wymagania) w postaci konkretnych wskazań. Zadań dla każdej fazy projektowej może być więcej lub mniej, zależnie od natury projektu. Prezentowane zestawienie jest tylko propozycją dla budowania ogólnego systemu AI, bez specyfikacji konkretnego rozwiązania.

Ogólny model procesu projektowania zawiera 6 etapów. Mimo że zadania są tu przedstawione w kolejności, nie muszą być sekwencyjne, ale mogą (np. jak w metodyce Agile) mieć formę cykliczna. Owe sześć ogólnych zadań oraz związane z nimi wskazania to:

1. **Określanie celów** – określenie do czego ma służyć system i jakie powinien mieć funkcjonalności.
   Wskazania:
   - Zaleca się, aby – jeśli to możliwe – w ramach oceny celów do współpracy z członkami zespołu deweloperskiego zaangażowany został profesjonalny etyk AI.
   - **Przewodnia rola człowieka**: sprawdź, czy zaproponowane cele projektu są zgodne z wymaganiami dotyczącymi ludzkiej autonomii. Naruszenia zasad etycznych dotyczą systemów, które m.in. ograniczają prawa człowieka, oszukują ludzi lub nimi manipulują, naruszają ich integralność cielesną lub psychiczną, powodują uzależnienie lub ukrywają przed użytkownikami fakt, że mają do czynienia z systemem sztucznej inteligencji.
   - **Prywatność i zarządzanie danymi**: sprawdź, czy cele są zgodne z wymaganiami dotyczącymi prywatności i zarządzania danymi i czy początkowe plany dotyczące wykorzystania danych osobowych i nienazwanych są sprawiedliwe i właściwe.
   - **Sprawiedliwość**: rozważ, czy naruszenie któregokolwiek z wymagań sprawiedliwości spowodowałoby znaczący uszczerbek w wymiarze społecznym lub politycznym, zmniejszenie sprawczości, jaką ludzie posiadają nad aspektami swojego życia, takimi jak praca czy styl życia, lub skutkowałoby prawdopodobną dyskryminacją lub stygmatyzacją.
   - **Dobrostan**: określ, czy cele systemu mogą spowodować, że ludzie doznają szkód fizycznych, psychicznych lub finansowych, lub też wspierają procesy, które mogą spowodować istotne szkody środowiskowe lub dotyczące procesów społecznych i instytucji (na przykład przez przyczynianie się do dezinformacji publicznej).
2. **Określanie wymagań** – rozwinięcie funkcjonalnych i niefunkcjonalnych wymagań. Zawiera ocenę ryzyka projektu, analizę kosztów i korzyści oraz plan projektu.
   Wskazania:

- Należy przeprowadzić ocenę etyczną proponowanej specyfikacji projektu, wybranych zasobów i infrastruktury oraz związanych z tym ewentualnych ograniczeń. Na przykład wybór technik uczenia głębokiego jako baza systemu AI, który wymaga wysokiej transparentności i możliwości wyjaśnienia, może zostać oceniony jako błędny.
- Upewnij się, że istotne wymagania etyczne właściwe dla danego projektu są uwzględnione na liście jego specyfikacji.

3. **Projektowanie wstępne (high-level design)** – opracowanie wysokopoziomowej architektury systemowej.

   Wskazania:

   - **Przewodnia rola człowieka**: zweryfikuj, czy wybrana architektura umożliwia stworzenie interfejsu, który pozostawia ludziom możliwość wyboru oraz zapewnia swobodę wyrażania się i dostępu do informacji.
   - **Prywatność i zarządzanie danymi:** upewnij się, że architektura rozwoju zawiera procesy, procedury i narzędzia, które zapewnią, że dane osobowe nie zostaną ujawnione w sposób naruszający prawo do prywatności. Upewnij się, że istnieją procesy zapewniające, że wybór danych dla systemu będzie sprawiedliwy, dokładny i pozbawiony stronniczości. Zaplanuj wstępną ocenę źródeł danych, zanim zostaną one wprowadzone do systemu. Zaprojektuj mechanizm dokumentowania i uzasadniania początkowego wyboru danych na potrzeby audytów. Opracuj formalne procesy weryfikacji i korekty stronniczości (lub błędów) w zaimportowanych danych.
   - **Sprawiedliwość**: sprawdź, czy projekt może zapewniać niektórym użytkownikom systemu lepsze funkcjonalności niż innym (jeśli tak, przygotuj formalne uzasadnienie dla takiego zróżnicowanego dostępu lub popraw projekt). Przeprowadź ocenę dostępności interfejsu oraz innych punktów dostępowych systemu. Upewnij się, że system spełnia standardy dostępności.
   - **Dobrostan**: przygotuj dokumentację, która wyjaśni, w jaki sposób skonstruowany system zapewni przyjazność dla środowiska. Oceń, czy system może potencjalnie spowodować szkody fizyczne dla ludzi lub mienia.
   - **Odpowiedzialność i nadzór**: zaprojektuj model etycznego zarządzania podczas rozwoju projektu. Skup się na mechanizmach, które umożliwią ludziom nadzór w trakcie prac nad rozwojem systemu AI. Model etycznego zarządzania musi uwzględniać następujące kwestie: Jak będzie wyglądać proces zarządzania? Jaka instytucja zapewni spełnienie wymagań etycznych (np. niezależny konsultant, rada programowa etyki w firmie)? Jakie będą jej uprawnienia? Jak zostanie wybrana? Jak można wykazać, że proces wyboru jest uczciwy i obejmuje różnorodność? Jakie

procedury zostaną zastosowane w przypadku konfliktu między instytucją etycznego nadzoru a deweloperami, inżynierami lub klientami? Dodatkowo zaprojektuj mechanizmy nadzoru ludzkiego oraz audytu zewnętrznego działające po wdrożeniu systemu.

