

**UNIWERSYTET KOMISJI EDUKACJI NARODOWEJ
W KRAKOWIE**

SZKOŁA DOKTORSKA

DYSCYPLINA: Filozofia

Maciej Czyszczoń

ROZPRAWA DOKTORSKA

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

Promotor: dr hab. Andrzej Dąbrowski, prof. UKEN

Kraków, 2026

**UNIVERSITY
OF THE NATIONAL
EDUCATION COMMISSION, KRAKOW**

DOCTORAL SCHOOL

DISCIPLINE: Philosophy

Maciej Czyszczon

DOCTORAL DISSERTATION

**From Minimal Procedural Addressability to Accountability: Conditions for
Attributing Responsibility to Artificial Intelligence Systems**

Supervisor: dr hab. Andrzej Dąbrowski, prof. UKEN

Krakow, 2026

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 rozstrzygania 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 predykccyjnej 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: $A2/A4 \times F2$, 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 zaostrza 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 WI. 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 rozstrzygnięcia 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.

Abstract

From Minimal Procedural Addressability to Accountability: Conditions for Attributing Responsibility to Artificial Intelligence Systems

The subject of the dissertation is an analysis of the conditions under which one may consider attributing weak procedural responsibility and strong accountability to artificial intelligence systems. The main research question concerns which recognizable conditions of minimal addressability an artificial intelligence system should meet for weak procedural responsibility to be attributed to it, and which additional conditions of stabilization, strong accountability, and institutional application determine the scope of that responsibility. The dissertation distinguishes the responsibility of artificial intelligence systems from both full human moral responsibility and the neutrality of a technical tool, arguing that a threshold model is more appropriate. Minimal procedural addressability is the first condition for formulating claims toward a system for justification, assessment, and correction; however, it does not determine full moral subjecthood, legal personhood, or independent institutional accountability.

The dissertation begins from the diagnosis that contemporary artificial intelligence systems, especially adaptive and generative systems, reveal the limitations of classical models of responsibility. Their actions may have social, legal, and moral significance, although the processes leading to decisions are often opaque, distributed, and difficult to reconstruct. A tension therefore arises between the impact of systems on goods protected by norms and the traditional requirement that responsibility be attributed to a subject capable of intentional action, understanding norms, and being subject to justified criticism. The dissertation does not assume, however, that the mere social impact of artificial intelligence systems is sufficient to recognize them as moral subjects. The decisive question is whether the system's decisions and justifications can be incorporated into public practices of asking "why?", assessing reasons, and correcting actions. In this sense, "procedural addressability" means the minimal possibility of treating the system as a point of reference in procedures of justification, audit, and decision correction.

Methodologically, the work is analytical and conceptual. The author conducts no empirical research or experiments and does not design technical solutions. Technological examples serve illustrative and heuristic purposes. In the case of IDx-

DR/LumineticsCore, published clinical findings and regulatory documentation provide evidence only for descriptive claims about the system; applying H1–H3 + WI remains a conceptual analysis and does not empirically validate the model. The basic methods are conceptual analysis, argument reconstruction, comparative assessment of positions, and testing the model's coherence. The work reconstructs disputes about subjecthood, free will, consciousness, agency, and responsibility, and then applies their findings to the philosophy of artificial intelligence. It draws on literature in moral philosophy, action theory, philosophy of mind, ethics of artificial intelligence, philosophy of technology, and artificial intelligence regulation. References to legal documents serve to verify the model's institutional applicability; they do not, however, replace philosophical argumentation.

Chapter 1 defines the research problem, working hypotheses, and scope of the work. The dispute centers on whether artificial intelligence systems are merely tools or, under certain conditions, may function as quasi-subjects in practices of responsibility. The chapter clearly distinguishes between causal and normative agency. Responsibility cannot be attributed solely based on the effects of an action. Three main hypotheses and an institutional condition are introduced. H1 defines minimal procedural addressability as the basis of weak procedural responsibility. H2 treats functional affectivity as a possible mechanism for stabilizing choices consistent with reasons, while K4 separately assesses diachronic stability and the reliability of the system's behavioral profile. H3 defines the threshold of strong accountability, which depends on metacognitive control and explicit resolution of conflicts between reasons. The WI condition sets the limits for applying the H1–H3 model in public, organizational, and regulatory practices. The model is prospective: it specifies conditions for possible responsibility rather than establishing that current artificial intelligence systems satisfy them.

Chapter 2 reconstructs the concept of subjecthood in selected philosophical accounts. The analysis covers three functions of the subject, namely being a bearer of states, being an agent of actions dependent on reasons, and being an addressee of claims and responsibility. The minimal conditions of practical responsibility do not require a strong ontology of the “deep self.” What is crucial is the public recognizability of the addressee, the possibility of justifying action, and reason-responsive control. The second-person perspective is especially important. Responsibility means participation in practices

of mutually demanding reasons, responses, criticism, justification, and correction. The chapter also considers the distinction between phenomenal consciousness and access consciousness. Phenomenal consciousness is important in cases of experiential harm, but it is not a necessary condition for minimal procedural addressability. H1 requires the functional recognition of norms, anticipation of the consequences of violating them, counterfactual responsiveness to reasons, and the provision of sufficient and reliable grounds for decisions that allow their connection to the outcome to be assessed and enable audit, challenge, and correction.

Chapter 3 defines diagnostic and operational criteria of artificial moral subjecthood. Criteria K1–K3 support the assessment of whether H1 is satisfied, whereas K4 separately assesses diachronic stability and the reliability of the system’s behavioral profile; they do not constitute additional hypotheses. K1 refers to the self-correction of norms and auditable correction of decisions. K2 denotes responsiveness to reasons that influence decisions. K3 concerns predictive reflexivity and the connection between deliberation and action. K4 denotes diachronic stability, that is, the durability of the decision profile and the possibility of reconstructing the history of changes. The chapter refers to James Moor’s typology of ethical agents and to Luciano Floridi and Jeff Sanders’s concept of Levels of Abstraction. The explicit-agent level provides a descriptive framework for the minimal threshold, but that typological classification does not replace demonstrating the capabilities required by H1 or satisfying the institutional conditions specified by WI. This approach eliminates both the excessive anthropomorphization of artificial intelligence and the reduction of responsibility to technical effectiveness.

Chapter 4 presents a pluralistic normative framework. The framework does not resolve the dispute between deontology, virtue ethics, and consequentialism, but orders their functions in the analysis of the actions of artificial intelligence systems. Deontology sets the limits of permissibility and protects the status of the addressees of norms. Virtue ethics, interpreted non-anthropomorphically, describes the stability and reliability of institutional practices in which artificial intelligence systems are designed, deployed, and supervised. Consequentialism resolves conflicts between reasons within the range of permissible variants of action. The dissertation proposes LexOrder: priority is given to the conditions of permissible action and normative addressability, followed by the

stability and reliability of practices; only after these conditions are met are permissible options compared in terms of consequences and risk. FairAim complements this original proposal with a requirement for the impartial selection of principles and priorities. PH1–PH3 provide the ethical and normative justification for the H1–H3 model but are not direct counterparts of the hypotheses. PH1 sets the limits of permissible action, whereas H1 specifies the capabilities required of a system; a permissible outcome does not, by itself, establish that H1 is satisfied. PH2 justifies the normative importance of stable and reliable practices; it is not the mechanism of functional affectivity specified by H2. PH3 conflicts between reasons within the range of permissible options but does not constitute the metacontrol required by H3.

Chapter 5 analyzes artificial subjecthood in classificatory and institutional terms. The distinction between an “agent” and a “subject” is crucial. An agent is a descriptive category referring to a system that exhibits interactivity, autonomy, and adaptability at a given Level of Abstraction, without any guarantee of normative status. A subject is a normative-institutional category; claims can be directed toward it, justifications can be expected from it, and responsibility can be enforced against it. This distinction shows that not every agent is a subject capable of bearing responsibility within institutional practices. The chapter presents a classificatory framework that links the ontological and functional dimensions of artificial subjecthood with conditions of intervention and interpretability. Requirements concerning meaningful human control and minimal interpretability are especially important. The analysis distinguishes the basic Delphi model, which predicts human moral judgments, from DelphiHYBRID, which combines predictions with symbolic rules. The explicitness of rules does not, by itself, establish that H1, F3, or H3 is satisfied. The conclusion of the chapter is cautious. An artificial intelligence system may function as a procedural quasi-subject only in a limited sense, after H1 and F3 have been demonstrated separately and appropriate R and I parameters have been ensured within WI. This does not mean granting full moral or legal personhood. On the contrary, it is meant to prevent responsibility from being removed from humans and institutions and sanctions from being shifted onto artifacts. Section 5.4 applies this framework to IDx-DR/LumineticsCore, classified as type I: $A2/A4 \times F2$, a normatively saturated instrument within an accountable sociotechnical arrangement. The case distinguishes system auditability and institutional accountability from agential procedural addressability:

satisfaction of H1 and H2 has not been established, and H3 remains unsatisfied considering the system's documented functionality.

Chapter 6 examines the emulation, evolutionary, and revolutionary scenarios of artificial intelligence development. These scenarios are not treated as empirical forecasts, but as analytical models that reveal different types of problems related to responsibility. The brain emulation scenario sharpens questions about continuity of identity, the multiplication of subjects, and the division of responsibility between emulated minds and the institutions that control them. The evolutionary scenario shows the gradual growth of complexity and the diffusion of responsibility in sociotechnical systems. The revolutionary scenario, associated with the idea of the singularity or intelligence explosion, functions as a warning boundary. It shows configurations in which the pace of system activity may exceed the capacities of human supervision, reflection, and correction. The chapter also develops an analysis of the responsibility gap. This gap may include the absence of a responsible subject but is not exhausted by that form: it is understood as a multidimensional violation of the conditions for attributing responsibility. The analysis uses qualitative metrics M1–M5, covering the identifiability of responsible actors, the auditability of reasons, reason-responsive control, the transparency and continuity of responsibility-attribution practices, and the relation between the pace of action or change and the time and knowledge available for reflection, control, and response. It distinguishes the capabilities of the technical system from the institutional conditions for applying the model; M4 concerns responsibility practices, whereas K4 assesses the stability of the system's behavioral profile. The answer to the responsibility gap is not to grant legal personhood to artificial intelligence, but to design models of co-responsibility in which the roles of humans, institutions, and technical systems are explicit, subject to audit, and open to correction.

Chapter 7 synthesizes the results and conducts a philosophical verification of H1–H3 and WI. H1 is confirmed as a necessary condition for minimal procedural addressability. Without the ability to recognize reasons, disclose them, and subject decisions to correction, there is no basis for directing claims for justification toward a system. H1, however, is not sufficient for strong accountability. H2 is justified as a possible mechanism for stabilizing choices consistent with reasons but is not a necessary condition of the minimal threshold. Diachronic stability and the reliability of the system's

behavioral profile are assessed separately through K4; the result of that assessment does not, by itself, establish that H2 is satisfied. H3 is confirmed as the threshold of strong accountability, requiring metacognitive control, explicit resolution of conflicts between reasons, and the ability to present justification from a second-order perspective. WI completes the model by showing that even the fulfillment of H1–H3 is not sufficient without publicly recognizable procedures of audit, supervision, correction, and attribution of responsibility to humans and institutions. The illustrative reference to IDx-DR/LumineticsCore shows that type I: $A2/A4 \times F2$ may exhibit auditability and extensive institutional embedding without demonstrated satisfaction of the agential thresholds. This result illustrates the model's distinctions; it does not empirically validate the model.

The result of the dissertation is a hierarchy of the moral statuses of artificial intelligence, formulated as a two-track model. The axis of agency orders procedural roles from a morally significant artifact, through a normatively saturated instrument, to a procedural quasi-subject. The axis of vulnerability to harm orders the possible scope of protection from no special protection, through relational protection, to a precautionary approach under uncertainty about harm. This hierarchy does not lead to the automatic granting of rights or personhood to artificial intelligence. Rather, it allows the duties of humans and institutions toward systems, users, and people affected by algorithmic decisions to be defined more precisely. The author's contribution consists in proposing a threshold model of responsibility for artificial intelligence systems. This model distinguishes minimal procedural addressability from strong accountability, separates agential conditions from conditions of institutional application, and fills a gap in the understanding of responsibility as a gradual and multidimensional phenomenon. It preserves ontological minimalism without lowering normative requirements. Responsibility remains a practice of public justification, correction, and accountability, whose ultimate burden rests on humans and the institutions that organize the operation of artificial intelligence systems.

Keywords: artificial intelligence, responsibility, procedural addressability, accountability, moral subjecthood, responsibility gap, AI ethics, meaningful human control, interpretability.

Spis treści

Rozdział 1. Wprowadzenie: pytanie badawcze, aparat pojęciowy, hipotezy i ograniczenia badania.....	20
1.1. Kontekst filozoficzny, technologiczny i społeczny problemu	20
1.2. Problem badawczy (główne pytanie badawcze)	24
1.3. Hipotezy H1–H3 oraz cele szczegółowe.....	29
1.4. Zadania badawcze i wkład autorski	30
1.5. Metody i źródła (analiza pojęciowa, analiza i ocena porównawcza).....	33
A. Metody badawcze	33
B. Źródła i uzasadnianie ustaleń.....	36
1.6. Układ dysertacji i logika wyvodu.....	37
1.7. Zakres i ograniczenia badania oraz definicje kluczowych pojęć	39
Rozdział 2. Podmiotowość w wybranych klasycznych i współczesnych koncepcjach filozoficznych	43
Wprowadzenie.....	43
2.1. Podmiot myślenia, poznania i działania.....	44
2.2. Podmiot moralny w koncepcjach klasycznych i współczesnych.....	48
2.3. Spory o świadomość, wolną wolę i odpowiedzialność	51
2.4. Stanowiska krytyczne i sceptycyzm ontologiczny.....	58
2.5. Implikacje dla sztucznej podmiotowości	62
Podsumowanie	67
Rozdział 3. Kryteria operacyjno-diagnostyczne sztucznej podmiotowości moralnej	70
Wprowadzenie.....	70
3.1. Typologia agentów etycznych (James Moor)	71
3.2. Sprawczość, intencjonalność i celowość: racje, plany i kryterium K2	76
3.3. Inteligencja, refleksja i przewidywanie skutków jako kryterium K3	81
3.4. Autonomia decyzyjna i autokorekta norm jako kryterium K1 dla H1	86
3.5. Stany afektywne, stabilizacja diachroniczna i wiarygodność praktyk przypisywania odpowiedzialności (H2/K4).....	91

3.6. Metakontrola i silna rozliczalność (H3).....	97
Podsumowanie	103
Rozdział 4. Pluralistyczne ramy etyczne: deontologia, cnoty, konsekwencjalizm – porównanie i implikacje normatywne.....	105
Wprowadzenie.....	105
4.1. Pluralizm etyczny: etyka cnót, deontologia, konsekwencjalizm, modele hybrydowe.....	107
4.1.1. Deontologia.....	108
4.1.2. Etyka cnót.....	108
4.1.3. Konsekwencjalizm	109
4.1.4. LexOrder.....	110
4.2. Kryteria oceny działań przypisywanych systemom AI (PH1–PH3).....	115
4.3. Strategie projektowania sztucznej moralności – podejścia odgórne (ang. top- down), oddolne (ang. bottom-up) i hybrydowe (ang. hybrid).....	120
4.4. Metody weryfikacji spełnienia kryteriów PH1–PH3 w systemach AI (ex ante / in-process / ex post)	125
Podsumowanie	130
Rozdział 5. Sztuczna podmiotowość: klasyfikacje i realizacje techniczne	132
Wprowadzenie.....	132
5.1. Agent a podmiot w filozofii technologii oraz w prawie	133
5.2. Klasyfikacje sztucznej podmiotowości (ontologiczne i funkcjonalne).....	141
5.3. Przykłady implementacji sztucznej moralności.....	152
5.4. [ILUSTRACJA] IDx-DR/LumineticsCore jako empirycznie zakotwiczony przypadek graniczny modelu H1–H3 + WI	165
5.4.1. Status przypadku i poziomy analizy.....	165
5.4.2. Ocena minimalnej adresowalności proceduralnej H1	167
5.4.3. H2/K4 oraz H3: stabilność i kontrola.....	168
5.4.4. WI, MHC, MLI i klasyfikacja $A \times F + (R, I)$	170
5.4.5. Kontrargument i mini-konkluzja.....	172
Podsumowanie	173

Rozdział 6. Scenariusze rozwoju AI, luka odpowiedzialności i modele współodpowiedzialności	175
Wprowadzenie.....	175
6.1. Emulacja mózgu.....	176
6.2. Model ewolucyjny.....	185
6.3. Model rewolucyjny – osobliwość (eksplozja inteligencji).....	194
6.4. „Luka odpowiedzialności” – analiza i jakościowe wymiary diagnostyczne M1– M5	205
6.4.1. Rozwój pojęcia „luki odpowiedzialności”	205
6.4.2. LSP i ZSP: luka jako naruszenie warunków przypisywania odpowiedzialności	209
6.4.3. Metryki M1–M5 jako wymiary luki odpowiedzialności	211
6.4.4. Rodzaje luk odpowiedzialności jako profile w przestrzeni M1–M5.....	213
6.4.5. Ograniczenia i normatywna funkcja metryk M1–M5	215
6.5. Zagrożenia i wyzwania: technologiczne, społeczne i regulacyjne.....	217
6.6. Modele współodpowiedzialności i łańd regulacyjny	224
Podsumowanie	232
Rozdział 7. Synteza i weryfikacja H1–H3 oraz WI: hierarchia statusów moralnych AI, implikacje i rekomendacje.....	234
7.1. Omówienie wyników i weryfikacja H1–H3 oraz WI	234
7.1.1. Epistemiczne i instytucjonalne ograniczenia weryfikacji H1–H3 oraz WI...	245
7.2. Hierarchia statusu moralnego AI	248
7.2.1. Hierarchia statusów AI a analogie pozaludzkie.....	257
7.3. Wnioski końcowe oraz rekomendacje badawczo-regulacyjne	259
Słownik pojęć i oznaczeń	271
Bibliografia	280

Spis tabel

Tabela 1. Schemat macierzy $A \times F$	145
Tabela 2. Ocena IDx-DR/LumineticsCore według modelu rozprawy	171

Rozdział 1. Wprowadzenie: pytanie badawcze, aparat pojęciowy, hipotezy i ograniczenia badania

1.1. Kontekst filozoficzny, technologiczny i społeczny problemu

W tradycyjnych ujęciach filozoficznych moralność uznawano za cechę wyróżniającą człowieka. Arystoteles określał człowieka mianem „zwierzęcia politycznego”, zdolnego do życia etycznego we wspólnocie. Immanuel Kant twierdził z kolei, że tylko ludzie mogą działać z pobudek moralnych, a nie jedynie pod wpływem instynktu czy skłonności. Z czasem zaczęto jednak dostrzegać potrzebę poszerzenia zakresu odpowiedzialności moralnej. Proces ten, określany jako poszerzanie wspólnoty moralnej, rozumiem jako historyczno-normatywny rozwój polegający na obejmowaniu ochroną moralną coraz szerszych kategorii bytów, od wolnych mężczyzn w greckiej polis, przez niewolników, kobiety i dzieci, po zwierzęta, elementy środowiska naturalnego, a dziś także byty wytworzone przez człowieka.

