

Streszczenie

Od minimalnej adresowalności proceduralnej do rozliczalności. Warunki przypisywania odpowiedzialności systemom sztucznej inteligencji

Przedmiotem rozprawy jest analiza warunków, w których można rozważyć przypisywanie systemom sztucznej inteligencji słabej odpowiedzialności proceduralnej oraz silnej rozliczalności. Główne pytanie badawcze dotyczy tego, jakie rozpoznawalne warunki minimalnej adresowalności powinien spełniać system sztucznej inteligencji, aby można było przypisać mu słabą odpowiedzialność proceduralną oraz jakie dodatkowe warunki stabilizacji, silnej rozliczalności i zastosowania instytucjonalnego określają zakres tej odpowiedzialności. Rozprawa odróżnia odpowiedzialność systemów sztucznej inteligencji od pełnej odpowiedzialności moralnej człowieka oraz neutralności technicznego narzędzia, wskazując, że właściwszy jest model progowy. Minimalna adresowalność proceduralna jest pierwszym warunkiem formułowania wobec systemu roszczeń o uzasadnienie, ocenę i korektę, nie przesądza jednak o pełnej podmiotowości moralnej, osobowości prawnej ani samodzielnej rozliczalności instytucjonalnej.

Rozprawa wychodzi od diagnozy, że współczesne systemy sztucznej inteligencji, zwłaszcza systemy adaptacyjne i generatywne, ujawniają ograniczenia klasycznych modeli odpowiedzialności. Ich działania mogą mieć znaczenie społeczne, prawne i moralne, choć procesy prowadzące do decyzji bywają nieprzejrzyste, rozproszone i trudne do odtworzenia. Powstaje więc napięcie między oddziaływaniem systemów na dobra chronione normami a tradycyjnym wymogiem przypisywania odpowiedzialności podmiotowi zdolnemu do działania intencjonalnego, rozumienia norm oraz podlegania uzasadnionej krytyce. Rozprawa nie zakłada jednak, że samo społeczne oddziaływanie systemów sztucznej inteligencji wystarcza do uznania ich za podmioty moralne. Rozstrzygające jest pytanie, czy decyzje i uzasadnienia systemu można włączyć do publicznych praktyk zadawania pytania „dlaczego?”, oceny racji oraz korekty działań. W tym sensie „adresowalność proceduralna” oznacza minimalną możliwość traktowania systemu jako punktu odniesienia w procedurach uzasadniania, audytu oraz korekty decyzji.

Pod względem metodologicznym praca ma charakter analityczno-koncepcyjny. Autor nie prowadzi badań empirycznych ani eksperymentów i nie projektuje rozwiązań technicznych.

Przykłady technologiczne pełnią funkcję ilustracyjną i heurystyczną. W przypadku IDx-DR/LumineticsCore opublikowane wyniki kliniczne i dokumentacja regulacyjna stanowią podstawę wyłącznie twierdzeń opisowych o systemie; zastosowanie H1–H3 + WI pozostaje analizą pojęciową i nie stanowi empirycznej walidacji modelu. Podstawowe metody to analiza pojęciowa, rekonstrukcja argumentów, ocena porównawcza stanowisk oraz sprawdzanie spójności modelu. Praca rekonstruuje spory o podmiotowość, wolną wolę, świadomość, sprawczość i odpowiedzialność, a następnie wykorzystuje ich ustalenia w filozofii sztucznej inteligencji. Odwołuje się do literatury z filozofii moralnej, teorii działania, filozofii umysłu, etyki sztucznej inteligencji, filozofii techniki oraz regulacji sztucznej inteligencji. Odwołania do dokumentów prawnych służą weryfikacji instytucjonalnej stosowalności modelu, jednak nie zastępują argumentacji filozoficznej.