- **Transparentność**: Zaprojektuj mechanizmy dokumentowania sposobu pozyskiwania, przechowywania i wykorzystywania danych umożliwiające przeprowadzenie audytu. Stwórz narzędzia do dokumentowania procesów rozwoju systemu, aby ludzie mogli je zrozumieć i ocenić podejmowane przez zespół decyzje w ramach procesów projektowania i rozwoju systemu. Upewnij się, że projekt zawiera mechanizmy, dzięki którym system AI będzie rejestrował swoje własne decyzje, aby mogły być one poddane ocenie przez ludzi.

4. **Zebranie i przygotowanie danych** – dotyczy zbierania, weryfikacji i czyszczenia danych.

   Wskazania:

   - **Prywatność i zarządzanie danymi:** gdy system AI przetwarza dane osobowe, przestrzegaj zasady minimalizacji danych: tylko dane istotne, adekwatne i absolutnie niezbędne powinny być przetwarzane przez system. Wszystkie dane osobowe muszą być przetwarzane w sposób zgodny z prawem, przejrzysty i sprawiedliwy.
   - **Sprawiedliwość**: przeanalizuj swoje dane treningowe i upewnij się, że są one reprezentatywne i nie są stronnicze w swojej reprezentacji różnych grup społecznych, na przykład poprzez nadreprezentację niektórych kategorii, brak różnorodności w reprezentacji lub ukryte stereotypowanie. Tam, gdzie stwierdza się możliwość szkodliwej stronniczości, zbuduj mechanizmy, które pozwolą jej uniknąć lub ją skorygują.
   - **Odpowiedzialność i nadzór**: wytwórz kulturę wspólnej odpowiedzialności za dane dostępne w organizacji. Upewnij się, że role i odpowiedzialności są jasno określone w zakresie zarządzania danymi w organizacji oraz że wszyscy pracownicy i interesariusze je rozumieją. Upewnij się, że zostało wyraźnie określone, jakiego rodzaju dane są potrzebne, jaką próbkę pobrano oraz że potrafisz wyjaśnić, do czego będzie ona wykorzystana.
   - **Transparentność**: przygotuj dokument, który szczegółowo opisuje, w jaki sposób projekt przestrzega wymagań ochrony danych. Musisz przeprowadzić analizę ryzyka etycznego związanego z przetwarzaniem danych i przygotować plan ograniczania ryzyka. Upewnij się, że potrafisz wytłumaczyć, w jaki sposób dane osobowe są używane, udostępniane i przechowywane.

5. **Projektowanie szczegółowe i budowa systemu** – rzeczywiste budowanie kompletnego działającego systemu.

Wskazania:

- **Prywatność i zarządzanie danymi:** jeśli generujesz nowe dane osobowe lub wrażliwe (np. poprzez szacowanie brakujących danych, generowanie pochodnych cech i nowych rekordów, integrację danych lub łączenie zestawów danych), może być konieczne uzyskanie dalszej świadomej zgody użytkownika. Takie dane powinny być także objęte co najmniej taką samą ochroną, jak wcześniej zebrane dane osobowe. Upewnij się, że istnieją procesy zapewniające ochronę jakości i integralności wszystkich istotnych danych, w tym środki weryfikacji, że zestawy danych nie zostały naruszone lub zhakowane oraz że system umożliwia osobom dostęp do ich danych, usunięcie ich z systemu i/lub poprawienie ewentualnych błędów. Dane mogą być manipulowane, uszkodzone, utracone lub niewłaściwie ujawnione w dowolnym systemie. Zaprojektuj procesy weryfikujące jakość zabezpieczenia danych pod względem etycznym.
- **Sprawiedliwość**: na tym etapie zapewnij stałą kontrolę nad ewentualnym wystąpieniem stronniczości algorytmicznej. Podczas prac nad interfejsem zadbaj o testy z różnymi interesariuszami.
- **Dobrostan**: przestrzegaj zasad efektywnego wykorzystywania zasobów i zrównoważonego zużycia energii.
- **Odpowiedzialność i nadzór**: staraj się wypracować w środowisku programistów kulturę pracy, w której rozwiązywanie kwestii etycznych jest postrzegane jako istotne, a nie kłopotliwe zadanie, które należy rozważyć dopiero po zakończeniu innych prac. Upewnij się, że w system wbudowane są mechanizmy kontrolne, które raportują wydajność oraz rejestrują podejmowane przez system decyzje. Stwórz mechanizmy, dzięki którym można ocenić zgłoszone obawy pracowników i osób trzecich i w razie konieczności podjąć odpowiednie działania.
- **Transparentność**: upewnij się, że kod jest dokumentowany (zgodnie z odpowiednimi językami i metodologią). Upewnij się, że dokumentacja ta jest dostępna i zrozumiała dla współpracowników. Upewnij się, że wiesz, w jakim stopniu decyzje podejmowane przez system mogą być zrozumiane i wytłumaczone, w tym także czy masz dostęp do wewnętrznego mechanizmu pracy modelu. Korzystaj z formalnych metodyk i narzędzi, aby zapewnić możliwość wyjaśniania (np. XAI). Jeśli system może prezentować fałszywe lub wprowadzające w błąd informacje, dodaj wymagania projektowe, które zminimalizują to ryzyko.

6. **Testowanie i ewaluacja** – testowanie systemu i jego ocena w perspektywie pierwotnych celów i wymagań.