Historia poszerzania wspólnoty moralnej nie przebiegała liniowo i nie prowadziła po prostu do coraz szerszego włączania kolejnych bytów. Była raczej zbiorem sporów, regresów i ponownych ustaleń granic. Emancypacja niewolników i kobiet wymagała nie tylko argumentów filozoficznych, lecz także presji społecznej i politycznej. Współczesne postulaty ochrony zwierząt oraz etyki środowiskowej również wywołały trwały spór, o czym świadczą filozoficzne krytyki koncepcji praw zwierząt i utylitaryzmu w odniesieniu do statusu zwierząt (Cohen, 1986; Carruthers, 1992; Frey, 1980), analizy podważające tezy biocentryzmu Paula Taylora (Evans, 2012), a także obrony antropocentrycznej polityki środowiskowej (Baxter, 1974; Passmore, 1980). Debata nad włączeniem artefaktów technologicznych do wspólnoty moralnej nie stanowi zatem końcowego etapu jednolitego procesu, lecz kolejną fazę historycznie zmiennego wyznaczania granic odpowiedzialności. Analogia ma tu charakter warunkowy. Jej wspólnym punktem jest normatywnie istotna sprawczość społeczna, to znaczy przewidywalny wpływ na dobra chronione normami. Taki wpływ rodzi roszczenia normatywne i uzasadnia przypisywanie odpowiedzialności. Zarazem analogia ta ma wyraźne granice. Przede wszystkim brak świadomości fenomenalnej oraz brak dowodów na autoteleologię, rozumianą jako samodzielne formułowanie i korygowanie celów wyższego rzędu, podważają możliwość prostego przeniesienia historycznych rozszerzeń na artefakty.

Obecnie formułuje się postulat warunkowego włączenia wybranych systemów sztucznej inteligencji (AI) do wspólnoty moralnej, jeżeli spełniają intersubiektywnie sprawdzalne kryteria sprawczości, rozumności i kompetencji moralnej (Coeckelbergh, 2020). Przesunięcie uwagi z ontologii na funkcję trafnie oddaje koncepcja „poziomów abstrakcji” (ang. *Levels of Abstraction*, LoA) Luciana Floridiego i Jeffa Sandersa (2004). Zamiast pytać, czym jest dany byt, pytam, jakie relacje wytwarza i jakie skutki wywołuje na określonym poziomie interakcji. W tym ujęciu system AI może aspirować do statusu podmiotu moralnego nie ze względu na swoją wewnętrzną istotę, lecz dlatego, że w praktyce działa jak podmiot podlegający ocenie moralnej. Ujęcie relacyjne osłabia znaczenie klasycznych trudności ontologicznych. Związek tej perspektywy z praktyką regulacyjną sygnalizuję tu jedynie skrótowo. Szersze omówienie wraz z analizą „luki odpowiedzialności” (ang. *responsibility gap*) przedstawiam w dalszej części pracy.¹

Skoro poszerzanie wspólnoty moralnej ujawnia względność dotychczasowych granic odpowiedzialności, należy przyjrzeć się klasycznym kryteriom podmiotowości, które wyznaczają normatywny horyzont dzisiejszych debat. Tradycyjnie podmiotowość moralną wiązano z wolną wolą, rozumnością, świadomością oraz zdolnością odróżniania dobra od zła (Kant, 2002; Strawson, 2008, s. 1–28). Osobą jest byt, który dzięki rozumowi praktycznemu ustanawia dla siebie prawo moralne; system AI, pozbawiony wolnej woli i refleksji nad sensem normatywnym, uchodzi zatem za narzędzie. W nurcie etyki utilitarystycznej, reprezentowanym między innymi przez Jeremy’ego Benthama i Petera Singera, kryterium podmiotowości moralnej stanowi zdolność do odczuwania cierpienia i przyjemności (ang. *sentience*). W niniejszej pracy odróżniam ją od funkcjonalnych wskaźników quasi-doznaniowych. Traktuję je wyłącznie jako techniczne odpowiedniki użyteczne w ocenie regulacyjnej, a nie jako wystarczające kryterium podmiotowości *sensu stricto*. U Arystotelesa punkt ciężkości spoczywa na mądrości praktycznej (gr. *phronesis*) i wyborze „ze względu na dobro” (Arystoteles, 2012). Przez „teleologię praktyczną” rozumiem tu rekonstrukcję arystotelesowskiej mądrości praktycznej w kategoriach operacyjnych, czyli trwałą zdolność adekwatnego rozpoznawania dobra w sytuacjach konfliktowych, która prowadzi do wyboru działań ze względu na dobro, a nie

¹ Przez sprawczość rozumiem trwałą zdolność realizowania celów mimo zakłóceń oraz gotowość do ich korygowania zgodnie z regułami systemu (testy: odporność zadaniowa, utrzymywanie celu, możliwość wyjaśnienia reguły sterującej). Rozumność praktyczna to zdolność formułowania uzasadnień opartych na regułach i konsekwentnego stosowania tych reguł w podobnych sytuacjach (testy: spójność decyzji, trafność rozstrzygnięć w konfliktach reguł). Kompetencja moralna oznacza rozpoznawanie konfliktów wartości oraz utrzymywanie stabilnego priorytetu zgodnego z przyjętą zasadą (testy: powtarzalność rozstrzygnięć zgodnych z deklarowaną zasadą; audyt uzasadnień *ex post*).

wyłącznie na wynik. Ujęcie to ma charakter przybliżony i nie przesądza o istnieniu odpowiednika w postaci doznań fenomenalnych; operuje pojęciem intencjonalności funkcjonalnej, rozumianej jako celowe ukierunkowanie systemu na określone wartości, bez uznawania świadomości fenomenalnej za warunek konieczny. Tak rozumiana intencjonalność rodzi dwa pytania. Po pierwsze, czy systemy AI mogą przejawiać te cechy mimo braku wewnętrznych przeżyć? Po drugie, czy sama symulacja oraz orientacja moralna wystarczają do przypisania systemom AI statusu podmiotu moralnego? Szczegółowe rozstrzygnięcie przedstawiam niżej, w którym odróżniam strategie interpretacyjne od intencjonalności ontologicznej. Przywołane kryteria nie tworzą jednolitego zbioru, lecz odrębne uzasadnienia podmiotowości. Sprowadzenie ich do wspólnego mianownika grozi eklektyzmem, który zaciera różnice między deontologią, konsekwencjalizmem i etyką cnót; dlatego dalej analizuję je oddzielnie, wskazując ich konsekwencje dla statusu moralnego AI.

Spór koncentruje się wokół pytania, czy systemy AI jedynie symulują zachowania podmiotów moralnych, czy też, przy spełnieniu intersubiektywnie weryfikowalnych kryteriów, mogą pełnić ich funkcję. Sceptycy (Johnson, 2006; Torrance, 2008; Himma, 2009; Hew, 2014) podkreślają brak podstaw do przypisywania systemom AI odpowiedzialności moralnej *sensu stricto*. Odpowiedzialność wymaga działania intencjonalnego, to znaczy podejmowania decyzji na podstawie rozpoznania i stosowania zasad etycznych, a nie na podstawie reakcji na dane wejściowe ani wykonywania instrukcji. Typologia Jamesa Moora (2006), obejmująca agentów o wpływie etycznym, implicytnych, jawnych i pełnych, wskazuje raczej na kontinuum niż na prostą dychotomię podmiotowości. Przyjmuję ją jako ramę operacjonalizacji. Zamiast opozycji „podmiot/nie-podmiot” posługuję się gradacją, która pozwala określać zakres i tryb przypisywania odpowiedzialności odpowiednio do wykazywanych zdolności. Taka stopniowalność pozwala przypisywać odpowiedzialność systemom spełniającym jedynie część klasycznych kryteriów; rama ta będzie kluczowa przy formułowaniu warunków minimalnych (por. Moor, 2006; zob. rozdz. 3 niniejszej dysertacji).²

Rzetelna ocena sceptycyzmu i funkcjonalizmu wymaga wskazania zakresu ich mocy wyjaśniającej. Sceptycy akcentują między innymi brak świadomości fenomenalnej

² Rozróżniam odpowiedzialność regulacyjną, przypisywaną rola i funkcjom w systemach socjotechnicznych, od odpowiedzialności moralnej *sensu stricto*, która wymaga intencjonalności i rozumności praktycznej. W dalszej części pracy termin „quasi-podmiot” odnoszę do pierwszej kategorii, natomiast „podmiot moralny” do drugiej.

w AI, co, ich zdaniem, stanowi zasadniczą granicę odróżniającą maszyny od istot moralnych i chroni przed pochopnym rozszerzaniem pojęcia podmiotowości. Przyjęcie tego warunku istotnie utrudnia jednak projektowanie schematów przypisywania odpowiedzialności wobec systemów o znaczącym wpływie społecznym, nawet jeśli niektóre reżimy prawne, na przykład odpowiedzialność zastępcza, pozostają dostępne. Rozwój głębokiego uczenia, przetwarzania języka naturalnego i systemów adaptacyjnych doprowadził do powstania artefaktów podejmujących decyzje na podstawie danych napływających w czasie rzeczywistym. Pojazdy autonomiczne, platformy rekomendacyjne i duże modele językowe, takie jak GPT-5, współtworzą dziś infrastrukturę o znaczącym, lecz trudnym do przewidzenia oddziaływaniu na sferę publiczną. Nieprzejrzystość procesu uczenia uniemożliwia pełną kontrolę *ex ante* oraz pełne wyjaśnienie *ex post*, co rodzi „lukę odpowiedzialności”, rozumianą tutaj jako brak podmiotu spełniającego łącznie przesłanki kontroli, przewidywalności i rozumności. Problem ten omawiam szerzej w dalszej części pracy (por. rozdz. 6).

Niezależnie od sceptycyzmu badania nad interakcją człowiek–komputer dokumentują spontaniczną personifikację artefaktów oraz skłonność do przypisywania niektórym systemom AI intencji i emocji (Turkle, 2011). Wymownym przykładem jest humanoidalna Sophia, której Arabia Saudyjska nadała symboliczne „obywatelstwo” w 2017 r. Był to gest wizerunkowy, pozbawiony skutków prawnych, który ujawnił siłę uznania społecznego, lecz nie ustanowił statusu normatywnego (Wróblewski & Fortuna, 2023). Uznanie społeczne, obejmujące oczekiwania, role i legitymizacje, odróżniam zatem od statusu normatywnego, który wymaga wykazania responsywności na racje oraz stosowania reguł w nowych przypadkach (Berger & Luckmann, 1966; Honneth, 1992). Przyjmowanie wobec artefaktów postawy intencjonalnej pozostaje przede wszystkim strategią przewidywania, a nie dowodem intencjonalności ontologicznej (Dennett, 1987). W tym kontekście Joanna J. Bryson ujmuje społeczną rolę systemów AI jako przedmiot decyzji etycznych i projektowych, a nie status wynikający automatycznie z ich możliwości technicznych (Bryson, 2018, s. 16, 21). Z perspektywy przyjętej w rozprawie samo przyznanie artefaktowi roli instytucjonalnej nie dowodzi jednak spełnienia warunków podmiotowości moralnej.

Na tym tle wyłania się spektrum stanowisk. Technoentuzjaści, tacy jak Ray Kurzweil oraz autorzy scenariuszy superinteligencji, tacy jak Nick Bostrom, przewidują możliwość powstania skrajnie potężnych systemów. Kwestia ich moralnej przewagi nad człowiekiem pozostaje jednak w tych ujęciach co najmniej sporna. Po stronie sceptycznej

występują różne argumenty: Deborah Johnson kwestionuje niezależną moralną sprawczość systemów komputerowych ze względu na brak odpowiednich stanów intencjonalnych i wolności działania; Steve Torrance i Kenneth Himma wiążą pełną sprawczość moralną ze świadomością lub zdolnością do odczuwania cierpienia i przyjemności; Patrick Hew akcentuje zależność reguł działania od zewnętrznych twórców; Joanna J. Bryson ostrzega natomiast przed automatycznym przekształcaniem możliwości technicznych w status normatywny (Johnson, 2006, s. 198–200; Torrance, 2008, s. 499–503; Himma, 2009, s. 24–28; Hew, 2014, s. 197; Bryson, 2018, s. 16, 21). Wspólna jest im zatem słabsza teza: wzrost sprawności technicznej nie stanowi sam przez się wystarczającej podstawy do przypisania podmiotowości moralnej.

Między tymi skrajnościami rozwijają się koncepcje warunkowej lub stopniowalnej podmiotowości, hierarchizujące statusy moralne (Gunkel, 2018) bądź odwołujące się do funkcjonalnych ról systemów AI. Luciano Floridi i Jeff Sanders (2004) proponują traktować niektóre systemy jako quasi-podmioty, jeżeli spełniają minimalne warunki agentowości i realizują funkcje o znaczeniu normatywnym. W dalszej analizie rekonstruuję i porównuję te stanowiska, odsyłając do operacjonalizacji opartej na typologii Moora oraz wskazanych wyżej warunkach minimalnych. Punktem wyjścia będzie wskazanie cech, które system AI musi wykazywać, aby można było rozważyć przypisanie mu statusu moralnego agenta. Prowadzi to do zasadniczego pytania badawczego: jakie właściwości są minimalnie konieczne, aby uznać system AI za moralnie odpowiedzialny podmiot i czy można je intersubiektywnie zidentyfikować w systemach istniejących lub projektowanych?

1.2. Problem badawczy (główne pytanie badawcze)

Przekształcanie AI z narzędzia obliczeniowego w systemy generatywno-adaptacyjne przesuwają pojęcie sprawczości w stronę autonomii instrumentalnej oraz ujawnia napięcia między nieprzejrzystością procesu decyzyjnego a tradycyjnymi modelami odpowiedzialności opartymi na racjonalnej i intencjonalnej sprawczości. Stawką pozostaje wskazanie podmiotu odpowiedzialności moralnej oraz określenie podziału odpowiedzialności między projektantami, operatorami a systemem. Wobec stopniowalności sprawczości oraz niepewności co do progu minimalnego proponuję porządek analityczny oparty na trzech powiązanych wymiarach: poznawczym, afektywnym i sprawczym.

Pytanie badawcze: jak ocenić, czy system AI osiąga poziom uzasadniający przypisanie mu odpowiedzialności moralnej oraz jakie cechy poznawcze, afektywne i sprawcze warunkują uznanie go za uczestnika relacji etycznoprawnych?

Pytanie to należy rozumieć w sposób prospektywny i progowy. Nie zakładam, że obecne systemy AI spełniają już warunki odpowiedzialności moralnej, podmiotowości moralnej ani silnej rozliczalności. Badam raczej, jakie warunki musiałby spełnić system przyszły, hipotetyczny albo bardziej zaawansowany, aby można było rozważać jego ograniczony udział w praktykach uzasadniania, audytu, korekty i rozliczania. W tym sensie praca nie dowodzi odpowiedzialności moralnej obecnych systemów AI, lecz rekonstruuje warunki możliwej odpowiedzialności proceduralnej i rozliczalności.

Teza robocza: minimalny, intersubiektywnie identyfikowalny próg adresowalności proceduralnej AI wyznacza H1, rozumiana jako funkcjonalna zdolność rozpoznawania norm, przewidywania konsekwencji ich naruszeń oraz powiązania rozstrzygnięcia z wystarczającymi i wiarygodnymi podstawami decyzji, które umożliwiają audyt, zakwestionowanie i korektę wyniku.³ H2 pełni funkcję stabilizującą, wzmacniając trwałość wyborów zgodnych z racjami, lecz nie stanowi warunku koniecznego dla progu minimalnego. H3 oznacza wyższy próg silnej rozliczalności, wymagający kontroli metapoznawczej oraz jawnego rozstrzygania konfliktów racji.

W systemach wykorzystujących uczenie maszynowe nieprzejrzystość może wynikać z tajemnicy organizacyjnej, ograniczonych kompetencji technicznych odbiorców lub złożoności obliczeń, która utrudnia ich ujęcie w kategoriach zrozumiałych dla człowieka (Burrell, 2016, s. 3–5). W przyjętej tu analizie trudności te ograniczają możliwość uzasadniania decyzji i oceny warunków odpowiedzialności; nie dowodzą jednak nieprzewidywalności każdego wyniku ani konieczności uznania systemu za podmiot moralny. W literaturze wskazuje się, że współczesne systemy spełniają minimalne warunki agentowości funkcjonalnej, tj. interaktywność, autonomię i adaptacyjność, ujmowane jako ciągły wymiar właściwości, co uzasadnia ich ocenę etyczną, lecz nie rozstrzyga o ich statusie podmiotowym (Floridi & Sanders, 2004; Moor,

³ Przez „adresowalność proceduralną” rozumiem możliwość traktowania danego systemu jako punktu odniesienia w procedurach oceny, uzasadniania, audytu oraz korekty decyzji. Termin nawiązuje do ang. *addressability*, lecz w niniejszej pracy ma znaczenie techniczne i jest węższe niż ogólna „odpowiedzialność” lub „rozliczalność”. System jest proceduralnie adresowalny wtedy, gdy można skierować wobec niego pytanie o racje działania, odtworzyć lub ocenić przedstawione uzasadnienie oraz powiązać wynik tej oceny z możliwością jego korekty. Pojęcie to nie zakłada świadomości fenomenalnej, osobowości moralnej ani samodzielnej odpowiedzialności prawnej systemu; oznacza jedynie próg uczestnictwa w publicznych praktykach uzasadniania i kontroli.

2006; Bryson, 2018). Rozróżniam sprawczość teleologiczną, czyli własność stopniowalną rozumianą jako zdolność samodzielnego wyznaczania i ponownego hierarchicznego porządkowania celów, agentowość funkcjonalną, czyli ciągły opis właściwości systemu mierzony interaktywnością, autonomią i adaptacyjnością oraz podmiotowość moralną, czyli status normatywny nadawany po przekroczeniu progu minimalnego. Dalej przez „sprawczość” rozumiem sprawczość teleologiczną; odróżniam ją od sprawczości kauzalnej, czyli czystego wpływu przyczynowego, który sam przez się nie ma znaczenia normatywnego. Zarazem brak spójnych, powszechnie akceptowanych wskaźników oceny rozwoju AI, odnoszących się do minimalnego progu podmiotowości moralnej, utrudnia stanowcze rozstrzygnięcia oraz praktykę regulacyjną (Winfield i in., 2014; Jobin, Ienca & Vayena, 2019).