Rozdział 1 określa problem badawczy, hipotezy robocze oraz zakres pracy. Spór koncentruje się na pytaniu, czy systemy sztucznej inteligencji są wyłącznie narzędziami, czy też w określonych warunkach mogą funkcjonować jako quasi-podmioty w praktykach odpowiedzialności. Rozdział jednoznacznie odróżnia sprawczość przyczynową od sprawczości normatywnej. Odpowiedzialności nie można przypisywać wyłącznie na podstawie skutków działania. Wprowadzono trzy główne hipotezy oraz warunek instytucjonalny. H1 określa minimalną adresowalność proceduralną jako podstawę słabej odpowiedzialności proceduralnej. H2 ujmuje afektywność funkcjonalną jako możliwy mechanizm stabilizujący wybory zgodne z racjami, natomiast K4 odrębnie ocenia stabilność diachroniczną i wiarygodność profilu działania. H3 definiuje próg silnej rozliczalności, uzależniony od kontroli metapoznawczej oraz jawnego rozstrzygnięcia konfliktów racji. Warunek WI wyznacza granice stosowania modelu H1–H3 w praktykach publicznych, organizacyjnych i regulacyjnych. Model ma charakter przyszłościowy: określa warunki możliwej odpowiedzialności, nie dowodzi ich spełnienia przez obecne systemy sztucznej inteligencji.

Rozdział 2 rekonstruuje pojęcie podmiotowości w wybranych koncepcjach filozoficznych. Analiza obejmuje trzy funkcje podmiotu, czyli bycie nośnikiem stanów, bycie sprawcą działań zależnych od racji oraz bycie adresatem roszczeń i odpowiedzialności. Minimalne warunki odpowiedzialności praktycznej nie wymagają silnej ontologii „głębokiego ja”. Kluczowe są publiczna rozpoznawalność adresata, możliwość uzasadnienia działania oraz kontrola zależna od

racji. Perspektywa drugiej osoby jest szczególnie istotna. Odpowiedzialność oznacza uczestnictwo w praktykach wzajemnego domagania się racji, odpowiedzi, krytyki, usprawiedliwienia oraz korekty. Rozdział uwzględnia też rozróżnienie między świadomością fenomenalną a świadomością dostępu. Świadomość fenomenalna jest ważna przy krzywdzie doznaniowej, lecz nie stanowi koniecznego warunku minimalnej adresowalności proceduralnej. H1 wymaga funkcjonalnego rozpoznawania norm, przewidywania konsekwencji ich naruszeń, kontrfaktycznego reagowania na racje oraz udostępniania wystarczających i wiarygodnych podstaw decyzji, pozwalających ocenić ich związek z wynikiem oraz umożliwiających audyt, zakwestionowanie i korektę.

Rozdział 3 określa operacyjno-diagnostyczne kryteria sztucznej podmiotowości moralnej. Kryteria K1–K3 wspierają ocenę spełnienia H1, natomiast K4 odrębnie ocenia stabilność diachroniczną i wiarygodność profilu działania; nie tworzą one dodatkowych hipotez. K1 to autokorekta norm oraz audytowalna korekta decyzji. K2 oznacza wrażliwość na racje, które wpływają na rozstrzygnięcia. K3 odnosi się do refleksyjności predykcyjnej oraz powiązania deliberacji z działaniem. K4 oznacza stabilność diachroniczną, czyli trwałość profilu decyzyjnego oraz możliwość rekonstrukcji historii zmian. Rozdział odwołuje się do typologii agentów etycznych Jamesa Moora oraz do koncepcji poziomów abstrakcji Luciana Floridiego i Jeffa Sandersa. Poziom agenta jawnego stanowi ramę opisu minimalnego progu, lecz sama kwalifikacja typologiczna nie zastępuje wykazania zdolności wymaganych przez H1 ani spełnienia warunków instytucjonalnych WI. Podejście eliminuje zarówno nadmierną antropomorfizację sztucznej inteligencji, jak i sprowadzanie odpowiedzialności do technicznej skuteczności.

Rozdział 4 przedstawia pluralistyczną ramę normatywną. Rama nie rozstrzyga sporu między deontologią, etyką cnót i konsekwencjalizmem, lecz porządkuje ich funkcje w analizie działań systemów sztucznej inteligencji. Deontologia wyznacza granice dopuszczalności i chroni status adresatów norm. Etyka cnót, interpretowana nieantropomorficznie, opisuje stabilność i wiarygodność praktyk instytucjonalnych, w których projektuje się, wdraża i nadzoruje systemy sztucznej inteligencji. Konsekwencjalizm rozstrzyga konflikty racji w obrębie dopuszczalnych wariantów działania. W rozprawie proponuje się porządek LexOrder: pierwszeństwo mają warunki dopuszczalności działań i adresowalności normatywnej, następnie stabilność i wiarygodność praktyk, a dopiero po ich spełnieniu porównuje się dopuszczalne opcje pod kątem