Wskazania:

- **Odpowiedzialność i nadzór**: upewnij się, że istnieją procesy, które umożliwiają stronom trzecim (takim jak dostawcy, konsumenci, dystrybutorzy) lub pracownikom zgłaszanie potencjalnych błędów lub stronniczości w systemie. Proces testowania powinien obejmować zrozumienie funkcjonalności systemu przez użytkowników końcowych. Zbadaj rozumienie przez użytkowników faktycznego celu systemu, kto może z niego korzystać oraz jakie są jego ograniczenia. Upewnij się, że użytkownicy otrzymują wyjaśnienia, które są dla nich zrozumiałe. Przeprowadź próbę przewidzenia skutków ubocznych działania systemu.
- **Transparentność**: upewnij się, że w systemie wbudowane są mechanizmy kontrolne, które sprawdzają wydajność, rejestrują decyzje podjęte przez system i jego funkcjonowanie. Przetestuj, czy użytkownicy rozumieją, że interakcja odbywa się z systemem AI i że jego decyzja lub rekomendacja jest wynikiem decyzji algorytmicznej. Upewnij się, że informacje o zdolnościach i ograniczeniach systemu są przekazywane użytkownikom w sposób jasny zrozumiały i pozwalający formułować realistyczne oczekiwania wobec AI.

Ethics by design oczami praktyka

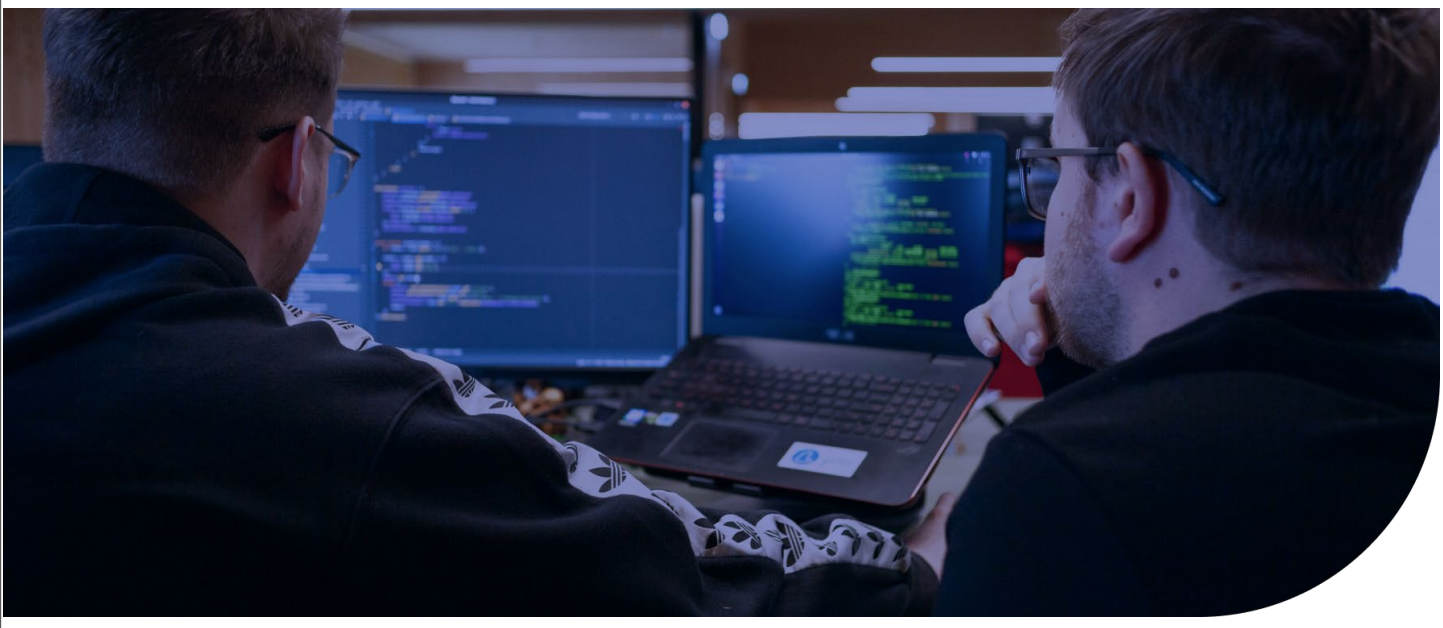
Podejście Ethics by Design, umożliwiając włączanie zagadnień etycznych do projektu AI już na poziomie jego planowania i budowania, ma fundamentalne znaczenie dla realizacji wymogów etyki AI. Traktowanie etyki jako dodatku jest w przypadku AI o tyle nieskuteczne, że wiele zagadnień etycznych rozbija się o kwestie danych – ich źródeł, przygotowania oraz wybranych metod modelowania. Oba te elementy są obecne na wczesnych etapach budowania rozwiązań AI.

Kluczowe znaczenie w przypadku podejścia Ethics by Design ma sposób operacjonalizacji wartości w postaci konkretnych wskazań dopasowanych do poszczególnych etapów procesu projektowego. Na poziomie abstrakcyjnie ujętych wartości łatwo jest się bowiem zgodzić co do zasadności wymogów etyki AI, jednak brak narzędzi pozwalających na ich operacjonalizację skutkował w wielu projektach niemożliwością efektywnego wdrożenia wymogów etyki na szeroką skalę.