Przyjmuję interdyscyplinarne ramy łączące główne nurty etyki, tj. deontologię, konsekwencjalizm i etykę cnót, z etyką informacyjną oraz prakseologią inżynierską. W tym ujęciu deontologia wiąże autonomię ze stanowieniem norm i samostwierdzeniem, konsekwencjalizm akcentuje przewidywanie skutków i minimalizację szkód, a etyka informacyjna rozszerza kategorie moralne na artefakty cyfrowe, jeżeli spełniają one określone warunki agentowości. Ponieważ sprawczość maszyn jest stopniowalna, postuluję normatywno-decyzyjny „próg jasnej granicy” (ang. *bright line threshold*), tj. umowną, lecz epistemicznie i prawnie uzasadnioną granicę, powyżej której wzrost właściwości agentowych pociąga za sobą określone skutki etyczne i prawne. Próg ten rozumiem jako regułę typu „jasna granica”, znaną z teorii prawa; ma on charakter decyzji normatywnej wspartej metrykami funkcjonalnymi, co umożliwia jego operacjonalizację bez negocjowania ciągłości sprawczości.

Wymiar poznawczy obejmuje minimalną zdolność rozpoznawania norm, przewidywania konsekwencji ich naruszeń oraz kontrfaktycznego porównywania alternatyw, bez wymogu deklaratywnej samoświadomości. Pełna kontrola metapoznawcza nie jest wymagana dla H1, lecz stanowi treść H3 jako warunek silnej rozliczalności. Z perspektywy postawy intencjonalnej intencjonalność przypisywana jest wtedy, gdy zachowanie agenta najlepiej wyjaśnia model przekonań i celów (Dennett, 1987). W niniejszym ujęciu „postawa intencjonalna” oznacza strategię interpretacyjną, natomiast „intencjonalność funkcjonalna” oznacza własność systemu intersubiektywnie identyfikowalną w zachowaniu oraz w audytowalnych protokołach decyzyjnych. Na progu minimalnym dla AI warunkiem koniecznym jest funkcjonalna kompetencja rozpoznawania norm i przewidywania konsekwencji ich naruszenia, nawet bez

deklaratywnej samoświadomości (Himma, 2009). Taki moduł „rozumienia norm” musi współdziałać z mechanizmami wnioskowania kontrfaktycznego oraz z możliwością wglądu w proces decyzyjny i jego audytu, aby umożliwiać ocenę alternatyw oraz korektę polityk działania. W wersji minimalnej weryfikacja obejmuje rozpoznawanie kolizji norm, przewidywanie sankcji w scenariuszach kontrfaktycznych oraz aktualizację polityk po naruszeniu, czyli uczenie oparte na błędach. Bez tej kompetencji nie sposób uzasadnić praktyki przypisywania nawet ograniczonej odpowiedzialności.

Wymiar afektywny definiuję jako zdolność do wartościowania zjawisk. Funkcjonalnie steruje on zachowaniem systemów – skłania do poszukiwania nagród i unikania zagrożeń (Picard, 1995; Winfield, 2024). Nie chodzi o *qualia*, lecz o afektywność funkcjonalną, czyli wzorce preferencji sprzężone z oceną sytuacji oraz regulacją reakcji. Kluczowe pozostaje rozróżnienie między afektywnością, rozumianą jako hedoniczne wartościowanie bodźców, a emocjami, ujmowanymi jako epizodyczne stany afektywne modulujące percepcję i działanie. Dalej używam terminu „afektywność funkcjonalna” jako kategorii zbiorczej; natomiast „emocje” określam jako epizodyczne przejawy tej afektywności. H2 nie jest warunkiem koniecznym do przekroczenia progu H1. Stabilny układ preferencji wartościujących wzmacnia jednak trwałość i przewidywalność wyborów zgodnych z normami, dlatego H2 pełni funkcję stabilizującą w modelu odpowiedzialności proceduralnej i zwiększa wiarygodność przypisywania odpowiedzialności w przypadkach granicznych.

Wymiar sprawczy oznacza stopniowalną zdolność systemu do realizowania, aktualizowania i porządkowania celów w zmiennych warunkach; rozumiem ją nie metafizycznie, jako wolną wolę, lecz funkcjonalnie, jako wewnętrzną procedurę wartościującą, która aktualizuje priorytety niezależnie od bieżących instrukcji operatora (Taddeo & Floridi, 2021). Rozróżniam dwie odmiany autonomii funkcjonalnej: (i) instrumentalną, czyli wykonywanie przydzielonych celów w zmiennym środowisku bez modyfikacji hierarchii celów; (ii) adaptacyjną, czyli ponowne samodzielne porządkowanie celów hierarchicznie, tj. aktualizację polityk działania przy ograniczonej zależności od operatora. Dla H1 istotne jest to, czy system może być adresatem oceny proceduralnej ze względu na rozpoznawanie norm i przewidywanie skutków działania. Dalej idące samosterowne porządkowanie celów i jawne rozstrzygnięcie konfliktów racji należą do H3 i uzasadniają silną rozliczalność.

Z zarysowanego kontekstu wynikają dwa główne zadania. Po pierwsze, konceptualna i techniczna operacjonalizacja minimalnych warunków podmiotowości

moralnej. Po drugie, opracowanie intersubiektywnie kontrolowalnych kryteriów oceny spełnienia tych warunków przez istniejące i projektowane systemy. Dotychczasowe propozycje, od „moralnego testu Turinga” po audyty funkcjonalne, pozostają fragmentaryczne i trudne do weryfikacji, co utrudnia kumulację wiedzy oraz wyznaczanie progów decyzyjnych (Winfield i in., 2014; Jobin, Ienca & Vayena, 2019). Poniższe hipotezy odpowiadają trzem lukom: brakowi pojęciowo sprawdzalnej intencjonalności funkcjonalnej (H1), niedookreśleniu roli afektywności (H2) oraz nadmiernemu akcentowaniu świadomości fenomenalnej kosztem hierarchicznego porządkowania celów w sposób funkcjonalny (H3). Formułuję trzy hipotezy robocze, które zostaną poddane rekonstrukcji i sprawdzeniu koherencyjnemu w dalszych rozdziałach:

H1: system AI może zostać uznany za minimalnie adresowalny proceduralnie i objęty słabą odpowiedzialnością proceduralną wtedy, gdy wykazuje funkcjonalną zdolność rozpoznawania norm, przewidywania konsekwencji ich naruszeń, reagowania na racje w trybie kontrfaktycznym oraz udostępniania wystarczających i wiarygodnych podstaw decyzji, umożliwiających ocenę ich związku z wynikiem, audyt, zakwestionowanie oraz korektę. H1 nie zakłada autonomii normatywnej w sensie silnym ani pełnej refleksyjności; wyznacza minimalny próg, od którego ocena systemu nie sprowadza się do opisu czysto kauzalnego.

H2: minimalna afektywność funkcjonalna nie stanowi warunku koniecznego do przekroczenia progu H1. Pełni funkcję stabilizującą i wzmacniającą, ponieważ zwiększa trwałość wyborów zgodnych z racjami, porządkuje przestrzeń decyzyjną w sytuacjach niepewności oraz ogranicza rozmycie celów. H2 może wzmacniać wiarygodność przypisywania odpowiedzialności, ale nie definiuje minimalnej adresowalności proceduralnej.

H3: silna rozliczalność systemu AI wymaga kontroli metapoznawczej, czyli zdolności monitorowania własnego procesu decyzyjnego, jawnego rozstrzygania konfliktów racji oraz przedstawienia audytowalnego uzasadnienia decyzji z perspektywy drugiego rzędu. H3 nie wymaga świadomości fenomenalnej; wymaga natomiast funkcjonalnych mechanizmów metakontroli, które pozwalają odróżnić stabilne reagowanie na racje od jedynie powierzchownej zgodności z regułą.

WI: warunek instytucjonalnego zastosowania modelu oznacza istnienie publicznie rozpoznawalnych procedur, ról, mechanizmów audytu oraz zasad rozliczalności, w których można w praktyce zastosować ustalenia H1–H3. WI nie jest dodatkowym warunkiem agentowym. Określa, czy i w jakim zakresie spełnienie H1–H3 może zostać przełożone na praktyki społeczne, prawne lub regulacyjne.

Przyjęty model H1–H3 + WI ma charakter przyszłościowy i progowy. H1–H3 nie opisują warunków faktycznie spełnianych przez obecne systemy AI, lecz wyznaczają progi, które musiałby spełnić system przyszły, hipotetyczny albo bardziej zaawansowany, aby można było mówić o jego ograniczonej odpowiedzialności proceduralnej i rozliczalności. WI nie dodaje kolejnej hipotezy dotyczącej sprawczości systemu, lecz określa warunki instytucjonalnego zastosowania modelu: publiczną rozpoznawalność procedur, możliwość audytu, korekty i odwołania oraz ostateczne przypisanie odpowiedzialności ludziom i instytucjom.

1.3. Hipotezy H1–H3 oraz cele szczegółowe

Oznaczenia H1–H3 są w niniejszej dysertacji zarezerwowane dla hipotez roboczych; poziomy pluralizmu hierarchicznego w rozdziale 4 oznaczam odrębnie jako PH1–PH3.

Minimalne i wzmocnione warunki przypisywania odpowiedzialności AI porządkuję za pomocą trzech hipotez: H1 określa próg minimalnej adresowalności proceduralnej i słabej odpowiedzialności proceduralnej; H2 określa funkcję stabilizującą i wzmacniającą, która nie jest warunkiem koniecznym dla H1; H3 określa próg silnej rozliczalności oparty na kontroli metapoznawczej i jawnym rozstrzygnięciu konfliktów racji.

Wniosek syntetyczny. H1 wyznacza minimalny próg adresowalności proceduralnej oraz słabej odpowiedzialności proceduralnej. H2 wzmacnia stabilność realizacji racji, ale nie jest warunkiem koniecznym dla H1. H3 wyznacza próg silnej rozliczalności poprzez kontrolę metapoznawczą oraz jawne rozstrzygnięcie konfliktów racji. WI określa warunki instytucjonalnego zastosowania modelu, czyli możliwość osadzenia H1–H3 w procedurach audytu, kontroli oraz rozliczalności.

W dalszej części pracy termin „racja” zachowuje przede wszystkim sens filozoficzny. Odnosi się do rekonstrukcji działania, odpowiedzialności oraz udzielania uzasadnień w odniesieniu do podmiotów ludzkich, zwłaszcza w rozdziale 2. W

odniesieniu do systemów AI używam tego terminu ostrożnie. Tam, gdzie groziłaby antropomorfizacja, stosuję określenia: „przesłanki normatywno-decyzyjne”, „podstawa decyzji”, „reguła pierwszeństwa” oraz „kolizja przesłanek normatywno-decyzyjnych”. Systemu AI nie traktuję więc jako bytu rozumiejącego racje w ludzkim sensie, lecz jako system, którego wynik może być, w określonych warunkach, zależny od jawnych i poddających się audytowi przesłanek normatywno-decyzyjnych.

W kolejnych rozdziałach H1–H3 pełnią funkcję ramy oceny modeli podmiotowości i odpowiedzialności AI. Rozdział 3 doprecyzowuje aparat pojęciowy i progi odpowiedzialności; rozdział 4 analizuje ich znaczenie normatywne; rozdział 5 odnosi je do klasyfikacji systemów AI, ich ograniczeń oraz hipotetycznego przypadku sprawdzającego; rozdział 6 wykorzystuje je do analizy luki odpowiedzialności, wymiarów diagnostycznych oraz modelu współodpowiedzialności.

1.4. Zadania badawcze i wkład autorski

Teza robocza niniejszego podrozdziału głosi, że model przypisywania odpowiedzialności moralnej systemom AI powinien zostać uporządkowany nie poprzez dodanie kolejnego poziomu agenta etycznego, lecz poprzez rozróżnienie trzech funkcji progowych: minimalnej adresowalności proceduralnej, stabilizacji motywacyjno-wartościującej oraz silnej rozliczalności. Pozwala to uniknąć dublowania kategorii refleksyjności i zachować ciągłość z typologią agentów etycznych Moora, traktowaną jako tło porządkujące, a nie jako pełną strukturę własnego modelu.

Pierwszą funkcję pełni H1. Oznacza ona próg minimalnej adresowalności proceduralnej systemu jako punkt odniesienia w procedurach oceny, uzasadniania i korekty oraz jako podstawę słabej odpowiedzialności proceduralnej. System spełniający H1 nie musi być pełnoprawnym podmiotem moralnym w sensie klasycznym. Musi jednak wykazywać funkcjonalną zdolność rozpoznawania norm, przewidywania konsekwencji ich naruszeń, reagowania na racje w trybie kontrfaktycznym oraz udostępniać wystarczające i wiarygodne podstawy decyzji, umożliwiające ocenę ich związku z wynikiem, audyt, zakwestionowanie oraz korektę. H1 wyznacza zatem minimalny próg, od którego przypisywanie odpowiedzialności uzyskuje podstawę proceduralną, choć nie oznacza jeszcze pełnej odpowiedzialności moralnej ani samodzielnej rozliczalności instytucjonalnej.

Drugą funkcję pełni H2. Minimalna afektywność funkcjonalna nie jest warunkiem koniecznym do przekroczenia progu H1. Pełni natomiast funkcję stabilizującą: wzmacnia

trwałość wyborów zgodnych z racjami, ogranicza przypadkowość reakcji i zwiększa odporność systemu na rozmycie celów w zmiennym środowisku. W tym sensie H2 nie definiuje progu minimalnego, ale podnosi wiarygodność przypisywania odpowiedzialności w przypadkach granicznych.

Trzecią funkcję pełni H3. Oznacza ona warunek silnej rozliczalności, obejmujący kontrolę metapoznawczą, jawne rozstrzyganie konfliktów racji oraz możliwość przedstawienia uzasadnienia decyzji z perspektywy drugiego rzędu. H3 przejmuje funkcję refleksyjności/metakontroli, która nie jest już ujmowana jako osobny poziom agenta, lecz jako wzmocniony warunek odpowiedzialności. System spełniający H3 powinien móc wskazać, dlaczego dana racja przeważała, jak rozpoznał konflikt między celem lokalnym a zasadą nadrzędną oraz w jaki sposób skorygował decyzję w świetle nowych racji.

Odrębne miejsce zajmuje WI, czyli warunek instytucjonalnego zastosowania modelu. H1–H3 opisują warunki agentowe i funkcjonalne, natomiast WI określa, czy istnieją publicznie rozpoznawalne role, procedury kontroli, mechanizmy audytu oraz zasady rozliczalności umożliwiające zastosowanie modelu w praktykach społecznych i prawnych. WI nie rozszerza treści H1–H3, lecz określa warunki ich praktycznego zastosowania.

Wkład autorski pracy polega zatem na uporządkowaniu modelu odpowiedzialności AI według progów H1–H3 oraz warunku instytucjonalnego WI. Model ten pozwala odróżnić minimalną adresowalność proceduralną od silnej rozliczalności, a także oddzielić warunki agentowe od warunków instytucjonalnego zastosowania. W dalszych rozdziałach: rozdział 3 rozwija aparat pojęciowy H1–H3, rozdział 4 analizuje ich znaczenie normatywne, rozdział 5 odnosi je do klasyfikacji i realizacji technicznych systemów AI, natomiast rozdział 6 wykorzystuje je do analizy luki odpowiedzialności, metryk oraz modelu współodpowiedzialności człowiek–AI.

Wkład ten należy rozumieć jako propozycję filozoficzno-normatywną, a nie jako twierdzenie, że istniejące systemy AI spełniają H1–H3. Model wskazuje, jakie warunki musiałyby zostać spełnione, aby przypisanie ograniczonej odpowiedzialności proceduralnej nie miało charakteru czysto metaforycznego ani nie sprowadzało się do odpowiedzialności ludzi za narzędzie. W tym sensie praca zachowuje minimalizm ontologiczny: nie przypisuje systemom AI świadomości fenomenalnej, moralnej samowiedzy ani ludzkiego rozumienia racji, lecz bada warunki ich możliwej adresowalności proceduralnej w praktykach instytucjonalnych.

Dalszy wkład autorski polega na doprecyzowaniu progów H1–H3 za pomocą kryteriów operacyjno-diagnostycznych K1–K4. Kryteria te porządkują analizę autonomii decyzyjnej i autokorekty norm, zależności decyzji od przesłanek normatywno-decyzyjnych, refleksyjności predykcyjnej oraz stabilności diachronicznej. Nie pełnią funkcji empirycznych testów systemów AI, lecz służą filozoficznej kontroli pojęć: pozwalają odróżnić rekonstruowalną zależność wyniku od podstawy decyzji, uzasadnienia i korekty od samej symulacji uzasadnień, skuteczności technicznej albo pozornej zgodności. Uzupełnia je miara mocy wyjaśniającej EPM, rozumiana jako narzędzie porządkowania eksperymentów myślowych i przypadków granicznych według spójności definicyjnej, odporności granicznej oraz potencjału normatywno-proceduralnego.

Istotnym elementem własnego ujęcia jest także pluralistyczna rama normatywna, w której deontologia, etyka cnót i konsekwencjalizm nie zostają połączone w jedną teorię moralności, lecz otrzymują odmienne funkcje w porządku uzasadniania. Wyrazem tego jest struktura LexOrder, która wyznacza kolejność oceny: najpierw dopuszczalność i adresowalność, następnie stabilność oraz wiarygodność praktyk, a dopiero potem rozstrzyganie kolizji przesłanek normatywno-decyzyjnych w świetle skutków, ryzyka i niepewności. Z tą strukturą powiązane są zasada FairAim, Zasada Stopniowej Weryfikowalności oraz Minimum Publicznego Uzasadnienia. Ich funkcją jest zabezpieczenie modelu przed utożsamieniem odpowiedzialności z deklaratywną zgodnością systemu z zasadami etycznymi.

Wkład pracy obejmuje ponadto rozróżnienie między opisową kategorią agenta a normatywno-instytucjonalną kategorią podmiotu. Rozróżnienie to zostaje rozwinięte w macierzy $A \times F + (R, I)$, która łączy klasyfikacje ontologiczne i funkcjonalne sztucznej podmiotowości z warunkami interwencji oraz interpretowalności. Dzięki temu quasi-podmiotowość AI zostaje ujęta jako ograniczony, proceduralny status zależny od publicznie kontrolowanych warunków instytucjonalnych. System może być punktem odniesienia dla roszczeń o uzasadnienie i korektę, lecz nie zastępuje ludzi ani instytucji jako końcowych adresatów odpowiedzialności.