skutków i ryzyka. FairAim uzupełnia tę autorską propozycję o wymóg bezstronnego ustalania zasad i priorytetów. PH1–PH3 stanowią etyczno-normatywną warstwę uzasadnienia modelu H1–H3, lecz nie są prostymi odpowiednikami hipotez. PH1 wyznacza granice dopuszczalności działań, natomiast H1 określa wymagane zdolności systemu; dopuszczalny wynik sam przez się nie dowodzi spełnienia H1. PH2 uzasadnia normatywne znaczenie stabilności i wiarygodności praktyk, nie jest jednak mechanizmem afektywności funkcjonalnej H2. PH3 porządkuje konflikty racji w obrębie dopuszczalnych opcji, nie stanowi jednak metakontroli właściwej H3.

Rozdział 5 analizuje sztuczną podmiotowość w ujęciu klasyfikacyjnym i instytucjonalnym. Rozróżnienie między „agentem” a „podmiotem” jest kluczowe. Agent to kategoria opisowa, odnosząca się do systemu wykazującego interaktywność, autonomię oraz adaptacyjność na określonym poziomie opisu, bez gwarancji normatywnego statusu. Podmiot jest kategorią normatywno-instytucjonalną; wobec niego można kierować roszczenia, oczekiwać uzasadnień oraz egzekwować odpowiedzialność. To rozróżnienie pokazuje, że nie każdy agent jest podmiotem, któremu można zasadnie przypisać odpowiedzialność w ramach praktyk instytucjonalnych. Rozdział przedstawia ramę klasyfikacyjną, łączącą ontologiczny i funkcjonalny wymiar sztucznej podmiotowości z warunkami interwencji oraz interpretowalności. Szczególne znaczenie mają wymogi dotyczące znaczącej kontroli człowieka oraz minimalnej interpretowalności. Analiza odróżnia podstawowy Delphi, przewidujący ludzkie osądy moralne, od DelphiHYBRID, który łączy predykcje z regułami symbolicznymi. Jawność reguł sama nie dowodzi spełnienia H1, F3 ani H3. Wniosek rozdziału jest ostrożny. System sztucznej inteligencji może pełnić rolę quasi-podmiotu proceduralnego tylko w ograniczonym sensie, po odrębnym wykazaniu H1 i F3 oraz zapewnieniu odpowiednich parametrów R i I w ramach WI. Nie oznacza to przyznania mu pełnej osobowości moralnej ani prawnej. Przeciwnie, ma zapobiegać zdejmowaniu odpowiedzialności z ludzi i instytucji oraz przerzucaniu sankcji na artefakty. Podrozdział 5.4 stosuje tę ramę do IDx-DR/LumineticsCore, sklasyfikowanego jako typ I: A2/A4 × F2, czyli normatywnie nasycony instrument w rozliczalnym układzie społeczno-technicznym. Przypadek pozwala odróżnić audytowalność systemu i rozliczalność instytucjonalną od agentowej adresowalności proceduralnej: nie wykazano spełnienia H1 ani H2, a H3 pozostaje niespełniona w świetle udokumentowanej funkcjonalności systemu.

Rozdział 6 bada emulacyjny, ewolucyjny i rewolucyjny scenariusz rozwoju sztucznej inteligencji. Scenariusze te nie są traktowane jako prognozy empiryczne, lecz jako modele analityczne ujawniające odmienne typy problemów związanych z odpowiedzialnością. Scenariusz emulacji mózgu zaostreza pytania o ciągłość tożsamości, powielanie podmiotów oraz rozdział odpowiedzialności między emulowanymi umysłami a instytucjami sprawującymi nad nimi kontrolę. Scenariusz ewolucyjny pokazuje stopniowe narastanie złożoności i rozmywania odpowiedzialności w układach socjotechnicznych. Scenariusz rewolucyjny, związany z ideą osobliwości lub eksplozji inteligencji, pełni funkcję granicy ostrzegawczej. Ukazuje bowiem konfiguracje, w których tempo działania systemów może przekraczać zdolności ludzkiego nadzoru, refleksji i korekty. Rozdział rozwija także analizę luki odpowiedzialności. Luka ta może obejmować brak odpowiedzialnego podmiotu, lecz nie wyczerpuje się w tej postaci: jest ujmowana jako wielowymiarowe naruszenie warunków przypisywania odpowiedzialności. Służą temu jakościowe metryki M1–M5, obejmujące identyfikowalność podmiotów odpowiedzialnych, audytowalność racji, kontrolę zależną od racji, przejrzystość i trwałość praktyk przypisywania odpowiedzialności oraz relację między tempem działania lub zmian a czasem i wiedzą dostępnymi na refleksję, kontrolę i reakcję. Analiza odróżnia zdolności systemu technicznego od instytucjonalnych warunków stosowania modelu; M4 dotyczy praktyk odpowiedzialności, podczas gdy K4 ocenia stabilność profilu działania systemu. Odpowiedzią na lukę odpowiedzialności nie jest nadawanie sztucznej inteligencji osobowości prawnej, lecz projektowanie modeli współodpowiedzialności, w których role ludzi, instytucji i systemów technicznych są jawne, podlegają audytowi i mogą zostać skorygowane.