Następnym krokiem ku dalszej konkretyzacji podejścia Ethics by Design byłoby rozpisanie wymagań dla określonych rodzajów projektów AI, takich jak budowanie modeli wizji, tekstowych czy tabularycznych, ze wskazaniem, jak można spełnić dany wymóg oraz podaniem przykładowych narzędzi. Opisane powyżej podejście oferuje nam przekrojową

analizę etyczną, jednak nadal dostosowanie jej do konkretnego projektu wymaga szerokiej współpracy etyków z wszystkimi podzespołami projektowymi, takimi jak inżynierowie danych czy architekci high-level, by zdefiniować, jak dokładnie spełnić określoną wytyczną. Aby zespoły projektowe niemające w swoich szeregach etyków nie były pozostawione same sobie, warto w przyszłości opracować choćby ogólne wytyczne narzędziowe i procesowe dla tych trzech typów projektów (wizja, tekst, dane tabelaryczne). To mogłoby doprowadzić zespoły programistyczne do wprowadzenia Ethics by Design do swoich procesów w szybszy i bardziej efektywny sposób.

Opracowano na podstawie materiałów: Ethics By Design and Ethics of Use Approaches for Artificial Intelligence oraz Annex 2 - Ethics By Design and Ethics of Use in AI and Robotics [w:] SIENNA D5.4: Multi-Stakeholder Strategy and Tools for Ethical AI and Robotics.



Operacjonalizacja wymogów trustworthy AI w organizacjach używających sztucznej inteligencji – analiza struktury wytycznych z projektu SHERPA

Maciej Chojnowski

Wymogi zawarte w unijnych Wytycznych w zakresie etyki dotyczących godnej zaufania sztucznej inteligencji są dość abstrakcyjne, na co zwrócili uwagę już sami ich twórcy, zamieszczając w rzeczonym dokumencie listę technicznych i pozatechnicznych metod ich urzeczywistniania (por. artykuł Koncepcja godnej zaufania sztucznej inteligencji jako wzorzec normatywny dla rozwoju AI w Unii Europejskiej – założenia, cele, wymogi w niniejszym raporcie). Jednym z użytecznych narzędzi umożliwiających operacjonalizację wymogów trustworthy AI w praktyce projektowej ukierunkowanej na rozwijanie AI jest podejście Ethics by Design opracowane w projekcie badawczym SIENNA (por. artykuł Wdrażanie etyki AI z użyciem podejścia Ethics by Design – założenia i mechanizm działania w niniejszym raporcie).

Jednak koncepcja godnej zaufania sztucznej inteligencji nie odnosi się wyłącznie do organizacji czy zespołów tworzących rozwiązania z obszaru AI, ale także do firm czy instytucji, które **używają** (bądź planują używanie) systemów sztucznej inteligencji. W tym drugim wypadku konkretyzacja wymogów trustworthy AI musi być więc pod różnymi względami odmienna od podejścia wypracowanego w ramach Ethics by Design.

Propozycję takiego szczegółowego rozwiązania stworzono w unijnym projekcie badawczym SHERPA (dokument pt. Guidelines for the Ethical Use of AI and Big Data Systems) i ona właśnie jest przedmiotem niniejszego artykułu. (Dla odróżnienia od Ethics by Design podejście to można określić mianem Ethically aligned Implementation, choć nie jest to wyrażenie używane w omawianym dokumencie). Należy jednak podkreślić, że poniższa analiza skupia się na **strukturze** rzeczonego rozwiązania, a dokładniej: na **logice** powiązań jego części składowych. Jest to bowiem najbardziej skomplikowany aspekt tej propozycji, który warto przedstawić krok po kroku. W związku z tym poszczególne (tzw. „finalne") wymogi stanowiące ostateczne rekomendacje zawarte w tym podejściu nie są tu analizowane. Wszyscy zainteresowani znajdą je w źródłowym dokumencie, w razie potrzeby korzystając z zawartych w niniejszym artykule podpowiedzi.

W opracowaniu SHERPA wyodrębniono kilka ważnych komponentów. Po pierwsze jest w nim mowa o **dwóch rodzajach wytycznych**: **ogólnych** (general guidelines) i **operacyjnych** (operational guidelines), których jednak nie należy traktować osobno, lecz w sposób komplementarny. Drugim istotnym elementem są **wymogi nadrzędne** (high-level requirements) oraz **wymogi szczegółowe** (sub-requirements). Trzecim – **model obrazujący wdrażanie** i **użytkowanie systemów IT w organizacjach** obejmujący dwa obszary: **strategii IT w organizacji** (IT governance) oraz **zarządzania usługami IT w organizacji** (IT management).

Zacznijmy od wymogów nadrzędnych oraz szczegółowych, stanowiących doprecyzowanie tych pierwszych. (Opracowanie SHERPA jest zgodne z unijnymi Wytycznymi dot. godnej zaufania sztucznej inteligencji, stąd też zawarte w nim siedem wymogów nadrzędnych generalnie pokrywa się – mimo niewielkich różnic – z pierwowzorem. Także wymogi szczegółowe są w dużym stopni zbieżne ze sposobem, w jaki poszczególne wymogi trustworthy AI zostały dookreślone w tamtym unijnym dokumencie [por. s. 26-28 niniejszego opracowania]). Są one następujące:

- **Wymóg nadrzędny 1**: Sprawczość, wolność i godność ludzka
  wymogi szczegółowe:
  - zapewnienie wolności pozytywnej
  - zapewnienie wolności negatywnej
  - ochrona godności człowieka
- **Wymóg nadrzędny 2**: Techniczna solidność i bezpieczeństwo
  wymogi szczegółowe:
  - cyberbezpieczeństwo i odporność na ataki
  - rozwiązania awaryjne (fall-back plan) i ogólne bezpieczeństwo
  - wiarygodność (reliability) i odtwarzalność (reproducibility)
- **Wymóg nadrzędny 3**: Prywatność i zarządzanie danymi
  wymogi szczegółowe:
  - poszanowanie prywatności
  - jakość i integralność danych
  - zapewnienie dostępu do danych i praw do nich – w tym prawa własności (ownership)
- **Wymóg nadrzędny 4**: Przejrzystość
  wymogi szczegółowe:
  - identyfikowalność (traceability)
  - wyjaśnialność (explainability)