Autorski charakter ma również ujęcie luki odpowiedzialności jako zjawiska stopniowalnego, a nie jako prostego braku jednego podmiotu odpowiedzialnego. Służą temu lokalna zasada przypisania odpowiedzialności, zasada superweniencji praktyk przypisywania odpowiedzialności oraz jakościowe wymiary diagnostyczne M1–M5. Wymiary te opisują identyfikowalność adresatów odpowiedzialności, dostępność i

audytowalność podstaw decyzji, kontrolę zależności decyzji od przesłanek normatywno-decyzyjnych, przejrzystość i trwałość praktyk przypisywania odpowiedzialności oraz relację między tempem działania lub zmian a czasem i wiedzą dostępnymi na reakcję, refleksję i korektę. Pozwala to ujmować różne typy luk odpowiedzialności jako odmienne układy deficytów w tej samej przestrzeni pojęciowej.

Do wkładu autorskiego należy także wykorzystanie scenariuszy emulacyjnego, ewolucyjnego i rewolucyjnego jako modeli analitycznych, a nie jako prognoz technologicznych. Scenariusze te służą pokazaniu, jak odmienne konfiguracje sztucznej podmiotowości generują różne problemy odpowiedzialności: rozgałęzienie tożsamości, narastającą złożoność układów socjotechnicznych oraz przekroczenie zdolności ludzkiego nadzoru przez tempo działania systemów. Na tym tle rozwijany jest model współodpowiedzialności człowiek–AI, którego celem jest utrzymanie rozliczalności ludzi i instytucji mimo rosnącej autonomii funkcjonalnej systemów sztucznej inteligencji.

Kończącym elementem wkładu jest dwutorowa hierarchia statusów moralnych AI oraz wynikające z niej rekomendacje badawczo-regulacyjne. Hierarchia ta rozdziela oś agencji, dotyczącą ról proceduralnych systemu, od osi pacjencji, dotyczącej ewentualnego zakresu ochrony. Pozwala to uniknąć utożsamienia rosnącej sprawczości AI z automatycznym przyznaniem jej praw, osobowości lub statusu pacjenta moralnego. Rekomendacje końcowe mają charakter filozoficzno-regulacyjny: wskazują na potrzebę znaczącej kontroli człowieka, minimalnej interpretowalności, jawnych procedur audytu, możliwości korekty oraz przeciwdziałania praniu odpowiedzialności. W tym sensie dysertacja proponuje model, który łączy minimalizm ontologiczny z wysokimi wymaganiami normatywnymi: nie antropomorfizuje systemów AI, ale też nie pozwala traktować ich oddziaływania jako moralnie neutralnego.

1.5. Metody i źródła (analiza pojęciowa, analiza i ocena porównawcza)

A. Metody badawcze

Trzon metodologiczny dysertacji stanowi analiza pojęciowa, ponieważ minimalne warunki podmiotowości moralnej sztucznej inteligencji mają charakter konceptualny. W tradycji filozofii analitycznej klarowne definicje są warunkiem rzetelnego rozstrzygania sporów normatywnych (Woleński, 1998). Sama precyzja terminologiczna jednak nie wystarcza; uzupełniam ją porównaniem różnych typów literatury i argumentów,

obejmującym (i) ustalenia psychologii moralnej przywoływane w literaturze, (ii) rekonstrukcje procedur deliberatywnych, (iii) analizę przypadków technologicznych (Winfield, Blum & Liu, 2014; Floridi & Cows, 2021). W odniesieniu do AI kluczowe pojęcia, takie jak „autonomia”, „intencjonalność” i „podmiotowość moralna”, są wieloznaczne; ich analiza semantyczna ujawnia ukryte założenia i zapobiega niejednoznacznościom pojęciowym w interdyscyplinarnej debacie (Floridi & Sanders, 2004).

W niniejszej pracy określenia takie jak „triangulacja argumentacyjna”, „sprawdzanie koherencyjne” czy „kontrola inferencyjna” mają znaczenie wyłącznie metodologiczne. Odnoszą się do procedur analizy pojęciowej i rekonstrukcji argumentów, a nie do prowadzenia badań empirycznych. Triangulację argumentacyjną rozumiem jako porównanie różnych rodzajów materiału teoretycznego – literatury filozoficznej, analiz z zakresu etyki sztucznej inteligencji oraz ustaleń badań empirycznych przywoływanych w literaturze – bez nadawania im funkcji dowodowej, w celu wzajemnej kontroli przyjmowanych definicji i wniosków. Tak rozumiana kontrola porównawcza zwiększa wiarygodność ustaleń i ogranicza ryzyko dowolności interpretacyjnej. Z kolei sprawdzanie koherencyjne oznacza ocenę spójności i konsekwencji przyjętych założeń w eksperymentach myślowych oraz w analizie porównawczej stanowisk. Model minimalnych warunków podmiotowości poddają trzem sprawdzeniom pojęciowo-proceduralnym: (i) spójności semantycznej, (ii) odporności na *petitio principii*, (iii) zdolności rozstrzygania przypadków granicznych (Matthias, 2004). W razie rozbieżności przyjmuję regułę rozstrzygania, zgodnie z którą pierwszeństwo ma koherencja definicyjna, następnie odporność graniczna, a na końcu potencjał predykcyjny.

Obok eksperymentów myślowych i modelowych przypadków analitycznych wykorzystuję w rozdziale 5 „empirycznie zakotwiczony przypadek graniczny” (ang. *empirically grounded boundary case*) IDx-DR/LumineticsCore. Opublikowane wyniki kliniczne i dokumentacja regulacyjna stanowią podstawę wyłącznie twierdzeń opisowych o systemie. Zastosowanie modelu H1–H3 + WI do tego przypadku pozostaje analizą pojęciową: nie służy empirycznej weryfikacji hipotez, lecz sprawdzeniu, czy model pozwala odróżnić ograniczoną autonomię diagnostyczną, wskaźniki stabilności technicznej, artefakty audytowe i osadzenie instytucjonalne od minimalnej adresowalności proceduralnej oraz silnej rozliczalności (zob. podrozdz. 5.4).

Analiza pojęciowa ma charakter diachroniczny i śledzi ewolucję znaczeń kluczowych terminów od ujęć klasycznych (Arystoteles, Kant), przez koncepcję osoby moralnej (Strawson), po współczesne modele agentów etycznych (Moor). Rekonstrukcja oddziela składniki definicyjne aktualne dla artefaktów od historycznego balastu i prowadzi do hipotezy o minimalistycznych warunkach koniecznych do uznania statusu podmiotu moralnego, określonych w kategoriach normatywnych. W rozdziałach 3–6 sprawdzam koherencyjnie tę hipotezę względem warunków H1–H3 oraz WI: zbieżności wzmacniają model progowy, a rozbieżności służą jego korekcie pojęciowej.

Rekonstrukcja argumentacji odtwarza logiczne ciągi przesłanek w tekstach klasycznych i współczesnych oraz ocenia ważność rozumowań i prawdziwość założeń. Zestawiam sceptycyzm (Johnson, 2006; Hew, 2014) ze stanowiskami stopniowalnej podmiotowości (Floridi & Sanders, 2004; Gunkel, 2018). Analiza porównawcza przebiega dwuetapowo: (1) konstruuje mapę stanowisk i relacji inferencyjnych, na przykład tezę o konieczności zdolności do odczuwania oraz funkcjonalistyczne tezy o wystarczalności funkcji; (2) sprawdzam konsekwencje stanowisk za pomocą eksperymentów myślowych modelujących artefakty spełniające różne konfiguracje H1–H3. Takie podejście pozwala wskazać argumenty zachowujące moc wyjaśniającą po przejściu z domeny biologicznej do technologicznej oraz te, które redukują się do intuicji metafizycznych pozbawionych skuteczności normatywnej. Aby ujednolicić i porównać wyniki tych eksperymentów, stosuję miarę mocy wyjaśniającej.

„Miara mocy wyjaśniającej” (ang. *Explanatory Power Measure*, EPM) to skala 0–3 służąca do uporządkowania siły argumentów generowanych w eksperymentach myślowych. Operacjonalizacja obejmuje trzy kryteria: (1) koherencję definicyjną, czyli zgodność wniosków scenariusza z przyjętymi definicjami (adresowalność, intencjonalność funkcjonalna, kontrola metapoznawcza), ewidencjonowaną w macierzy zależności pojęciowych: definicje operacyjne – parametry scenariuszy – oceny miary mocy wyjaśniającej (EPM); (2) odporność graniczną, czyli utrzymanie mocy wyjaśniającej w obliczu kontrprzykładów, np. „agenta o wpływie etycznym pozbawionego refleksyjności (metaregulacji)”; (3) potencjał predykcyjny, czyli liczbę i jakość intersubiektywnie weryfikowalnych implikacji normatywnych i procedur praktycznych, np. dla audytów funkcjonalnych. Termin „predykcyjny” rozumiem normatywno-proceduralnie, nie zaś empiryczno-przyczynowo; nie odwołuję się do popperowskiej falsyfikowalności empirycznej. EPM porządkuje moc wyjaśniającą

konfiguracji H1–H3 i WI, a nie mierzy empirycznie stopnia osiągnięcia osobnego poziomu refleksyjnego.

B. Źródła i uzasadnianie ustaleń

Dobór literatury w niniejszej pracy opiera się na trzech uzupełniających się grupach źródeł. Po pierwsze, wykorzystuję teksty kanoniczne z filozofii moralnej i teorii umysłu – od Arystotelesa, przez Johna Locke’a, Davida Hume’a i Immanuela Kanta, po współczesnych autorów, takich jak P. F. Strawson, John Rawls, Harry Frankfurt czy Daniel Dennett. Po drugie, podstawę stanowią recenzowane publikacje z zakresu etyki sztucznej inteligencji, filozofii techniki oraz infoetyki – w szczególności prace Luciana Floridiego i Jeffa Sandersa, Virginii Dignum, Jamesa Moora, Joanny Bryson, Davida Gunkela czy Andreasa Matthiasa. Po trzecie, uwzględniam dokumenty normatywne i programowe, które kształtują współczesny dyskurs regulacyjny – raporty Komisji Europejskiej oraz wytyczne Grupy Ekspertów Wysokiego Szczebla ds. Sztucznej Inteligencji (HLEG on AI), rezolucje Parlamentu Europejskiego, a zwłaszcza Rozporządzenie Parlamentu Europejskiego i Rady (UE) 2024/1689 z dnia 13 czerwca 2024 r. w sprawie ustanowienia zharmonizowanych przepisów dotyczących sztucznej inteligencji (AI Act). W odniesieniu do literatury technicznej i analiz polityki publicznej przyjmuję wymóg aktualności – korzystam przede wszystkim z publikacji po 2000 roku – natomiast w przypadku filozofii klasycznej kryterium stanowi kanoniczność, rozumiana jako udokumentowany i trwały wpływ na rozwój refleksji nad podmiotowością moralną. Przyjmowana perspektywa liberalno-indywidualistyczna jest jawnie deklarowana, a jej dominację równoważy krytyczna analiza stanowisk alternatywnych, obejmujących m.in. ujęcia relacyjne i konstruktywistyczne.

Praca ma charakter analityczno-koncepcyjny; dlatego podstawowymi wymogami pozostają rygor koherencji, adekwatność pojęciowa oraz kompletność wyjaśnienia. Wyniki pozostają jednak zależne od przyjętych definicji i obciążone ryzykiem ekwiwokacji między tradycjami teoretycznymi. Zastosowana „miara mocy wyjaśniającej” (EPM) pełni funkcję porządkującą, nie zaś empirycznie weryfikacyjną; aby ograniczyć ryzyko *petitio principii* przy doborze kryteriów, ujawniam słownik i macierz zależności pojęciowych oraz zapewniam audytowalność procedur oceny. Kompletność modelu rozumiem jako zdolność klasyfikowania znanych i racjonalnie przewidywalnych przypadków przez konfiguracje H1–H3 oraz WI, bez dodawania odrębnego poziomu do typologii Moora. Zakres stosowalności wniosków dotyczy

artefaktów spełniających minimalne warunki określone w hipotezie roboczej oraz analizowanych konfiguracji H1–H3.

1.6. Układ dysertacji i logika wyvodu

Przyjęty układ ma charakter warstwowy oraz rekonstrukcyjno-syntetyczny. Analiza rozwija się od rekonstrukcji pojęciowej i genealogicznej, przez operacjonalizację i porównawczą ocenę stanowisk, aż po modelowanie implikacji normatywnych i regulacyjnych. Metody – analiza pojęciowa, rekonstrukcja argumentów i krytyczna komparacja – zostały ułożone tak, aby wynik każdego etapu stanowił założenie robocze dla kolejnego. Konkluzje z etapu n kalibrują etap $n+1$ w sposób jednokierunkowy, nigdy odwrotnie. Dzięki temu maleje ryzyko błędnego koła w uzasadnieniu (*petitio principii*), a przejrzystość wnioskowania wzrasta (Popper, 1959). Skuteczność tego rozwiązania wynika z tego, że każdy etap wprowadza nowe, logicznie niezależne ustalenia (definicje, kryteria, wskaźniki), które następnie znajdują zastosowanie na kolejnym poziomie.

Rozdział 1 pełni funkcję wprowadzenia. Ustanawia aparat pojęciowy, formułuje hipotezy robocze oraz określa ograniczenia badania. Rozdział 2 rekonstruuje klasyczne ujęcia podmiotowości (Arystoteles, Kant, Strawson), wydobywając przesłanki sporów o sprawczość moralną. Rozdział 3 przedstawia gradacyjne kryteria minimalnej (super)podmiotowości moralnej (SPM) oraz ich operacjonalizację (por. podrozdz. 3.1). Rozdział 4 porządkuje ramy normatywne (etyczne) w perspektywie pluralistycznej (etyka cnót, deontologia, konsekwencjalizm). Rozdział 5 przedstawia klasyfikację systemów AI oraz analizuje ich ograniczenia w odniesieniu do układu H1–H3 + WI, ze szczególnym uwzględnieniem przypadków granicznych, ograniczeń audytowalności i funkcji kontrastów negatywnych. Oprócz przykładów implementacji sztucznej moralności obejmuje przypadek IDx-DR/LumineticsCore, oparty na opublikowanych danych i dokumentacji, wykorzystany do odrębnej oceny H1, H2/K4, H3 oraz WI. Obecne systemy AI nie są tam traktowane jako przykłady spełnienia modelu, lecz jako przypadki pokazujące różnice między autonomią operacyjną, stabilnością techniczną, audytowalnością, osadzeniem instytucjonalnym i agentową adresowalnością proceduralną. Rozdział 6 analizuje lukę odpowiedzialności, możliwe miary oceny oraz gradacyjny model współodpowiedzialności człowiek–AI w świetle ram regulacyjnych, zwłaszcza AI Act.

Na tym tle wyodrębniam dwa filary pracy. Filar I stanowi progowy model odpowiedzialności AI oparty na H1–H3 oraz WI: H1 wyznacza minimalną

adresowalność proceduralną i słabą odpowiedzialność proceduralną, H2 stabilizuje profil decyzji w czasie, H3 wyznacza silną rozliczalność poprzez kontrolę drugiego rzędu i jawne określanie reguł pierwszeństwa, a WI określa warunki instytucjonalnego zastosowania modelu. Model ten należy odczytywać przyszłościowo: jako układ warunków dla przyszłych, hipotetycznych albo bardziej zaawansowanych systemów, a nie jako opis spełniany obecnie przez istniejące systemy AI. Filar II stanowi gradacyjny model współodpowiedzialności człowiek–AI, rozwinięty w rozdziale 6 i osadzony w ramach unijnych instrumentów etyczno-prawnych.

Funkcję pozytywnego przypadku sprawdzającego w pracy pełni hipotetyczny system oceny odstępstw w obszarze pomocy społecznej. Nie jest to opis rzeczywistego urzędu, konkretnej procedury prawnej ani projektu technicznego. Przypadek ten służy jako modelowy przykład analityczny, pozwalający sprawdzić, jak H1–H3 + WI działałyby w odniesieniu do systemu przygotowującego warunkowe rozstrzygnięcie, wskazującego podstawę decyzji, ujawniającego przesłanki normatywno-decyzyjne, przedstawiającego regułę pierwszeństwa oraz wyznaczającego punkty audytu. Ostateczna odpowiedzialność pozostaje jednak po stronie ludzi i instytucji.

Obok niego analizuję IDx-DR/LumineticsCore jako przypadek graniczny oparty na dokumentacji. Przypadek ten nie przedstawia pełniejszej konfiguracji H1–H3 + WI i nie służy empirycznemu potwierdzeniu modelu. Jego funkcją jest pokazanie, że audytowalność systemu, techniczna powtarzalność i rozliczalność instytucjonalna mogą występować bez wykazanej agentowej responsywności na racje i bez kontroli drugiego rzędu (zob. podrozdz. 5.4).

Dla spójności lektury każdy rozdział kończy się zwięzłym podsumowaniem zawierającym wnioski przejściowe, które wprost pokazują, jak uzyskane ustalenia uzasadniają kolejny krok. Ustalenia historyczno-pojęciowe z rozdziału 2 uzasadniają doprecyzowanie progów H1–H3 w rozdziale 3; wyniki tej operacjonalizacji, ujęte w perspektywie pluralistycznych ram normatywnych rozdziału 4, kształtują klasyfikację i analizę realizacji technicznych w rozdziale 5; ta analiza przygotowuje grunt pod ujęcie luki odpowiedzialności, metryk oraz modelu współodpowiedzialności w rozdziale 6; rozdział 7 weryfikuje spójność H1–H3, WI oraz modelu współodpowiedzialności.

1.7. Zakres i ograniczenia badania oraz definicje kluczowych pojęć

Dysertacja korzysta wyłącznie z procedur analityczno-koncepcyjnych właściwych filozofii analitycznej, ujawniając przesłanki normatywne, lecz bez roszczenia do empirycznej falsyfikacji ani triangulacji empirycznej. Trafność wniosków oceniam przez spójność logiczną i adekwatność pojęciową, a nie przez konfrontację z danymi empirycznymi. Rezygnacja z badań empirycznych zwiększa przejrzystość definicji, ale zarazem zawęża zakres ich zastosowania. W rozdziałach 5–7 wskazuję obszary, w których symulacje agentowe, audyty etyczne oraz teoretyczne studia przypadków mogą w przyszłych badaniach interdyscyplinarnych służyć do sprawdzenia koherencji i praktycznej adekwatności przyjętych hipotez. W niniejszej dysertacji odniesienia te mają charakter ilustracyjny i heurystyczny, a nie dowodowy ani implementacyjny (Winfield, Blum & Liu, 2014).