Rozdział 7 syntetyzuje wyniki i przeprowadza filozoficzną weryfikację H1–H3 oraz W1. H1 zostaje potwierdzona jako warunek konieczny minimalnej adresowalności proceduralnej. Bez zdolności rozpoznawania racji, ich ujawniania i poddawania decyzji korekcie nie ma podstaw, by kierować do systemu roszczenia o uzasadnienie. H1 nie wystarcza jednak do silnej rozliczalności. H2 uzyskuje uzasadnienie jako możliwy mechanizm stabilizujący wybory zgodne z racjami, lecz nie jest warunkiem koniecznym minimalnego progu. Stabilność diachroniczna i wiarygodność profilu działania podlegają odrębnej ocenie K4, której wynik nie dowodzi sam przez się spełnienia H2. H3 zostaje potwierdzona jako próg silnej rozliczalności, wymagający kontroli metapoznawczej, jawnego rozstrzygania konfliktów racji oraz zdolności przedstawienia

uzasadnienia z perspektywy drugiego rzędu. WI dopełnia model, wskazując, że nawet spełnienie H1–H3 nie wystarcza bez publicznie rozpoznawalnych procedur audytu, nadzoru, korekty i przypisania odpowiedzialności ludziom oraz instytucjom. Ilustracyjne odwołanie do IDx-DR/LumineticsCore pokazuje, że typ I: $A2/A4 \times F2$ może wykazywać audytowalność i rozbudowane osadzenie instytucjonalne bez wykazanego przekroczenia progów agentowych. Wynik ten ilustruje rozróżnienia modelu, nie stanowi jego empirycznej walidacji.

Końcowym rezultatem rozprawy jest hierarchia statusów moralnych sztucznej inteligencji ujęta w dwutorowym modelu. Oś sprawczości porządkuje role proceduralne od artefaktu istotnego moralnie, przez instrument normatywnie nasycony, aż po quasi-podmiot proceduralny. Oś podatności na krzywdę porządkuje możliwy zakres ochrony od braku szczególnej ochrony, przez ochronę relacyjną, po ostrożne podejście w warunkach niepewności dotyczących krzywdy. Hierarchia ta nie prowadzi do automatycznego przyznawania sztucznej inteligencji praw ani osobowości. Pozwala natomiast precyzyjniej określać obowiązki ludzi i instytucji wobec systemów, użytkowników oraz osób dotkniętych skutkami decyzji podejmowanych przez algorytmy. Wkład autorski rozprawy polega na zaproponowaniu modelu progowego odpowiedzialności systemów sztucznej inteligencji. Model ten odróżnia minimalną adresowalność proceduralną od silnej rozliczalności, oddziela warunki sprawcze od warunków zastosowania instytucjonalnego oraz uzupełnia lukę w rozumieniu odpowiedzialności jako zjawiska stopniowalnego i wielowymiarowego. Zachowuje minimalizm ontologiczny, lecz nie obniża wymagań normatywnych. Odpowiedzialność pozostaje praktyką publicznego uzasadniania, korekty i rozliczania, której ostateczny ciężar spoczywa na ludziach oraz instytucjach organizujących działanie systemów sztucznej inteligencji.

Słowa kluczowe: sztuczna inteligencja, odpowiedzialność, adresowalność proceduralna, rozliczalność, podmiotowość moralna, luka odpowiedzialności, etyka AI, znacząca kontrola człowieka, interpretowalność.