    - zapewnienie odpowiednich informacji i komunikacji
- **Wymóg nadrzędny 5**: Różnorodność, brak dyskryminacji i bezstronność (fairness)

  wymogi szczegółowe:
    - unikanie i minimalizowanie niesprawiedliwej stronniczości (bias)
    - zapewnienie bezstronności i unikanie dyskryminacji
    - inkluzywność w relacjach z interesariuszami
- **Wymóg nadrzędny 6**: Dobrostan jednostkowy, społeczny i środowiskowy

  wymogi szczegółowe:
    - zrównoważone i przyjazne środowiskowo systemy AI i big data
    - troska o dobrostan jednostkowy, relacje społeczne i spójność społeczną (cohesion)
    - troska o demokratyczne i silne instytucje
- **Wymóg nadrzędny 7**: Odpowiedzialność (accountability)

  wymogi szczegółowe:
    - możliwość kontroli (auditability)
    - minimalizowanie i raportowanie negatywnego oddziaływania
    - wewnętrzne i zewnętrzne ramy governance
    - kwestie zadośćuczynienia za ewentualne szkody
    - zapewnienie nadzoru ze strony człowieka

Na tym poziomie oba rodzaje wymogów pozostają jeszcze dość ogólne, co nie pozwala na ich operacjonalizację (dlatego właśnie na końcu wprowadzony zostaje trzeci rodzaj wymogów, które tutaj dla ułatwienia nazwiemy „finalnymi” – więcej na ten temat za chwilę).

Przejdźmy teraz do zawartego w podejściu SHERPA modelu wdrażania i użytkowania systemów IT w organizacjach. Jest on połączeniem znanych modeli COBIT oraz ITIL, i obejmuje dwa obszary: strategii IT w organizacji (IT governance) oraz zarządzania usługami IT w organizacji (IT management), z których drugi zawiera pięć etapów. Można to przedstawić następująco:

- **obszar strategii IT w organizacji** – domena zarządu
- **obszar zarządzania usługami IT w organizacji** – domena kierownictwa wykonawczego
    - Strategia zarządzania operacyjnego IT
    - Zamawianie albo projektowanie systemu
    - Wdrożenie (implementation) i uruchomienie (deployment) systemu
    - Obsługa systemu
    - Monitorowanie, ewaluacja i udoskonalanie systemu

Jak połączyć ze sobą wskazane powyżej komponenty, tzn. wymogi oraz obszary? Rozwiązanie przyjęte w opracowaniu SHERPA na pierwszy rzut oka może się wydawać nieco skomplikowane. Nie pomaga także niezbyt fortunna decyzja terminologiczna polegająca na wprowadzeniu różnych rodzajów wymogów. Zanim do tego przejdziemy, trzeba jednak jeszcze wyjaśnić, na czym polega wspomniany uprzednio podział na wytyczne ogólne i operacyjne.

Wytyczne ogólne odnoszą się do obszaru strategii IT (IT governance) oraz obszaru zarządzania usługami (IT management) – w tym jego pięciu etapów – i, jak sama nazwa wskazuje, są dość generyczne. Obejmują dziewięć wymogów, które w następnym kroku – czyli w wytycznych operacyjnych – zostają uzupełnione o kolejne 62 wymogi (w sumie owych wymogów jest 72, przy czym niektóre z nich są dodatkowo rozszerzone o podpunkty a, b, c itd.).

Możliwa trudność w zrozumieniu tych relacji wynika po pierwsze stąd, że wytyczne operacyjne (tak jak wytyczne ogólne) także odnoszą się do owych pięciu etapów zarządzania usługami; po drugie zaś z faktu, że na poziomie wytycznych operacyjnych zostają przywołane – opisane wcześniej – wymogi nadrzędne (high-level requirements) oraz szczegółowe (sub-requirements), co może być mylące pod względem terminologicznym (ostatecznie mamy bowiem do czynienia z trzema kategoriami wymogów). Jest to jednak przede wszystkim trudność dotycząca klarowności prezentacji podejścia SHERPA, a nie jego sensowności czy użyteczności. Dla ułatwienia będę owe ostatecznie sformułowane 72 wymogi nazywał poniżej **„finalnymi”** i wyróżniał pogrubieniem.

Omawiane zależności można schematycznie przedstawić w następujący sposób:

- Wytyczne ogólne:

  9 „finalnych” wymogów obejmujących:

  - obszar strategii IT (**wymóg 1**[24])
  - obszar zarządzania usługami i jego pięć etapów (**wymogi 2**–**9**):
    - strategia zarządzania operacyjnego IT (**wymogi**: **2, 3** i **4**)

[24] Jak już wspomniałem, celowo pomijam w tym schemacie treść konkretnych „finalnych” wymogów, ponieważ chodzi mi o przedstawienie relacji zachodzących między poszczególnymi elementami w podejściu SHERPA. W przypadku poszczególnych 62 „finalnych” wymogów zawartych w wytycznych operacyjnych staram się też podawać, do jakich etapów zarządzania usługami IT się odnoszą (co również ma na celu ukazanie związków między elementami składowymi omawianego podejścia), chyba że w danym wypadku jest ich zbyt wiele – wówczas zaznaczam, że konkretne „finalne” wymogi dotyczą różnych etapów zarządzania usługami IT.