Z tego powodu przykłady obecnych systemów AI pełnią w pracy przede wszystkim funkcję kontrastową. Pokazują ograniczenia istniejących układów wobec H1–H3 + WI, lecz nie stanowią dowodu na to, że obecne systemy osiągają próg odpowiedzialności proceduralnej ani próg silnej rozliczalności. Jedyne pozytywne przypadki pełniejszej konfiguracji modelu ma charakter hipotetyczny i modelowy. IDx-DR/LumineticsCore pełni odmienną funkcję: jest udokumentowanym przypadkiem granicznym, służącym sprawdzeniu mocy rozróżniającej modelu, a nie jego empirycznemu wdrożeniu lub potwierdzeniu.

W dalszej analizie „sztuczna inteligencja” oznacza system interaktywny, autonomiczny i adaptacyjny w klasycznym ujęciu (Russell & Norvig, 2021), przy czym interaktywność rozumiem jako zdolność do dwukierunkowej wymiany informacji i oddziaływania z otoczeniem, autonomię jako podejmowanie decyzji bez bieżącego nadzoru człowieka w granicach wyznaczonych celów, a adaptacyjność jako uczenie się i modyfikowanie sposobów działania na podstawie danych. Rozważam wyłącznie niebiologiczne artefakty działające samodzielnie w zakresie realizacji powierzonych celów; w pracy termin „artefakt” oznacza sztucznie skonstruowany system techniczny, zaprojektowany i wytworzony w procesie inżynierskim przez człowieka, którego funkcjonowanie opiera się na przetwarzaniu informacji oraz który w toku działania wykazuje zdolność interakcji z otoczeniem, podejmowania decyzji bez nadzoru oraz uczenia się i modyfikowania sposobów działania.

Definicja wyklucza: (1) automaty reaktywne pozbawione zdolności uczenia się, (2) cyfrowe odwzorowania świadomości biologicznej, (3) zbiorowości ludzi wspomaganych przez AI oraz hybrydy biologiczno-technologiczne. Poza zakresem pracy pozostają także systemy, które nie spełniają minimalnych wymogów interaktywności, autonomii i adaptacyjności (Floridi & Sanders, 2004), a także byty biologiczne inne niż ludzie oraz problematyka podmiotowości zwierząt. Ogranicza to możliwość porównań z przypadkami granicznymi, takimi jak moralna sprawczość niemowląt czy samoświadomych zwierząt; ewentualne odniesienia do nich mają jedynie charakter heurystycznych kontrastów pojęciowych. „Szare strefy”, takie jak systemy socjotechniczne i implanty neuronowe, omawiam jedynie marginalnie jako kierunki dalszych badań (zob. rozdz. 6 niniejszej dysertacji) i powracam do nich podczas rekonstrukcji sporów granicznych.

Ograniczenie zakresowe do systemów interaktywnych, autonomicznych i adaptacyjnych dotyczy pozytywnych przypisań agentowości oraz podmiotowości. W funkcji kontrastowej mogą zostać przywołane także istniejące systemy, których publiczna dokumentacja nie pozwala wykazać wszystkich trzech własności, o ile analiza służy wyłącznie rozróżnieniu poszczególnych progów modelu i nie prowadzi do pozytywnego przypisania H1–H3. Taką funkcję pełni przypadek IDx-DR/LumineticsCore, w odniesieniu do którego publiczne materiały nie pozwalają stwierdzić adaptacyjności po wdrożeniu.

Termin „studium przypadku” oznacza tu wyłącznie eksperyment myślowy albo modelowy przypadek analityczny w sensie filozofii nauki. Nie chodzi o empiryczne studium przypadku, które obejmuje zbieranie danych oraz analizę obserwacyjną. W niniejszej pracy studium przypadku służy sprawdzeniu konsekwencji pojęciowych przyjętych definicji oraz ujawnieniu ich implikacji normatywnych; materiał badawczy stanowią konsekwencje logiczne wynikające ze spójnego zastosowania zdefiniowanych warunków do modelowej reprezentacji artefaktu.

Przez „empirycznie zakotwiczony przypadek graniczny” rozumiem natomiast istniejący system opisany na podstawie opublikowanych danych i dokumentacji, który służy do analitycznego sprawdzenia konsekwencji modelu bez prowadzenia przez autora badań empirycznych. Materiał źródłowy ma moc dowodową wobec twierdzeń opisowych o systemie, natomiast zastosowanie H1–H3 + WI pozostaje argumentem filozoficznym i oceną koherencyjną.

Znaczenia kluczowych terminów i funkcje oznaczeń zebrano w części „Słownik pojęć i oznaczeń”, umieszczonej po rozdziale 7, przed bibliografią. Zestawienie ma charakter pomocniczy: porządkuje przyjęte rozróżnienia i odsyła do ich rozwinięcia w odpowiednich podrozdziałach, nie wprowadzając dodatkowych warunków modelu H1–H3 + WI.

W warstwie normatywnej odwołuję się do trzech tradycji: deontologii, konsekwencjalizmu i etyki cnót. Analizy nie obejmują perspektyw postkolonialnych ani niezachodnich, np. konfucjańskich, co ogranicza uniwersalność wniosków. Wszystkie wnioski normatywne mieszczą się w kanonie etyki zachodniej: deontologicznej, konsekwencjalistycznej i aretologicznej, rozwijanej w myśli europejskiej i północnoamerykańskiej od starożytności po współczesność. Przyjmuję zatem, że: (i) definicje wartości, obowiązku i sprawczości zakładają liberalno-indywidualistyczną koncepcję osoby; (ii) minimalne kryteria podmiotowości moralnej rekonstruuje wyłącznie w tym paradygmacie; (iii) zastosowanie ich do innych tradycji, np. konfucjańskiej, ubuntu czy etyk rdzennych, wymaga odrębnej analizy porównawczej. To samoograniczenie zakresu ma zapewnić spójność pojęciową i zapobiec nieuzasadnionym uogólnieniom międzykulturowym.

Szczegółową charakterystykę „luki odpowiedzialności”, wraz z propozycją jej pojęciowego uporządkowania i implikacjami regulacyjnymi, przedstawiam w rozdziale 6; tutaj jedynie sygnalizuję, że problem ten wyznacza górną granicę mocy wyjaśniającej zastosowanej metodologii. Ponieważ przyjęta metodologia nie dostarcza empirycznych narzędzi do przypisywania odpowiedzialności, wnioski w tym zakresie mają przede wszystkim charakter uzasadniający, a nie rozstrzygający. Dalsze badania interdyscyplinarne mogłyby zostać uzupełnione o empiryczne studia przypadków oraz testy funkcjonalne oceniające praktyczną przydatność zaproponowanych kryteriów; nie należą one jednak do metody niniejszej dysertacji. W tym sensie ograniczenia wskazane w podrozdziale 1.5 wyznaczają zarazem program dalszych dociekań w filozofii AI.

Uzupełnieniem eksperymentów myślowych będzie systematyczna rekonstrukcja „sporów granicznych”, czyli przypadków spełniających kryteria jedynie częściowo, np. artefaktów o autonomii instrumentalnej lub systemów pozbawionych perspektywy diachronicznej. Analiza tych quasi-podmiotów pełni dwie funkcje. Tworzy mechanizm kontroli inferencyjnej, ponieważ brak rozstrzygającego argumentu wymusza rewizję przynajmniej jednego kryterium oraz wyznacza pole rozszerzeń modelu dla przyszłych

badań interdyscyplinarnych, czyniąc tę konstrukcję odporną na zarzut arbitralności mimo braku empirycznych narzędzi weryfikacyjnych.

Zamknięcie rozdziału 1 wyznacza przejście od analizy współczesnych wyzwań konceptualnych, technologicznych i normatywnych związanych z możliwością sztucznej podmiotowości moralnej do rekonstrukcji podstaw pojęciowych, które konstytuują to zagadnienie w obrębie tradycji filozoficznej. Aby ocenić, czy i w jakim sensie artefakty technologiczne mogą aspirować do statusu podmiotów moralnych, trzeba najpierw precyzyjnie ustalić, czym jest „podmiot” w sensie epistemologicznym, mentalnym i sprawczym. Rozdział 2 pełni funkcję fundamentu pojęciowego. Na nim opierają się dalsze analizy warunków moralnej sprawczości AI: od kartezjańskiego cogito i kantowskiej autonomii woli po stanowiska relacyjne i konstruktywistyczne, ze szczególnym uwzględnieniem przesłanek uzasadniających przypisywanie odpowiedzialności moralnej. Dalsze rozdziały należy zatem czytać jako analizę warunków możliwej odpowiedzialności proceduralnej i rozliczalności, a nie jako twierdzenie, że obecne systemy AI już spełniają te warunki. Dopiero na tym gruncie możliwe będzie krytyczne zbadanie, czy w systemach sztucznej inteligencji mogą wystąpić cechy analogiczne, a także jakie implikacje to pociąga za sobą dla redefinicji granic wspólnoty moralnej.

Rozdział 2. Podmiotowość w wybranych klasycznych i współczesnych koncepcjach filozoficznych

Wprowadzenie

Rozdział 2 analizuje funkcjonalny i normatywny problem podmiotowości. Interesują mnie warunki, w których sensowne jest orzekanie o jedności poznawczej podmiotu, przypisywanie autorstwa czynów oraz przyznawanie statusu adresata norm (adresowalności normatywnej), rozumianej jako zdolność bycia adresatem norm moralnych i prawnych. Przyjmuję perspektywę porównawczą. Zestawiam trzy klasyczne ujęcia (Arystoteles, Kant, P. F. Strawson) z nurtami krytycznymi i sceptycznymi (Hume, Parfit, Galen Strawson, Metzinger). Celem jest rekonstrukcja minimalnych warunków podmiotowości odpornej na spory ontologiczne. Rozdział wypełnia lukę: brak jednolitego ujęcia łączącego trzy tradycyjne funkcje podmiotu. Są nimi: (1) nośnik stanów i perspektywy, (2) sprawca działań zależnych od racji oraz (3) adresat roszczeń i odpowiedzialności.

Teza przewodnia brzmi: rdzeń podmiotowości ma charakter funkcjonalno-relacyjny. Innymi słowy, o tym, czy zasadne jest mówienie o podmiocie w sensie normatywnym, decydują konkretne funkcje, a nie jego metafizyczny status. O statusie normatywnym decydują trzy elementy: jedność poznawcza, sprawczość rozumiana jako responsywność na racje oraz adresowalność związana z odpowiedzialnością praktyczną. Wynika stąd zasada ontologicznej obojętności minimalnych warunków odpowiedzialności: o przypisywaniu sprawczości decydują identyfikowalność podmiotu („kto?”), uzasadnialność i kontrola – nie zaś przyjęcie określonej ontologii „głębokiego ja”.

Rozdział 2 ma układ problemowy. W podrozdziale 2.1 rekonstruuje pojęcie podmiotu jako nośnika stanów i sprawcy działań oraz pokazuję, że nie wymaga to tezy o substancjalnym „ja”. W podrozdziale 2.2 przechodzę do podmiotu moralnego: wyodrębniam minimum adresowalności oraz silną rozliczalność. W 2.3 omawiam świadomość i kontrolę, w 2.4 – krytyki ontologiczne, w 2.5 – implikacje dla sztucznej podmiotowości.

2.1. Podmiot myślenia, poznania i działania

W tym podrozdziale pytam o minimalne warunki „podmiotu”, które pozwalają mówić o jedności myślenia i odpowiedzialnym działaniu. Nie odwołuję się przy tym do tez o substancjalnym „ja”. Teza robocza: podmiotowość ma charakter funkcjonalno-relacyjny. To struktura utrzymywana w czasie, złożona z rozpoznawalnych zdolności poznawczych i wykonawczych, które zapewniają kontrolę sprawczą zależną od racji. „Kontrola sprawczą zależną od racji” oznacza wrażliwość deliberacyjną i wykonawczą na uzasadnione racje, przejawiającą się w planowaniu, realizacji i korekcie działań. Rozumiem ją jako kontrolę wykonawczą, czyli zdolność do implementacji planów, obejmującą m.in. hamowanie reakcji i adaptację. Dla zachowania precyzji rozróżniam ramę funkcjonalno-relacyjną (perspektywa teoretyczna) od jej profilu funkcjonalno-relacyjnego (operacjonalizacja za pomocą obserwowalnych wskaźników).

W tradycji arystotelesowskiej podmiot – *hypokeimenon* – jest tym, „o czym” orzeka się własności oraz stanowi podstawę zarówno zmiany, jak i trwałości. W *O duszy* Arystoteles formułuje ujęcie hylemorficzne: *psyche* jako forma organizmu żywego pełni funkcję zasady życia oraz źródła funkcji poznawczych i ruchowych (Arystoteles, 2003, „DA” II 1, 412a20–412b9; por. II 3, 414a31–b1). W *Etyce nikomachejskiej* pojęcia rozumnego wyboru (*proairesis*) i dobrowolności wiążą autorstwo czynów z dyspozycjami charakteru oraz praktycznym rozumem, który kształtuje celowość działania (Arystoteles, 2012, „EN” III 1–5, 1109b30–1115a3; por. II 1, 1103a15–1103b25). Historycznie rzecz ujmując, „podmiot” pełni tu rolę zarówno substratu tożsamości organizmu, jak i źródła celowo ukierunkowanych aktów.

Kantowskie przejście od substancjalnego „ja” do transcendentalnej funkcji jednoczącej doświadczenie przesuwają punkt ciężkości pojęcia podmiotu. Jedność apercpepcji – słynne „ja myślę” – nie opiera się na hipotezie substancjalnego podmiotu, lecz stanowi warunek możliwości doświadczenia: wszelkie przedstawienia muszą dać się zjednoczyć w samoświadomości jednego aktu myślenia (Kant, 2001, B131–B132). Po pierwsze, jedność poznawcza pełni funkcję formalnego warunku doświadczenia; po drugie, ponieważ stosowanie pojęć można weryfikować i uzasadniać, kontrola nad regułami jest z natury publiczna i realizuje się w praktykach wzajemnego domagania się uzasadnień i korekt. W ujęciu Strawsona „osoba” stanowi punkt odniesienia dla

publicznych kryteriów identyfikacji – tego, kto doświadcza i działa – nie przesądzając jednak o ontologicznym statusie „ja” (Strawson, 2003, s. 99–102). Według McDowella doświadczenie jest od początku pojęciowe. Zdolności rozumne, które umożliwiają sądy, zarazem kształtują identyfikację podmiotu (McDowell, 1996, s. 48–50, 190–191). Jedność nie powstaje „nad danymi”, lecz tkwi w sposobie, w jaki świat staje się dostępny dla rozumu.

Aby uniknąć niejasności i nieporozumień, wprowadzam następujące rozróżnienia pojęciowe. Przez „jedność poznawczą” rozumiem formalno-regulową syntezę przedstawień w ramach „ja myślę”, poddającą się intersubiektywnej kontroli za pośrednictwem praktyk uzasadniania i korekty. „Jedność podmiotowa” to właściwość systemu, który utrzymuje owe funkcje: zarówno jedność poznawczą, jak i spójny porządek sprawczy, obejmujący planowanie, koordynację oraz rewizję w czasie. Termin „podmiot” stosuję w ujęciu ogólnym, odnoszącym się zarówno do bytów naturalnych, jak i sztucznych. Z kolei „osoba”, rozumiana w duchu Strawsona, stanowi kategorię pierwotną, do której przynależą predykaty zarówno cielesne, jak i mentalne. W przypadkach niekanonicznych (np. systemy sztuczne) odwołuję się do publicznie rozpoznawalnego agenta sztucznego jako funkcjonalnego odpowiednika osoby.

Klasyczny punkt odniesienia dla tej rekonstrukcji stanowi kantowska formuła: „ja myślę, musi móc towarzyszyć wszystkim moim wyobrażeniom” (Kant, 2001, § 16, B131–B132). W niniejszym wywodzie formuła ta pełni funkcję historyczno-rekonstrukcyjną: nie oznacza, że H1 wymaga transcendentnej samoświadomości lub świadomości fenomenalnej, lecz wskazuje klasyczny model jedności przedstawień, wobec którego można opisać funkcjonalną adresowalność racji.

Przejsie od analizy jedności do kwestii sprawczości wymaga ukazania, w jaki sposób „jedność” przekłada się na autorstwo czynu. U Anscombe działanie intencjonalne definiuje teleologiczny porządek odpowiedzi na pytanie „dlaczego?”, a więc działanie podlegające opisowi odwołującemu się do racji sprawcy (Anscombe, 2000, s. 9–12; por. s. 11, 14–15). Davidson dodaje komponent przyczynowy, wskazując, że racje – przekonania i pragnienia – mogą być przyczynami działania, co pozwala zharmonizować wyjaśnienie normatywne z mechanizmem sprawczym (Davidson, 1963, s. 685–686). Bratman dopełnia obraz w wymiarze diachronicznym: intencje, rozumiane jako plany, koordynują działania w czasie, rozstrzygają konflikty celów i umożliwiają współdziałanie

dzięki zgraniu (ang. *meshing*) podplanów (Bratman, 1987, s. 2; s. 15–16; s. 64–71). Razem ujęcia te pokazują trzy aspekty sprawstwa: normatywny, przyczynowy i diachroniczny. W niniejszej pracy przyjmuję, że zakres uzasadnionego przypisania autorstwa czynu jest tym większy, im silniej responsywność na racje koreluje ze spójną i planową organizacją celów.

Anscombe ujmuje to przejście w formule, zgodnie z którą „intentional actions are ones to which a certain sense of the question «Why?» has application” — w przekładzie roboczym: „działania intencjonalne to działania, do których stosuje się pewien sens pytania «dlaczego?»” (Anscombe, 2000, s. 11, § 6). Cytat wzmacnia tezę, że intencjonalność działania ma strukturę racji, a nie tylko strukturę zdarzenia kauzalnego.

Wniosek można ująć następująco: jedność poznawcza zakłada funkcjonalną syntezę przedstawień według reguł, poddawaną intersubiektywnej kontroli w praktykach uzasadniania i korekty. Ponieważ poprawność stosowania pojęć wymaga publicznej weryfikacji, identyfikacja nośników tych aktów opiera się na kryteriach publicznych. W praktykach społecznych nośnik ten rozpoznaje się jako osoba w sensie Strawsona. W przypadkach niekanonicznych ten nośnik przyjmuje postać publicznie identyfikowalnego agenta, stanowiącego równoważny punkt odniesienia dla przypisań sprawczości. Działanie intencjonalne jest wyznaczane przez racje oraz organizowane przez plany, które koordynują je w czasie (por. Anscombe, 2000, s. 9–12; Davidson, 1963, s. 685–686; Bratman, 1987, s. 2; s. 15–16; s. 64–71). Wynika z tego, że minimalne warunki podmiotowości praktycznej spełnia każdy system, który utrzymuje jedność poznawczą w sensie regułowej syntezy, stanowi publicznie identyfikowalny nośnik sprawczości oraz wykazuje kontrolę zależną od racji w planowaniu, realizacji i rewizji działań.

Aby uniknąć niejasności operacyjnych, należy doprecyzować pojęcia robocze. Przez „politykę decyzyjną” rozumiem ustaloną regułę wyboru działania w określonej klasie sytuacji. Stabilność polityk w czasie nie wyklucza wrażliwości na kontrprzykłady; przeciwnie, zakłada ich uwzględnianie w ramach spójnej korekty planów. Odróżniam „adresowalność epistemiczną” od „adresowalności praktycznej”. „Adresowalność epistemiczną” definiuję jako możliwość domagania się uzasadnienia przekonań, a „adresowalność praktyczną” – jako możliwość żądania korekty działań i przypisywania sprawczości. Ponieważ praktyki uzasadniania mają z natury charakter wzajemny,

minimalna „adresowalność epistemiczna” stanowi warunek konieczny dla „adresowalności praktycznej”.

Tak określone minimum pozwala mówić o podmiotach niekanonicznych, w tym o systemach złożonych i rozproszonych. Jeżeli pojęcie „osoby” traktuje się jako kategorię praktyczną, to w odpowiedzi na pytanie „kto działa?” najważniejsze są stabilne kryteria przypisywania autorstwa czynów oraz możliwość ich publicznej weryfikacji albo ich funkcjonalne odpowiedniki. Jeżeli „jedność poznawcza” stanowi warunek funkcjonalny, to kluczowe znaczenie zyskuje stabilność porządku pojęciowego oraz zdolność do jego utrzymania w obliczu nowych danych. Jeżeli „kontrola sprawcza zależna od racji” sprowadza się do reagowania na uzasadnienia i korekty planów, to na tym etapie wystarczy odwołać się do podrozdziału 2.3, gdzie wskazuję klasy ocen wymagające świadomości fenomenalnej.

Na tej podstawie można wyróżnić pojęciowe kryteria jedności podmiotowej. Warunek jedności spełnia system, który: (1) utrzymuje koherencję ocen wobec różnych reprezentacji tego samego stanu rzeczy, (2) zachowuje stabilne polityki, a zarazem reaguje na kontrprzykłady oraz (3) pozwala na publiczną rekonstrukcję ciągu racji prowadzących do działania – na minimalnym poziomie audytowalności. Dwa pierwsze kryteria mają charakter konieczny, a wszystkie trzy łącznie są wystarczające, aby mówić o jedności podmiotowej. Trzecie kryterium wskazuje minimalny próg normatywnej transparentności, rozumianej jako możliwość oceny działań na podstawie dostępnych uzasadnień w praktykach przypisywania sprawczości oraz rozliczania.

Zamykając tę część, przechodzę do podrozdziału 2.2, w którym zawężam perspektywę do podmiotu moralnego i ukazuję, jak arystotelesowska rozumność praktyczna, kantowska autonomia oraz strawsonowskie praktyki reaktywne przekładają się na warunki „adresowalności normatywnej” i odpowiedzialności. Analiza koncentruje się na wymiarze poznawczo-normatywnym i praktycznym podmiotowości; nie rozstrzygam tu statusu fenomenologicznego „ja” ani nie formułuję silnych twierdzeń metafizycznych. W odniesieniu do systemów sztucznych przyjmuję epistemicznie oszczędne kryteria jedności i sprawczości, które można uznać za wystarczające w ramach większości praktyk odpowiedzialności i oceny, natomiast klasy ocen odwołujące się do krzywdy doznaniowej omawiam osobno w podrozdziale 2.3.

2.2. Podmiot moralny w koncepcjach klasycznych i współczesnych

W perspektywie arystotelesowskiej status moralny opiera się na rozumie praktycznym, cnotach charakteru oraz rozumnym wyborze. Czyny są „nasze”, gdy są dobrowolne i wynikają ze stabilnych dyspozycji kształtowanych przez praktykę, wychowanie i deliberację. Dobrowolność to nie tylko brak przymusu. Oznacza działanie zgodne z rozsądną oceną tego, co jest najlepsze w danej sytuacji. Podmiot moralny to byt, którego działania ujawniają stabilną, rozsądnie ukierunkowaną strukturę dyspozycji. Pozwala to przypisać mu autorstwo i uzasadnić ocenę moralną. Schemat jest prosty: czyn przypisujemy, jeśli jest dobrowolny i podlega ocenie rozumu praktycznego. Dobrowolność i rozsądnosc wynikają z dyspozycji ukształtowanych przez nawyk i refleksję. Podmiot moralny ujawnia stabilną strukturę dyspozycji. Zastosowanie tej koncepcji najlepiej widać u sprawcy o stałej tożsamości praktycznej. Ograniczeniem pozostaje brak rozstrzygnięć w przypadkach rozproszonych i zbiorowych oraz brak jasnego rozróżnienia między „byciem ocenianym” a „byciem właściwym adresatem roszczeń drugiej osoby”. Przypisywalność wyjaśnia, dlaczego wiążemy czyny ze sprawcą. Aby sprawca był adresatem roszczeń drugiej osoby, potrzebna jest kompetencja do udzielania racji. To otwiera przejście do ujęcia kantowskiego oraz współczesnych analiz udzielania racji.

Nawiązując do arystotelesowskich ujęć, u Kanta definicja osoby moralnej wiąże się z autonomią: osoba stanowi prawo moralne (samolegisłacja) i podporządkowuje wolę imperatywowi kategorycznemu (Kant, 2002, AA 4:429–430; Kant, 2004, s. 57–72). Kant stwierdza: osobę należy traktować jako cel sam w sobie, nie tylko jako środek. Podmiot moralny to byt, wobec którego kierowane są roszczenia normatywne: prawo do szacunku oraz zakaz instrumentalności. Schemat: (1) Prawo moralne jest dostępne rozumowi praktycznemu. (2) Zdolność podporządkowania się temu prawu tworzy autonomicznego nosiciela powinności. Wniosek: podmiot moralny podporządkowuje wolę rozumowi i posiada godność jako cel (Kant, 2002, AA 4:431–434; Kant, 2004, s. 161–163). Autonomia, interpretowana funkcjonalnie, nie wymaga substancjalnego „ja”, lecz stabilnych mechanizmów samoregulacji: rozpoznawania reguł, podporządkowania działań, rewizji motywów oraz udzielania racji pod publiczną kontrolą. Adresowalność jako relacja między osobami tworzy prawo żądania racji oraz jej ocenę. Publiczna

audytowalność jest warunkiem formalnym relacji z drugą osobą, a nie tylko retorycznym dodatkiem.

Podążając dalej, Strawson przesuwając dyskusję o odpowiedzialności z metafizycznych sporów na grunt społecznych praktyk uzasadniania i rozliczania. Odpowiedzialność nie zależy od rozstrzygnięć o wolnej woli; wywodzi się z nastawień reaktywnych – urazy, wdzięczności, oburzenia, przeprosin – które kształtują relacje społeczne. Te postawy podlegają regułom zawieszania przypisywania odpowiedzialności w przypadku przymusu, niedojrzałości, choroby oraz innych okoliczności. Podmiot moralny to uczestnik relacji wzajemnego adresowania i odpowiadania, którego status podlega regułom przypisywania, zawieszania i przywracania odpowiedzialności. Schemat: społeczne praktyki uzasadniania i rozliczania konstytuują normy przypisywania odpowiedzialności; posiadają własne kryteria zawieszania i przywracania; podmiot moralny uczestniczy w relacjach adresowania i odpowiadania oraz podlega regułom zawieszania i przywracania statusu.

Rozszerzając ten wątek, zakres tego ujęcia jest szeroki: perspektywa drugiej osoby obejmuje również przypadki, w których kwestia wolności pozostaje otwarta. Warunek stosowalności wobec podmiotów zbiorowych można ująć następująco: rozszerzenie kategorii podmiotu moralnego na agentów instytucjonalnych jest uzasadnione tylko wtedy, gdy instytucja spełnia trzy wymogi. Po pierwsze, dysponuje zinstytucjonalizowanymi kanałami udzielania racji (publicznego uzasadniania działań). Po drugie, podlega rozpoznawalnym sankcjom; po trzecie, działa zgodnie z publicznymi regułami zawieszania i przywracania przypisania odpowiedzialności. Podsumowując: relacyjność i praktyki konstytuują ramę, w której rozliczalność staje się możliwa i podlega audytowi.

W ostatnich dekadach pojęcie „odpowiedzialności” rozbito na kategorie porządkujące zakres i wymóg kompetencji. Przypisywalność (ang. *attributability*): czyny i stany przypisujemy podmiotowi, gdy odzwierciedlają jego charakter lub postawy, nawet bez sankcji relacyjnych (Watson, 1996; Smith, 2005). Udzielanie racji (ang. *answerability*): adresat musi uzasadniać swoje działania; inni oczekują korekty w przypadku sprzeczności, co wymaga audytowalności publicznej. Rozliczalność (ang. *accountability*): podmiot jest adresatem roszczeń i sankcji; według Darwalla pojawia się naprzeciw drugiej osoby, której roszczenie wyrasta z norm relacyjnych (Darwall, 2006;

Wallace, 1998). Scanlon wyróżnia trzy wymiary oceny moralnej: dopuszczalność, znaczenie czynu, obwinianie, które nie sprowadzają się do jednego testu (Scanlon, 2008). Porządkujący schemat: przypisywalność $\rightarrow$ (+ udzielanie racji) $\rightarrow$ (+ reguły drugiej osoby) = rozliczalność. Przypisywalność definiuje, co można przypisać podmiotowi jako wyraz postaw; udzielanie racji wyjaśnia, dlaczego podmiot publicznie uzasadnia i koryguje swoje działania; reguły drugiej osoby, uprawnienia i sankcje przekształcają to w rozliczalność. Każdy element jest potrzebny, ale żaden sam nie wystarcza. Ciągłość dyspozycji nie gwarantuje obowiązku publicznego uzasadnienia; autonomia formalna nie określa reguł zawieszania przypisywania odpowiedzialności; praktyki bez minimalnych kompetencji adresata prowadzą do formalizmu. Model responsywności na racje (ang. *reasons-responsiveness*),⁴ wskazujący relację między kontrolą a zdolnością reagowania na uzasadnienia i kontrfaktyczne warianty działania. Responsywność na racje nie zastępuje wymogu autonomii, lecz go operacjonalizuje, czyniąc go sprawdzalnym w ocenie kontroli. To napięcie stanowi oś podrozdziału 2.3.

W tym kontekście wprowadzam rozróżnienie między minimalnymi warunkami adresowalności a wystarczającymi warunkami rozliczalności. Minimalne warunki adresowalności są następujące: podmiot moralny w sensie minimalnym spełnia trzy wymagania. Po pierwsze, rozpoznaje i stosuje reguły ogólne do przypadków jednostkowych w sposób pojęciowy (a nie czysto asocjacyjny). Po drugie, potrafi przyznawać rację oraz korygować swoje przekonania i intencje w świetle kontrargumentów, zachowując przy tym publiczną audytowalność. Po trzecie, zachowuje minimalną ciągłość identyfikacyjną i przewidywalność profilu działania, co umożliwia sensowne przypisanie czynów i postaw temu samemu adresatowi. Nie jest to jeszcze stabilność diachroniczna oceniana przez K4, lecz słabszy warunek identyfikacji podmiotu odpowiedzi w praktykach adresowania roszczeń. Minimum jest spełnione, gdy podmiot potrafi: (1) wyjaśnić, dlaczego dana reguła stosuje się w danej sytuacji oraz (2) zmienić swoje działanie po przedstawieniu trafnego kontrprzykładu. Przejście od minimalnej adresowalności do silnej rozliczalności wymaga rozdzielenia poziomu systemowego i instytucjonalnego. Po stronie technicznego systemu H1 musi pozostać spełniona, a dodatkowo H3 wymaga monitorowania własnego procesu decyzyjnego, rozpoznawania

⁴ Termin „responsywność na racje” (*reasons-responsiveness*) pochodzi z teorii Fischera i Ravizy i oznacza zdolność agenta do reagowania na uzasadnienia normatywne oraz do modyfikowania działania w świetle rozpoznanych racji, także kontrfaktycznych. Koncepcja ta stanowi kompatybilistyczną próbę operacjonalizacji kontroli sprawczej (zob. Fischer & Ravizza, 1998, s. 28–91, 207–240).

konfliktu przesłanek normatywno-decyzyjnych, jawnego wskazywania reguły pierwszeństwa oraz przedstawiania audytowalnego uzasadnienia drugiego rzędu. Planowanie, hamowanie reakcji i zwykle rozstrzyganie konfliktów celów pozostają formami kontroli pierwszego rzędu i same nie wystarczają do H3. Po stronie układu społeczno-technicznego WI wymaga osadzenia systemu w publicznie rozpoznawalnych regułach drugiej osoby, w tym w procedurach zawieszania i przywracania przypisywania odpowiedzialności, ocenianych pod względem proporcjonalności oraz symetrii uprawnień stron. Krótko mówiąc, aby można było traktować system jako „odpowiedzialny przed kimś” w silniejszym sensie, musi on nie tylko sensownie odpowiadać na pytanie „dlaczego?”, lecz także spełniać próg H3, a jego zastosowanie musi podlegać WI.

Wniosek roboczy polega na tym, że status podmiotu moralnego można wiarygodnie określić bez odwoływania się do daleko idących założeń dotyczących substancjalnego „ja”. Wystarcza zestaw funkcjonalno-relacyjnych kompetencji poznawczo-praktycznych oraz włączenie się do sieci roszczeń drugiej osoby. Stanowisko to rozwija i łączy trzy tradycje: od Arystotelesa przejmuje ciągłość charakteru i rozum praktyczny, od Kanta – formalny kształt normatywności i autonomię, od Strawsona – konstytutywną rolę praktyk i postaw reaktywnych. Nota zakresowa wskazuje, że przedstawiona rekonstrukcja dotyczy normatywnego sensu podmiotowości i nie rozstrzyga sporów o tożsamość osobową ani o „głębką” wolność metafizyczną; zakłada natomiast, że dla rozumienia odpowiedzialności miarodajne są praktyki jej przypisywania. To prowadzi do podrozdziału 2.3, w którym rozważę, czy i kiedy świadomość fenomenalna jest konieczna dla adresowalności i rozliczalności oraz jaki zakres kontroli sprawczej można uznać za wystarczający. Posłużę się przy tym analizą pojęciową oraz analizą przypadków granicznych.

2.3. Spory o świadomość, wolną wolę i odpowiedzialność

W kontekście odpowiedzialności nie pytam o tajemniczą „moc” wyboru. Interesuje mnie zdolność do udziału w wymianie racji: ujawnianie motywów, rozumienie kontrargumentów, rewizja planów i porządkowanie zamiarów pod publiczną ocenę. Zdolność tę nazywam „adresowalnością normatywną”. Jej funkcjonalnym odpowiednikiem po stronie dostępu do racji może być forma świadomości zapewniająca

ustrukturyzowany dostęp do treści myśli i reguł działania, umożliwiająca wewnętrzną deliberację i korektę – tę formę określam jako świadomość dostępu (ang. *access consciousness*; A-świadomość). Świadomość fenomenalna (ang. *phenomenal consciousness*; P-świadomość), czyli subiektywne doświadczenie – jak to jest coś przeżywać – może wzmacniać ocenę w kontekstach krzywdy doznaniowej, lecz sama w sobie nie stanowi uniwersalnego kryterium statusu podmiotowego. W tej perspektywie punkt ciężkości przesuwają się z metafizycznych spekulacji na widoczną z zewnątrz praktykę uzasadniania i korekty w ramach przypisywania odpowiedzialności, a więc na odpowiedź na pytanie, czy ktoś jest zdolny uczestniczyć w wymianie uzasadnień, przyjmując zarzut i odpowiednio zmienić swoje działanie. Tak rozumiana adresowalność łączy dyskusje o świadomości i wolnej woli. Prowadzi do roboczej tezy, że spór o „ostateczne” podstawy nie przesądza o sensowności praktyk społecznych przypisywania odpowiedzialności. Uzasadnienie tej tezy rozwijam poniżej.⁵ Doprecyzowanie to nie oznacza, że H1 zawiera warunek świadomościowy: H1 oznacza minimalną adresowalność proceduralną, a A-świadomość pełni tu funkcję funkcjonalnego ekwiwalentu dostępności racji lub kryterium pomocniczego w ocenie tej adresowalności.

Rozróżnienie między P-świadomością – czyli jakościami „jak to jest” związanymi z przeżywaniem stanów – a A-świadomością – ustrukturyzowanym dostępem do treści myśli i reguł działania, warunkującym deliberację, ocenę racji, korektę planów oraz hamowanie reakcji sprzecznych z uznanymi zasadami – ma znaczenie nie tylko terminologiczne, lecz także funkcjonalne i normatywne. Dla uzasadnionego przypisywania odpowiedzialności kluczowa jest ta forma świadomości, która udostępnia rozumowi treści myśli: pozwala wiązać reprezentacje z regułami, ujawnia racje oraz dostosowuje zachowanie do zarzutów i perspektyw innych (Block, 1995, s. 227–232; Dennett, 2016, rozdz. 3–4; Levy, 2014, rozdz. 2, s. 14–37; rozdz. 6, s. 109–131). Fenomenalność, o ile ją przyjmuję, może pełnić funkcję wskaźnika wrażliwości wówczas, gdy przedmiotem oceny jest doświadczenie bólu lub poniżenia; nie stanowi jednak warunku uczestnictwa w praktykach uzasadniania. A-świadomość jest warunkiem koniecznym jedynie wtedy, gdy analizuję świadomościowy dostęp podmiotu do racji; w

⁵ „Pozorna responsywność” odnosi się do sytuacji, w których raportowane racje nie stanowią rzeczywistej przyczyny decyzji, lecz są wynikiem konfabulacji lub efektem „dostrajania do wskaźnika”. Z perspektywy adresowalności normatywnej sama zgodność z wynikiem nie jest więc wystarczająca; konieczny jest dający się wykazać związek przyczynowy między rozpoznaną racją a dokonanym wyborem, możliwy do ujawnienia w ramach przyjętego poziomu abstrakcji (LoA) oraz poddawany ocenie w praktykach uzasadniania (por. Nisbett & Wilson, 1977; Floridi & Sanders, 2004).

modelu H1 jej rolę może pełnić funkcjonalny ekwiwalent dostępności racji. W praktykach udzielania racji i rozliczalności instytucjonalnej taka dostępność racji bywa także wystarczająca do racjonalnego przypisywania odpowiedzialności proceduralnej. Nie rozstrzyga jednak o „ostatecznej” winie.