- zamawianie albo projektowanie systemu (**wymogi**: **5, 5a, 5b, 5c, 5d i 5e**)
- wdrożenie i uruchomienie systemu (**wymogi**: **6, 6a, 6b, 6c, 6d, 6e**)
- obsługa systemu (**wymóg 7**)
- monitorowanie, ewaluacja i udoskonalanie systemu (**wymogi**: **8** i **9**)

- Wytyczne operacyjne:

62 „finalne" wymogi obejmujące:
obszar zarządzania usługami i jego pięć etapów ujętych w powiązaniu z wymogami nadrzędnymi i wymogami szczegółowymi (**wymogi**: **10**–**72**):

- Wymóg nadrzędny 1: Sprawczość, wolność i godność ludzka
  - wymóg szczegółowy: zapewnienie wolności pozytywnej i ludzkiej sprawczości (**wymóg 10** – dotyczy wszystkich pięciu etapów zarządzania usługami IT)
  - wymóg szczegółowy: zapewnienie wolności negatywnej (**wymóg 11** – dotyczy wszystkich pięciu etapów zarządzania usługami IT)
  - wymóg szczegółowy: ochrona godności człowieka (**wymóg 12** – dotyczy wszystkich etapów zarządzania usługami IT poza etapem obsługi systemu)
- Wymóg nadrzędny 2: Techniczna solidność i bezpieczeństwo
  - wymóg szczegółowy: cyberbezpieczeństwo i odporność na ataki (**wymogi**: **13** i **14** – dotyczą wszystkich pięciu etapów zarządzania usługami IT)
  - wymóg szczegółowy: rozwiązania awaryjne i ogólne bezpieczeństwo (**wymogi**: **15** i **16** – dotyczą wszystkich pięciu etapów zarządzania usługami IT)
  - wymóg szczegółowy: odpowiedniość, wiarygodność i odtwarzalność (**wymogi**: **17** i **18** – dotyczą wszystkich pięciu etapów zarządzania usługami IT, oprócz etapu strategii zarządzania operacyjnego w przypadku wymogu 17)
- Wymóg nadrzędny 3: Prywatność i zarządzanie danymi
  - wymóg szczegółowy: poszanowanie prywatności (**wymogi**: **19**–**29** – dotyczą różnych etapów zarządzania usługami IT)
  - wymóg szczegółowy: jakość i integralność danych (**wymogi**: **30** i **31** – dotyczą różnych etapów zarządzania usługami IT)
  - wymóg szczegółowy: zapewnienie dostępu do danych i praw do nich – w tym prawa własności (**wymogi**: **32**–**34** – dotyczą różnych etapów zarządzania usługami IT)

- wymóg szczegółowy: zapewnienie ochrony praw do danych – w tym prawa własności (**wymóg 35** – dotyczy dwóch etapów zarządzania usługami IT: etapu 2 –zamawianie albo projektowanie systemu, oraz etapu 3 – wdrożenie i uruchomienie systemu)

- Wymóg nadrzędny 4: Przejrzystość
  - wymóg szczegółowy: identyfikowalność (**wymogi**: **36** i **37** – dotyczą różnych etapów zarządzania usługami IT)
  - wymóg szczegółowy: wyjaśnialność (**wymogi**: **38**–**41** – dotyczą różnych etapów zarządzania usługami IT)
  - wymóg szczegółowy: zapewnienie odpowiednich informacji i komunikacji (**wymogi**: **42**–**44** – dotyczą różnych etapów zarządzania usługami IT)
- Wymóg nadrzędny 5: Różnorodność, brak dyskryminacji i bezstronność
  - wymóg szczegółowy: unikanie i minimalizowanie niesprawiedliwej stronniczości (**wymogi**: **45a, 45b, 46**–**49** – dotyczą różnych etapów zarządzania usługami IT)
  - wymóg szczegółowy: zapewnienie bezstronności i unikanie dyskryminacji (**wymogi**: **50**–**54** – dotyczą różnych etapów zarządzania usługami IT)
  - wymóg szczegółowy: inkluzywność w relacjach z interesariuszami (**wymogi**: **55** i **56** – dotyczą różnych etapów zarządzania usługami IT)
- Wymóg nadrzędny 6: Dobrostan jednostkowy, społeczny i środowiskowy
  - wymóg szczegółowy: zrównoważone i przyjazne środowiskowo systemy AI i big data (**wymóg 57** – dotyczy wszystkich pięciu etapów zarządzania usługami IT, a szczególnie etapu 3: uruchomienie systemu i kolejnych)
  - wymóg szczegółowy: troska o dobrostan jednostkowy (**wymogi**: **58** i **59** – dotyczą wszystkich pięciu etapów zarządzania usługami IT)
  - wymóg szczegółowy: troska o dobrostan społeczny (**wymogi**: **60** i **61** – dotyczą różnych etapów zarządzania usługami IT)
  - wymóg szczegółowy: troska o demokratyczne i silne instytucje (**wymóg 62** – dotyczy 3 i 5 etapu zarządzania usługami IT, czyli wdrożenia i uruchomienia systemu oraz jego monitorowania)
- Wymóg nadrzędny 7: Odpowiedzialność
  - wymóg szczegółowy: możliwość kontroli (**wymogi**: **63**–**65** – dotyczą wszystkich pięciu etapów zarządzania usługami IT)
  - wymóg szczegółowy: minimalizowanie i raportowanie negatywnego oddziaływania (**wymogi**: **66** i **67** – dotyczą różnych etapów zarządzania usługami IT)

- wymóg szczegółowy: wewnętrzne i zewnętrzne ramy governance (**wymogi**: **68** i **69** – dotyczą różnych etapów zarządzania usługami IT)
- wymóg szczegółowy: kwestie zadośćuczynienia za ewentualne szkody (**wymóg 70** – dotyczą trzeciego etapu zarządzania usługami IT, czyli wdrożenia i uruchomienia systemu)
- wymóg szczegółowy: zapewnienie nadzoru ze strony człowieka (**wymogi 71** i **72** – dotyczą różnych etapów zarządzania usługami IT)

Poszczególne „finalne" wymogi różnią się objętością czy szczegółowością. Niektóre z nich są dość zwięzłe (np. wymóg 1 dotyczący obszaru strategii i obowiązków ciążących na zarządzie), inne zaś znacznie bardziej rozbudowane (np. wymóg 45a, który dotyczy oszacowania niesprawiedliwej stronniczości systemu).