Rozróżnienie to można zwięźle ująć za Blockiem: „P-conscious content is phenomenal, whereas A-conscious content is representational” (Block, 1995, s. 232). W modelu H1 A-świadomość pełni co najwyżej funkcję pomocniczego opisu dostępności racji albo jej funkcjonalnego ekwiwalentu; nie jest składnikiem minimalnej adresowalności proceduralnej.

Adresowalność ma charakter stopniowalny. Im stabilniejszy i bogatszy dostęp do racji działania oraz im lepiej sprawca radzi sobie z konfliktem celów, hamowaniem reakcji i integracją planów, tym silniejsze staje się roszczenie, by odpowiadał w trybie uzasadniania. Im słabszy ten dostęp, tym bardziej uzasadnione jest zawieszenie lub ograniczenie przypisywania odpowiedzialności (Levy, 2014, rozdz. 2, s. 14–37; rozdz. 6, s. 109–131). Przyjmuję regułę progową: zawieszam przypisywanie odpowiedzialności, gdy brak jest minimalnej zdolności rozpoznawania i stosowania reguł; osłabiam ją natomiast, gdy zdolność ta jest zachowana, lecz chwiejna albo ograniczona kontekstowo.

Kusi teza, że odpowiedzialność wymaga „realnych” alternatyw – zgodnie z Zasadą Alternatywnych Możliwości (ang. *Principle of Alternative Possibilities*, PAP), sformułowaną przez H. G. Frankfurta w artykule *Alternate Possibilities and Moral Responsibility* (The Journal of Philosophy, 1969, s. 829–839). Przeciwno temu przemawiają jednak tzw. przypadki frankfurtowskie: jeżeli decyzja wypływa z uprzednio ukształtowanych motywów, z którymi sprawca się utożsamia, a ingerencja zewnętrzna jedynie „gwarantuje” wynik, to ocena odpowiedzialności nie traci podstaw, nawet jeśli alternatywa została wyłączona. Trzeba jednak zastrzec, że przypadki frankfurtowskie jedynie osłabiają, a nie unieważniają zasadę PAP. W kontekstach „przebłysków wolności” (ang. *flickers of freedom*) oraz ominięcia (ang. *bypassing*) ciężar przypisywania odpowiedzialności może bowiem rozkładać się inaczej.

Ominięcie to ingerencja, która omija proces legitymowania i internalizacji mechanizmu sprawczego przez agenta (np. manipulacja, wszczepienie polityk decyzyjnych). W takich przypadkach warunek historyczny, czyli warunek własności

mechanizmu i przyjęcia odpowiedzialności, stanowiący podstawę kompatybilistycznej koncepcji kontroli zależnej od racji, nie zostaje spełniony (Fischer & Ravizza, 1998). Tzw. repliki miękkiej linii wskazują, że nawet przy synchronicznej responsywności na rację brak „dobrej historii” podważa przypisanie pełnej odpowiedzialności (Mele, 2019; McKenna, 2008). Intuicję tę wzmacniają argumenty z manipulacji, pokazujące ciąg przypadków, w których spełnienie wymogów kompatybilistycznych współwystępuje z nieakceptowalną genezą mechanizmu sprawczego (Pereboom, 2001). Dlatego w ramach zasady superweniencji praktyk przypisywania odpowiedzialności (ZSP), której warunki oznaczam niżej jako ZSP(i)–ZSP(iii), przyjmuję regułę braku ominięcia: przypisania odpowiedzialności są zasadne tylko wtedy, gdy mechanizm powstał i daje się korygować w trybie, który nie pomija kompetencji rozpoznawania i ważenia racji (por. podrozdział 2.4, gdzie syntetycznie porządkuję warunki ZSP(i)–ZSP(iii)).

W ujęciu kompatybilistycznym „kontrola” oznacza przede wszystkim responsywność na racje oraz przyjęcie mechanizmu sprawczego za własny (warunek własności mechanizmu lub utożsamienia się z nim).⁶ Sprawca jest odpowiedzialny, jeżeli w odpowiednich kontekstach jego system decyzyjny reaguje na rozpoznawalny wachlarz racji lepszych i gorszych, a sam sprawca przyjmuje wobec tego systemu postawę identyfikacji oraz gotowości do bycia adresatem krytyki. Aby odpowiedzialność miała pełny charakter, konieczny jest także warunek historyczny: mechanizm nie może być ukształtowany przez manipulację, która uniemożliwia jego racjonalne nabycie i późniejszą korektę (Fischer & Ravizza, 1998, rozdz. 2, s. 28–61; rozdz. 3, s. 62–91; rozdz. 8, s. 207–239).

Naturalnie nie wszyscy akceptują takie ujęcie. Argument z konsekwencji wskazuje, że jeśli determinizm jest prawdziwy, nasze działania stanowią nieuchronny skutek przeszłości i praw natury – a więc czynników, nad którymi nie mamy władzy.⁷ Jak zatem można ponosić „głęboką” odpowiedzialność? Libertarianie podkreślają rolę działań „samokształtujących”, podejmowanych w warunkach nieusuwalnej otwartości,

⁶ „Ownership condition” (warunek własności mechanizmu) oznacza u Fischera i Ravizy (1998) wymóg, by sprawca utożsamiał się ze swoim systemem decyzyjnym i przyjmował go jako własny. Bez tego identyfikacyjnego komponentu responsywność synchroniczna nie wystarcza do pełnego przypisania odpowiedzialności.

⁷ Wskazuje się, że jeśli determinizm jest prawdziwy, to każdy czyn stanowi konieczny skutek przeszłości i praw natury; ponieważ nie ma władzy ani nad przeszłością, ani nad prawami, nie ma także władzy nad ich konsekwencjami, w tym nad własnym działaniem, co podważa „głęboką” kontrolę wymaganą przez odpowiedzialność w sensie zasłużenia (por. van Inwagen, 1983, s. 56).

które mają nadawać sprawcy źródłowość wyboru: nie tylko reaguje on na racje, lecz inicjuje moment „tworzący” nowy kierunek życia (van Inwagen, 1983, s. 56; Kane, 1996, s. 126–135). Replika twardej linii i replika miękkiej linii⁸ podtrzymują intuicję, że odpowiedzialność wymaga czegoś więcej niż samej responsywności na racje – wymaga autorstwa niewykłanego w prawo przyczynowe albo w szczęście moralne.⁹

Przedstawiciele twardego inkompetybilizmu twierdzą, że niezależnie od tego, czy prawdziwy jest determinizm, czy indeterminizm, nie ma wolnej woli w sensie wystarczającym do ponoszenia „ostatecznej” winy. Skoro tak, praktyki przypisywania odpowiedzialności należy nastawić na skutki i prewencję: uzasadnione są te formy reakcji, które chronią dobra, kształtują zachowania i stabilizują zaufanie, a nie te, które odwołują się do odpłaty opartej na zasługiwaniu (Pereboom, 2001, s. 123–157).¹⁰ Inni autorzy wiążą odpowiedzialność z rozumnością: sprawca odpowiada wtedy, gdy działa z zachowaną zdolnością rozpoznawania i ważenia wartości; brak tej zdolności usprawiedliwia wyłączenie lub złagodzenie odpowiedzialności (Wolf, 1990, s. 68–90).

W tle sporu obecny jest tzw. argument podstawowy Galena Strawsona.¹¹ Głosi on, że jeżeli pełna, „ostateczna” odpowiedzialność miałaby wymagać bycia *causa sui* – czyli całkowitej odpowiedzialności za to, kim się jest – to nikt nie spełniałby tego warunku (Strawson, 1994, s. 5–7). Innymi słowy, nikt nie może być ostatecznie odpowiedzialny za własne motywy i dyspozycje, gdyż zawsze kształtują się one w oparciu o wcześniejsze uwarunkowania, na które sprawca nie miał wpływu. W konsekwencji podważone zostają wyłącznie roszczenia do odpowiedzialności

⁸ „Replika miękkiej linii” (ang. *soft-line reply*) polega na twierdzeniu, że w przypadkach frankfurtofskich zawsze pozostaje pewien „przeblask wolności” (ang. *flicker of freedom*) – minimalna alternatywna możliwość wystarczająca, by zachować odpowiedzialność sprawcy.

⁹ Pojęcie „moral luck” (szczęście moralne) zostało spopularyzowane przez Thomasa Nagela (1979) i Bernarda Williamsa (1981). Oznacza ono sytuacje, w których zakres przypisywanej odpowiedzialności zależy od czynników pozostających poza kontrolą sprawcy – takich jak konsekwencje działania, okoliczności zewnętrzne czy warunki urodzenia.

¹⁰ Pojęcie „desert” (dosł. zasługa) jest terminem technicznym w filozofii moralnej. Oznacza roszczenie do nagrody lub kary ze względu na to, co ktoś uczynił, niezależnie od jego skutków. W sporach o odpowiedzialność rozróżnia się odpowiedzialność pragmatyczną (np. prewencja, ochrona społeczna) od odpowiedzialności opartej na zasłudze (ang. *desert-based responsibility*), która zakłada, że sprawca „zasłużył” na odpłatę w sensie moralnym. Krytycy, m.in. Galen Strawson i Derk Pereboom podważają możliwość takiej odpowiedzialności, wskazując, że wymagałaby ona bycia *causa sui* (ostatecznym źródłem własnego istnienia).

¹¹ Utrzymuje się, że aby być ostatecznie odpowiedzialnym za czyn, należałoby być ostatecznie odpowiedzialnym za to, jakim się jest; to zaś wymagałoby bycia *causa sui*, co jest niemożliwe. Wniosek: ostateczna odpowiedzialność moralna (oparta na zasługaniu) jest niemożliwa (por. Strawson, 1994, s. 5–7).

retributywnej, czyli odpowiedzialności w sensie odpłaty (ang. *retributive responsibility*), opartej na idei zasłużonej winy i kary. Nie oznacza to jednak unieważnienia praktyk odpowiedzialności zakorzenionych w innych uzasadnieniach, takich jak ochrona społeczna, korekta zachowań czy podtrzymywanie relacji opartych na normach społecznych.

Odpowiedzi na argumenty sceptyczne odwołują się do logiki praktyk stosowanych jeszcze, zanim zostanie postawione pytanie o ostateczne fakty metafizyczne. Praktyki przypisywania odpowiedzialności trwają, bo koordynują i stabilizują relacje moralne. Publiczne uzasadnienia zapewniają przewidywalność, budują zaufanie i umożliwiają naprawę krzywd – także przy utrzymującej się niepewności metafizycznej. P. F. Strawson przypomina, że nasze relacje moralne są naznaczone postawami reaktywnymi; wycofuje się je nie z powodu odkryć dotyczących przyczynowości, lecz z powodu zaniku warunków rozumnej adresowalności (Strawson, 2019, s. 35–57). R. Jay Wallace rekonstruuje tę intuicję w języku reguł: winy nie przypisuje się tam, gdzie u źródła działania leży brak zdolności podporządkowania się racjom wymaganym przez praktykę; przypisuje się ją natomiast tam, gdzie ktoś mógł i powinien był działać w świetle tych racji, a tego nie uczynił (Wallace, 1998, s. 125). Stephen Darwall dodaje perspektywę drugiej osoby: odpowiedzialność rozumie on jako wzajemne roszczenie do uzasadnienia, korekty i szacunku (Darwall, 2006, rozdz. 4, s. 65–90; rozdz. 5, s. 91–118).

W tym sensie postawy reaktywne można rozumieć jako praktyki wyrażające wzajemny wymóg międzyosobowego uznania, a ich zawieszenie jako ograniczenie uczestniczącej relacji odpowiedzialności (Strawson, 2008, s. 17). Parafraza ta wzmacnia relacyjny sens odpowiedzialności praktycznej bez wprowadzania dodatkowego sporu metafizycznego.

Z powyższych analiz wyłania się trójdzielny wniosek porządkujący kwestie odpowiedzialności praktycznej. Po pierwsze, minimalnym warunkiem jest dostępność racji: u podmiotów świadomych realizuje ją A-świadomość, a w modelu H1 jej funkcjonalny ekwiwalent umożliwia publiczne ujawnianie i korygowanie racji. Po drugie, kontrola polega przede wszystkim na responsywności wobec racji oraz na właściwej relacji do własnych mechanizmów (uwzględniającej także warunek historyczny), a nie na dostępie do „realnych” alternatyw w sensie zasady PAP. Przypadki

frankfurtowskie pokazują, że brak takich alternatyw nie podważa przypisywania odpowiedzialności, o ile zachowana jest odpowiednia relacja sprawcy do własnych mechanizmów i nie dochodzi do ominięcia. Po trzecie, nawet jeśli zawiesi się roszczenia do „ostatecznej” winy – pod wpływem argumentu z konsekwencji, libertariańskiej koncepcji źródłowości czy sceptycyzmu *causa sui* – pozostaje szeroki i normatywnie bogaty obszar odpowiedzialności praktycznej. Rozumiem ją jako odpowiedzialność za udzielanie uzasadnień oraz jako rozliczalność instytucjonalną w ramach reguł drugiej osoby. Tego rodzaju odpowiedzialność nie jest „okrojona” wersją wielkiej metafizycznej idei; stanowi raczej to, czego rzeczywiście wymaga się od uczestników praktyk wzajemnego uzasadniania, w warunkach wspólnego życia, sporów i naprawiania krzywd.

To rozstrzygnięcie przesuwaa akcent z żądania metafizycznego uzasadnienia odpowiedzialności na pragmatyzm normatywny. Zamiast pytać, czy sprawca jest ostatecznym autorem samego siebie, pytam, czy jest stroną, do której sensownie można skierować żądanie: „uzasadnij” albo „odwołaj” oraz czy nasza wspólna praktyka, w której słowa te mają moc, jest prowadzona zgodnie ze sprawiedliwymi regułami. Nie oznacza to rezygnacji z ambicji filozoficznej; przeciwnie, jest to wybór, by rozumność i wzajemne zobowiązania przedłożyć nad spory o tajemniczą „iskrę” wolności oraz by filozofia moralna nie czekała na ostateczne rozstrzygnięcia metafizyczne, zanim dostarczy kryteriów działania i oceny.

Na tym tle wyraźniej widać granice dotychczasowej argumentacji. Kwestię relacji między fenomenalnością a miarami krzywdy doznaniowej, która wymaga odrębnej analizy fenomenologiczno-normatywnej, pozostawiam otwartą: inaczej bowiem należy oceniać odpowiedzialność tam, gdzie w grę wchodzi naruszenie reguły, a inaczej tam, gdzie znaczenie ma doznana krzywda. Skoro „adresowalność normatywna” nie wymaga założenia substancjalnego „ja”, kolejnym krokiem jest pytanie, jak broni się ona wobec krytyk samej idei osoby – redukcjonizmu, animalizmu, stanowisk braku „ja” oraz eliminatywizmu. Jeżeli podstawowe praktyki przypisywania odpowiedzialności mają zostać zachowane – wzajemne domaganie się uzasadnień i gotowość do ich korygowania – przy osłabieniu metafizycznych założeń dotyczących podmiotu, należy wykazać, że da się je utrzymać także na gruncie mniej zobowiązującej ontologii.

2.4. Stanowiska krytyczne i sceptycyzm ontologiczny

Po ustaleniach z wcześniejszych partii wywodu stawiam pytanie, jaki minimalny zakres pojęcia „ja” trzeba zachować, aby mogło ono pełnić funkcję w praktykach przypisywania odpowiedzialności, jeśli potraktować sceptycyzm ontologiczny poważnie. Podmiotowość normatywna opiera się na dwóch elementach: (1) zdolności bycia adresatem roszczeń uzasadniania i korekty (adresowalności) oraz (2) sprawowaniu kontroli zależnej od racji. Nie pytam o to, czy „ja” istnieje jako byt *sui generis*. Pytam, czy – i kiedy – można sensownie traktować kogoś jako uczestnika praktyk drugiej osoby: adresata uzasadnień, korekt i usprawiedliwień. Interesuje mnie nie tyle natura „ja”, ile warunki sensownego uczestnictwa w praktykach drugiej osoby. Pytam, czy stanowiska takie jak redukcjonizm, animalizm, eliminatywizm czy koncepcje braku „ja” podważają warunki adresowalności, czy tylko zmieniają opis ich tła ontologicznego (Parfit, 1984, s. 299–300; Olson, 1997, s. 94–123; Hume, 2023, s. 319–331; Dennett, 2016, rozdz. 3–4; Metzinger, 2019, s. 9–11)? Na potrzeby dalszego wywodu przyjmuję ustalenia z podrozdziału 2.3, zgodnie z którymi obrona odpowiedzialności opiera się na A-świadomości i responsywności na racje, bez konieczności odwoływania się do tezy o realnych alternatywach w sensie zasady PAP (por. Frankfurt, 1969, s. 829–839; Fischer & Ravizza, 1998, s. 28–61, 62–91, 207–239; Levy, 2014, s. 14–37, 109–131).

Teza robocza głosi, że nawet jeśli odrzuci się metafizyczny obraz „głębokiego ja”, nie zostaje utracone to, co konstytuuje odpowiedzialność praktyczną. Proponuję Zasadę superweniencji praktyk przypisywania odpowiedzialności (ZSP). Praktyki przypisywania odpowiedzialności superweniują na trzech warunkach: (i) identyfikowalności adresata („kto?”), (ii) audytowalności racji oraz (iii) kontroli zależnej od racji. Jeżeli warunki ZSP(i)–ZSP(iii) są spełnione, przyjęta ontologia podmiotu (np. „wiązka własności”, „organizm”) nie decyduje o tym, czy i w jaki sposób uczestnicy praktyk rozliczają się wzajemnie. Dla ważności przypisywania odpowiedzialności wystarcza spełnienie tych warunków. ZSP nie rozstrzyga natury „ja”, lecz porządkuje elementy praktyk przypisywania odpowiedzialności, które nie podlegają negocjacji wraz ze zmianami metafizycznych teorii.

Przyjmuję, że odpowiedzialność praktyczna nie wymaga zasady PAP (Frankfurt, 1969, s. 829–839; Fischer & Ravizza, 1998, s. 28–61, 62–91). Argumenty

frankfurtowskie przesuwają akcent z „mocy alternatyw” na mechanizmy decyzyjne: reagujące na racje, rozpoznawane jako „nasze” i poddające się publicznej korekcie. W tym sensie pole odpowiedzialności opiera się na regułach udzielania racji i rozliczalności, a nie na roszczeniu do *causa sui* (Strawson, 2019, s. 35–57; Wallace, 1998, s. 51–83; Darwall, 2006, s. 87–118; Pereboom, 2001, s. 127–157). Ramę tę traktuję jako kryterium przewodnie analizy stanowisk: w każdym przypadku pytam o status warunków ZSP(i)–ZSP(iii).