Tym, co w praktyce odróżnia omawiane tu podejście SHERPA od podejścia Ethics by Design wg SIENNA (notabene oba projekty były prowadzone pod kierownictwem tego samego grona badaczy), jest większy wysiłek w przypadku tego pierwszego podejścia potrzebny do zmapowania poszczególnych „finalnych" wymogów z konkretnymi etapami projektowymi (w tym przypadku: etapami zarządzania usługami IT). Zastosowanie owych wytycznych będzie więc wymagało od osób zainteresowanych dodatkowej pracy, co jednak może zwiększyć gwarancję, że cały proces zostanie zrealizowany w sposób przemyślany i dostosowany do specyfiki i potrzeb konkretnego zespołu projektowego czy danej organizacji.

Opracowano na podstawie: Shaping the ethical dimensions of smart information systems – a European perspective (SHERPA), Guidelines for the Ethical Use of AI and Big Data Systems.

Rekomendacje dla administracji centralnej w zakresie wdrażania etycznej sztucznej inteligencji

Robert Sroka

Wymogi godnej zaufania sztucznej inteligencji (trustworthy AI) dotyczą nie tylko sektora prywatnego, ale również publicznego. W wielu przypadkach sektor publiczny powinien być nie tylko przykładem, lecz również wymagać od swoich partnerów prywatnych rozwijania i wykorzystania AI zgodnie z uznanymi zasadami etycznymi. Dlatego warto, aby sektor publiczny, w szczególności administracja oraz instytucje publiczne rozważyły wdrożenie poniższych rekomendacji.

I. Powołanie Centralnej Rady Etycznej AI dla administracji państwowej i samorządowej oraz instytucji publicznych

Powołanie na poziomie ministerialnym lub KPRM opiniodawczo-doradczej Centralnej Rady Etycznej AI dla administracji państwowej i samorządowej oraz instytucji publicznych odpowiedzialnej za koordynację i monitorowanie etycznych aspektów wdrażania AI w administracji oraz instytucjach publicznych:

> Utworzenie w ramach Centralnej Rady Etycznej AI międzyresortowego oraz eksperckiego komitetu sterującego, który będzie regularnie spotykać się, aby monitorować aktualne i planowane projekty, omawiać postępy, rozwiązywać problemy etyczne i koordynować działania mitygujące negatywne konsekwencje etyczne.

Przykładowy zakres zadań i kompetencji Centralnej Rady Etycznej AI dla administracji państwowej i samorządowej oraz instytucji publicznych:

1. Standardy wdrożeniowe etycznego AI dla administracji oraz instytucji publicznych:
   a. Standaryzacja procesów - opracowanie jednolitych standardów i protokołów wdrożeniowych dla wszystkich instytucji, aby zapewnić spójność i zgodność z najlepszymi praktykami etycznymi na bazie już wypracowanych międzynarodowych standardów;
   b. Wytyczne technologiczne - wypracowanie wytycznych technologicznych, które będą ułatwiać integrację systemów AI z istniejącymi systemami IT w różnych instytucjach;

c. Wytyczne konsultacyjne – wypracowanie procesu etycznej konsultacji wstępnej oraz okresowej wypracowywanych lub wdrażanych przez administrację i instytucje publiczne rozwiązań AI z Centralną Radą Etyczną;

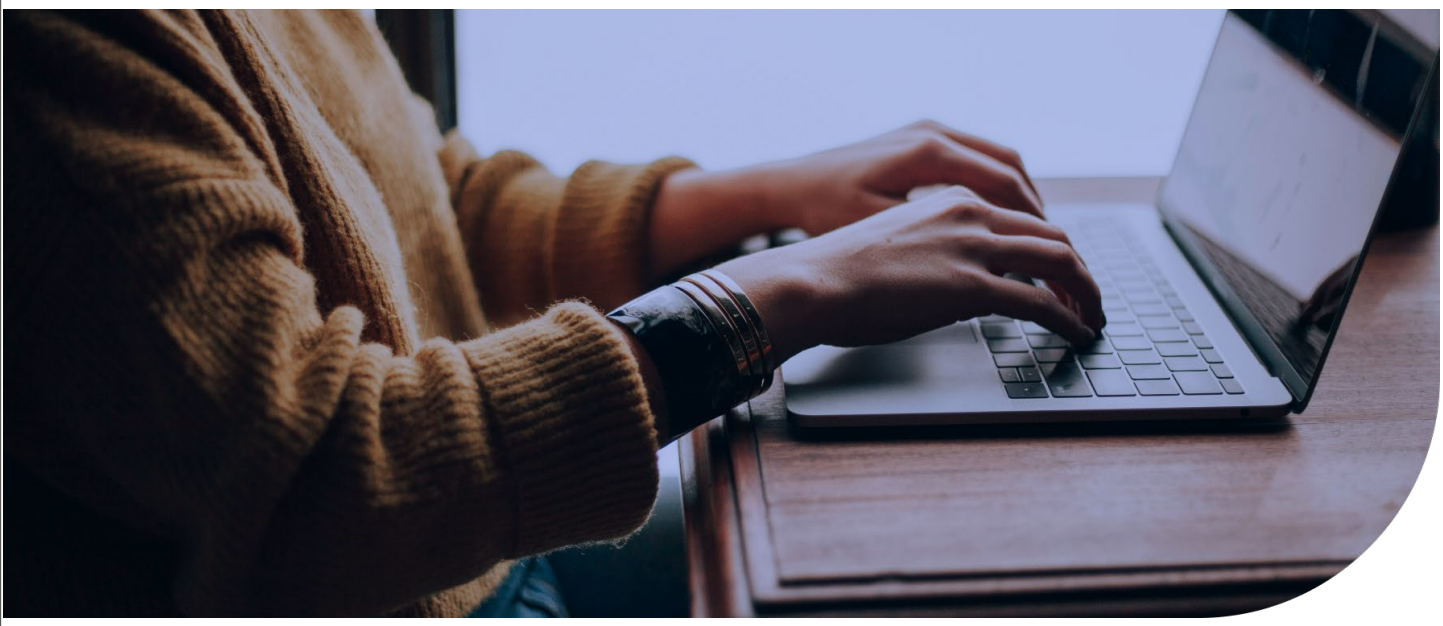
2. Systemy Monitorowania i Ewaluacji:
   a. Platforma monitorowania - Wdrożenie zintegrowanej platformy monitorowania, która pozwoli na śledzenie działania i wydajności systemów AI w zakresie spełniania standardów etycznych;
   b. Metryki - Ustanowienie kluczowych wskaźników wydajności (KPI) i metryk do oceny skuteczności i wpływu społecznego wdrażanych rozwiązań AI;
   c. Przygotowanie listy najważniejszych instytucji państwa wykorzystujących systemy AI istotnych dla oddziaływania społecznego (m.in. ZUS, NFZ, policja, sądy) oraz wstępna i okresowa ocena etyczna prowadzona przez niezależny panel ekspertów;
3. Współpraca Międzyinstytucjonalna:
   a. grupy robocze - Tworzenie tematycznych grup roboczych, które będą zajmować się specyficznymi wyzwaniami wdrożeniowymi w różnych sektorach, takich jak zdrowie, sądownictwo, czy administracja;
   b. Wymiana doświadczeń - Organizowanie regularnych spotkań i warsztatów w celu wymiany doświadczeń i najlepszych praktyk między instytucjami;
4. Programy szkoleniowe i rozwój kompetencji:
   a. Opracowanie i wdrożenie programów szkoleniowych dla pracowników publicznych na wszystkich szczeblach, aby zwiększyć ich kompetencje w zakresie etyki AI;
   b. Certyfikacje: Wprowadzenie certyfikacji dla specjalistów zajmujących się AI w sektorze publicznym, aby zapewnić odpowiedni poziom wiedzy i umiejętności w zakresie etycznego rozwoju, wdrażania i wykorzystywania AI;
5. Monitorowanie spełniania wymagań etycznych:
   a. Regularny monitoring - Przeprowadzanie regularnych ocen systemów AI, aby ocenić ich zgodność z ustalonymi etycznymi standardami i wytycznymi;
   b. Raportowanie - Ustanowienie regularnego raportowania o postępach i wynikach wdrożeń AI do Centralnej Rady Etycznej AI oraz odpowiednich organów nadzorczych;
6. Udział społeczny i transparentność:
   a. Konsultacje społeczne i eksperckie - Organizowanie konsultacji eksperckich i publicznych w celu zbierania opinii i sugestii od obywateli dotyczących wdrażanych rozwiązań AI;
   b. Transparentność działań - Zapewnienie transparentności wdrażania AI poprzez regularne publikowanie raportów i aktualizacji dotyczących;
7. Reagowanie na nadużycia i nieprawidłowości:

a. System zgłaszania nieprawidłowości – ustanowienie mechanizmu zgłaszania nadużyć i nieprawidłowości etycznych rozwiązań AI wykorzystywanych przez administrację i instytucje publiczne;
b. Prowadzenie postępowań wyjaśniających – opracowanie procedury przyjmowania i rozpatrywania zgłoszeń.

II. **Wdrożenie do reguł udzielania pomocy publicznej, w tym konkursów grantowych, udzielanych przedsiębiorcom, instytucjom publicznym i administracji wymogów uwzględniania aspektów etycznych**:
a. Na etapie opracowania i oceniania zgłoszenia konkursowego;
b. Na etapie realizacji finansowego przy wsparciu pieniędzy publicznych projektu wykorzystującego AI.

III. **Wsparcie badań w zakresie etyki AI:**
a. Stypendia - Finansowanie badań nad etycznymi aspektami AI oraz wspieranie innowacji, które promują odpowiedzialne wykorzystanie AI;
b. Granty – wsparcie centrów badawczych i think tanków skoncentrowanych na etycznym rozwoju AI.

IV. **Edukacja społeczna:**
Kampania edukacyjna – przeprowadzenie kampanii społecznej edukującej w zakresie etycznego wykorzystywania rozwiązań AI przez administrację i instytucje publiczne.



Godna zaufania AI

Grupa robocza ds. Sztucznej Inteligencji przy Ministerstwie Cyfryzacji

Warszawa, styczeń 2025

Ministerstwo Cyfryzacji | Tu tworzymy przyszłość

GRAI GRUPA ROBOCZA DS. SZTUCZNEJ INTELIGENCJI