W pierwszej kolejności należy rozważyć redukcjonizm psychologiczny. W ujęciu klasycznym nie istnieje żaden „dodatkowy fakt” o tożsamości poza relacjami ciągłości i łączności psychologicznej: tożsamość opiera się na sieci pamięci, intencji i dyspozycji deliberacyjnych (Parfit, 1984, s. 299–300). Scenariusze myślowe wskazują, że praktycznie istotna jest ciągłość relacyjna (cierpienie, zobowiązania, projekty), a nie substancjalny podmiot (Parfit, 1984, s. 253–255, 299–300; por. Dennett, 2016, rozdz. 3–4). Z perspektywy ZSP redukcjonizm wzmacnia ZSP(ii)–ZSP(iii); ZSP(i) wymaga jedynie utrzymania rozpoznawalnego profilu w czasie.

Animalizm zakłada, że osoby są przede wszystkim organizmami; trwają tak długo, jak długo trwa żywy organizm, nawet przy zerwaniu łączności psychologicznej (Olson, 1997, s. 94–123). Stanowisko to zabezpiecza ZSP(i), lecz samo nie gwarantuje ZSP(ii)–ZSP(iii): identyfikacja biologiczna nie zastępuje publicznej ocenialności racji ani responsywności wobec racji (Strawson, 2008, s. 1–28; Wolf, 1990, s. 73–90). W tym sensie animalizm rozwiązuje problem nośnika, ale nie gwarantuje warunków adresowalności.

Stanowiska eliminacyjne i koncepcje braku „ja” idą krok dalej: traktują „ja” jako konstrukt lub model (Hume, 2023, s. 319–331; Dennett, 2016, rozdz. 3–4; Metzinger, 2019, s. 9–11). Odwołuję się zatem do „warunku tła” (WT): minimalna A-świadomość lub jej funkcjonalny ekwiwalent zapewniający dostępność treści decyzyjnych dla audytu i korekty (w skrócie: możliwość raportowania/korekty racji oraz hamowania reakcji sprzecznych z uznanymi racjami). Gdy WT oraz ZSP(i)–ZSP(iii) są spełnione, „eliminacja” dotyczy ciężaru metafizycznego, nie zaś funkcji moralnej (Dennett, 2016, rozdz. 3–4; Metzinger, 2019, s. 9–11).

Powyższe analizy prowadzą do następującej rekonstrukcji. Po pierwsze, praktyki przypisywania odpowiedzialności kształtują się w przestrzeni wzajemnych roszczeń i odpowiedzi (Strawson, 2008, s. 1–28; Darwall, 2006, s. 87–118). Po drugie, aby praktyki przypisywania odpowiedzialności miały sens, adresat musi dysponować A-świadomością albo funkcjonalnym ekwiwalentem dostępności racji oraz kontrolą zależną od racji (Block, 1995, s. 227–247; Dennett, 2016, rozdz. 3–4; Levy, 2014, s. 14–37, 109–131; Fischer & Ravizza, 1998, s. 28–61, 62–91, 207–239). Po trzecie, żadne z rozważanych ujęć „ja” nie unieważnia automatycznie ZSP(i)–ZSP(iii); czynią to dopiero konkretne deficyty WT – brak możliwości identyfikacji podmiotu, brak audytu racji, brak responsywności (Fischer & Ravizza, 1998, s. 207–239; Wolf, 1990, s. 73–90). Z powyższych rozważań wynika zasada, którą można określić mianem „Zasady ontologicznej obojętności minimalnych warunków odpowiedzialności” (ZO).¹² Głosi ona, że z punktu widzenia ważności przypisywania odpowiedzialności różne ontologie „ja” są równoważne, jeśli tylko nie naruszają minimalnych warunków dostępności racji, identyfikowalności racji, audytowalności racji i kontroli zależnej od racji. ZO koryguje silne roszczenia animalizmu – pogląd, że samo wskazanie «kto?» automatycznie rozstrzyga o zakresie odpowiedzialności («za co?») – oraz skrajnego redukcjonizmu, który zakłada, że wystarcza jakakolwiek łączność psychologiczna, nawet przy braku audytu racji. Warto w tym miejscu doprecyzować, że ZO nie zakłada normatywnej równoważności ontologii we wszystkich aspektach – nie obejmuje ona rozkładu sankcji, kompensacji ani polityk naprawczych, lecz odnosi się wyłącznie do ważności przypisywania odpowiedzialności.

Przypadki graniczne – takie jak rozszczepienia tożsamości, zaburzenia świadomości czy systemy sztuczne, w których trudno o antropomorficzne utożsamienie się z mechanizmem decyzyjnym – potwierdzają potrzebę precyzyjnej operacjonalizacji warunków przypisywania odpowiedzialności. ZSP(i) wymaga jedynie, aby w ramach danej wspólnoty praktyk można było stabilnie wskazać „kto?”. ZSP(ii) żąda możliwości rekonstrukcji uzasadnień (choćby skrótowych lub częściowo automatycznych), ich publicznej krytyki i korygowalności; za wystarczające uchodzą: spójne uzasadnienia *ex*

¹² „Zasada ontologicznej obojętności minimalnych warunków odpowiedzialności” rozwija intuicje obecne u Parfita (1984, s. 215–216) – brak „dalszego faktu” o tożsamości, u P. F. Strawsona (2019, s. 35–57) – prymat praktyk reaktywnych nad metafizyką oraz u Wallace’a (1998) – normatywne ujęcie odpowiedzialności. Zbieżne idee pojawiają się także u kompatybilistów takich jak Dennett (1984, 2016) i Levy (2014), którzy podkreślają, że przypisywanie odpowiedzialności wymaga funkcjonalnych kompetencji, a nie substancjalnego „ja”.

post, ślad decyzji (log), dostrzegalna poprawa decyzji i działań w odpowiedzi na korektę oraz konsekwentne hamowanie reakcji sprzecznych z uznanymi racjami. ZSP(iii) wymaga sprawdzalnej responsywności na racje w różnych kontekstach: hamowania impulsów, hierarchizacji celów i aktualizacji planów w odpowiedzi na zarzuty. Gdy warunki te ulegają osłabieniu, słabnie również podstawa przypisywania odpowiedzialności; gdy całkowicie zanikają, ustaje przypisywanie odpowiedzialności. Zależność ta nie wymaga dodatkowych założeń metafizycznych – wystarcza opis społecznych praktyk uzasadniania i rozliczania, w których dokonuje się przypisania odpowiedzialności.

W tle wszystkich tych sporów powraca kwestia wolnej woli (van Inwagen, 1983, s. 56; Frankfurt, 1969, s. 829–839). Jeśli przyjąć argumenty frankfurckie przeciw konieczności realnych alternatyw działania (tj. możliwości postąpienia inaczej), wówczas ciężar wyjaśnienia odpowiedzialności spoczywa nie na „metafizycznej mocy” wyboru, lecz na tym, w jaki sposób mechanizmy decyzyjne reagują na racje i pozostają normatywnie rozpoznawane jako „nasze” – to znaczy rozpoznane jako własne i poddawane publicznej korekcie (Fischer & Ravizza, 1998, s. 62–91; Levy, 2014, s. 109–131). W tym ujęciu redukcjonizm psychologiczny i narracyjny ułatwiają opis tak rozumianej kontroli; animalizm pozostaje wobec niej neutralny, koncentrując się jedynie na nośniku; eliminatywizm zaś nie musi jej podważać, o ile WT i ZSP(i)–ZSP(iii) pozostają spełnione. Sceptycyzm – w postaci „argumentu z konsekwencji” i „argumentu podstawowego” (van Inwagen, 1983, s. 56; Strawson, 1994, s. 5–7) – osłabia roszczenia dotyczące „ostatecznej winy”. Nie podważa jednak odpowiedzialności praktycznej, opartej na udzielaniu racji i rozliczalności. Sfera rozliczalności działa bowiem w ramach reguł drugiej osoby, a nie na podstawie pretensji do *causa sui* (Strawson, 2019, s. 35–57; Darwall, 2006, s. 87–118; Wallace, 1998, s. 51–83; Pereboom, 2001, s. 174–179). Dlatego zamiast pytać o niemożliwy do spełnienia warunek, by podmiot był całkowitym źródłem samego siebie (*causa sui*), pytam, czy sensownie można skierować do niego żądanie uzasadnienia i czy potrafi on na nie odpowiedzieć.

Powyższe ustalenia przygotowują grunt pod analizę w podrozdziale 2.5 poświęconą implikacjom dla sztucznej podmiotowości. Konieczne jest przełożenie ZSP i WT na wymogi operacyjne, które posłużą do oceny przypadków modelowych i granicznych – zarówno ludzkich, jak i sztucznych. Kryteria identyfikowalności „kto?”,

audytowalności racji i kontroli zależnej od racji można sformułować tak, by zachować wrażliwość na kontekst (medyczny, prawny, technologiczny), a jednocześnie uniknąć fetyszyzacji narzędzi: to praktyki wyznaczają miary, a nie odwrotnie. Ogólny wniosek z tej części wywodu brzmi: odpowiedzialność nie wymaga substancjalnego „ja”, lecz podmiotów zdolnych do posługiwania się racjami i poddawania ich ocenie publicznej.

2.5. Implikacje dla sztucznej podmiotowości

Powyższa analiza prowadzi do tezy, że odpowiedzialność może być rozumiana przede wszystkim w kategoriach relacyjnych: polega na zobowiązaniu do rozumnego uzasadniania swoich czynów wobec innych, którzy są zdolni nas rozumieć i oceniać. Dopóki zachowany jest wspólny horyzont racji, spór o metafizyczne fundamenty podmiotowości pozostaje kwestią wtórną. Sprawcą jest byt, który potrafi ujawniać i korygować racje w odpowiedzi na pytania, zarzuty oraz roszczenia innych. W tym miejscu pojawia się pytanie, jakie konsekwencje ma takie ujęcie dla systemów sztucznej inteligencji. Skoro odpowiedzialność można rozumieć bez konieczności odwoływania się do substancjalnego „ja”, należy sprawdzić przede wszystkim, czy system AI spełnia próg H1: rozpoznaje normy, przewiduje konsekwencje ich naruszeń, reaguje na racje w trybie kontrfaktycznym oraz udostępnia wystarczające i wiarygodne podstawy decyzji, umożliwiające audyt, zakwestionowanie i korektę. Stabilność diachroniczna jego profilu działania podlega odrębnej ocenie K4, natomiast publicznie audytowalne praktyki, role i procedury przypisywania odpowiedzialności należą do WI. W tym sensie rozważania z poprzednich podrozdziałów stanowią podstawę do zbadania, czy i w jakim zakresie można przypisać systemom AI status podmiotu moralnego.

Niniejszy podrozdział podsumowuje wnioski z analiz ontologiczno-genealogicznych i normatywnych przeprowadzonych w podrozdziałach 2.1–2.4 oraz koncentruje się na filozoficznych implikacjach dla pojęcia sztucznej podmiotowości. Z podrozdziału 2.1 wynika, że jedność podmiotową można ujmować funkcjonalno-relacyjnie, jako regułowe porządkowanie przedstawień i planów. To umożliwia identyfikację „kto?” i przypisanie autorstwa czynu. Podrozdział 2.2 doprecyzowuje minimalne i silne warunki statusu podmiotu moralnego w kategoriach przypisywalności, udzielania racji oraz rozliczalności. Podrozdział 2.3 wskazuje, że dla większości praktyk przypisywania odpowiedzialności wystarcza świadomość dostępu (A-świadomość) albo jej funkcjonalny ekwiwalent oraz kontrola rozumiana jako responsywność wobec racji. Podrozdział 2.4 pokazuje, że spory o „głębokie ja”

(redukcjonizm, animalizm, stanowisko braku „ja”, eliminatywizm) nie podważają podstawowych warunków przypisywania odpowiedzialności, o ile spełnione są kryteria identyfikowalności, uzasadnialności i kontroli. Na tym tle „sztuczna podmiotowość” nie zakłada tezy ontologicznej o istnieniu substancjalnego „ja”, lecz określa zakres możliwej partycypacji systemów sztucznych w praktykach uzasadniania i rozliczania.

Dla większej przejrzystości wyводу przyjmuję tu konwencję wynikową: kolejne partie analizy kończą się sformułowaniem „wyniku”, który porządkuje ustalenia cząstkowe i wskazuje ich implikacje.

Wynik pierwszy dotyczy „pokory ontologicznej”: wiele klasycznych sporów o substancjalność „ja” może pozostać nierozstrzygniętych bez szkody dla sensowności praktyk przypisywania odpowiedzialności. Wynika to z faktu, że społeczne praktyki przypisywania odpowiedzialności opierają się na postawach reaktywnych i wzajemnych roszczeniach, które kształtują rozumienie winy i usprawiedliwienia. W tej perspektywie akcent przesuwają się z „realnych” alternatyw na zdolność do udzielenia racji i korekty zachowania w świetle przedstawionych argumentów. Przypadki frankfurckie pokazują, że odpowiedzialność nie zawsze wymaga „realnych” alternatyw działania, o ile zachodzi adekwatny związek między czynem a mechanizmem sprawczym uznanym przez sprawcę za „własny”. Dlatego pytanie o status „ja” ustępuje miejsca pytaniu, czy i kiedy dany system wchodzi w krąg praktyk relacji drugiej osoby jako adresat i dawca racji (por. Strawson, 2008, s. 1–28; Frankfurt, 1969, s. 829–839).

Wynik drugi dotyczy świadomości. Dla przypisywania odpowiedzialności kluczowa jest A-świadomość – ustrukturyzowany dostęp do treści potrzebnych do deliberacji i uzasadniania. P-świadomość staje się konieczna jedynie wówczas, gdy treść doświadczenia fenomenalnego wchodzi w opisy krzywdy (Block, 1995; Levy, 2014; por. Dennett, 2016). Minimalny udział w praktykach odpowiedzialności wymaga dostępności racji oraz zdolności do ich korekty; u podmiotów świadomych tę dostępność opisuje A-świadomość, natomiast w modelu H1 może ją realizować funkcjonalny ekwiwalent dostępności racji. Warunek ten obejmuje: jawność (zapis racji), audytowalność (zewnętrzna kontrola) i wyjaśnialność („dlaczego X, a nie Y”). Skoro warunki te czynią uzasadnienia dającymi się publicznie oceniać, wyznaczają one minimalny próg udzielania racji, który, przy spełnieniu pozostałych wymogów, jest konieczny (choć niewystarczający) dla adresowalności w relacji z drugą osobą (por. Block, 1995, s. 227–247; Levy, 2014).

Trzeci wynik dotyczy relacyjnego charakteru norm moralnych. Perspektywa drugiej osoby (Darwall) akcentuje wzajemne uprawnienia do żądania racji i udzielania odpowiedzi: roszczenia moralne zakładają współustanowioną przestrzeń „my”, w której strony uznają się za kompetentnych adresatów roszczeń i uzasadnień moralnych. W odniesieniu do sztucznej podmiotowości pytanie „czy X jest podmiotem?” przekształca się w pytanie: „czy rozsądnie można zwrócić się do X z roszczeniem uzasadnienia i korekty oraz czy X potrafi odpowiedzieć w sposób dający się publicznie ocenić?”. Jak wskazano wyżej, spełnienie kryteriów świadomości dostępu wyznacza próg udzielania racji i przygotowuje warunki dla adresowalności, która dodatkowo wymaga uprawnionych roszczeń drugiej osoby do oceny i korekty (por. Darwall, 2006, rozdz. 4–5; Wallace, 1998).

Wynik czwarty dotyczy ciągłości tożsamości osobowej w czasie. Dyskusja w podrozdziale 2.4 – od redukcjonizmu psychologicznego, przez animalizm i eliminatywizm, po stanowiska braku „ja” – prowadzi do wniosku, że w praktykach przypisywania odpowiedzialności ważniejsza od „dodatkowego faktu” o osobowej esencji jest ciągłość rozrachunku, czyli stała możliwość rekonstrukcji „kto za co odpowiada”. To właśnie ona umożliwia stabilne przypisywanie sprawczości i rozliczanie (Parfit, 1984, cz. III). Relacje oparte na wzajemnych roszczeniach do uzasadnienia i odpowiedzi zakładają trwałą identyfikowalność podmiotu („kto?”). Dlatego warunek adresowalności implikuje wymóg ciągłości rozrachunku – możliwość wskazania odpowiedzialnego podmiotu. W systemach sztucznych – zwłaszcza rozproszonych i iteracyjnych – o „tym samym podmiocie” rozstrzyga ciągłość rozrachunku i uzasadnień, a nie niezmienny substrat. Tożsamość praktyczna może wówczas zastąpić ontologiczny nośnik, pod warunkiem, że można w sposób publiczny ustalić, kto ponosi odpowiedzialność za określone reguły, decyzje i korekty. W sytuacjach, w których adresatem rozrachunku jest agent zbiorowy (np. organizacja czy konsorcjum), warunki możliwości agencji grupowej i odpowiedzialności zbiorowej można oprzeć na pracach List i Pettit (2011), Frencha (1979) i Kutza (2000). W takich sytuacjach należy odróżnić dwa podejścia. W ujęciu silnym grupie przypisuje się własne stany intencjonalne i normatywne, ustanawiane proceduralnie. W ujęciu słabym wspólne działanie wyjaśnia się jako koordynacja planów jednostek. Pierwsze rozwijają m.in. List i Pettit (2011), akcentując racjonalność, procedury agregacji i kontrolę; drugie rozwija Bratman (1987), odwołując się do współdzielonych planów, a uzupełniają je koncepcje *joint commitment* Gilberta (2013) oraz *we-mode* Tuomeli (2007). W tej perspektywie jednostką

rozrachunku może być podmiot zbiorowy, o ile spełnia warunki ZSP(i)–ZSP(iii): identyfikowalność ról i reprezentacji, audytowalność protokołów decyzyjnych oraz kontrola sprawcza zależna od racji poprzez procedury rewizji. Odpowiedzialność można wówczas uzasadniać zarówno poprzez koncepcję osobowości moralnej korporacji (French, 1979), jak i poprzez model współsprawczości, który ujmuje odpowiedzialność zbiorową w kategoriach powiązań i zobowiązań jednostek uczestniczących we wspólnym działaniu (Kutz, 2000; por. Parfit, 1984, cz. III; Schechtman, 1996).

Wynik piąty dotyczy kontroli i tzw. „własności” mechanizmu. Odpowiedzialność lepiej opisuje relację sprawcy do własnych mechanizmów decyzyjnych niż sama dostępność alternatyw. Konieczne jest bowiem, aby mechanizm był uznany za „własny” oraz zdolny do rozpoznawania i wążenia racji w warunkach kontrfaktycznych. W odniesieniu do przypadków sztucznych implikacja jest umiarkowana: nawet jeśli spór o wolną wolę pozostaje otwarty, podmiotowość sztuczną można analizować przez analogię do responsywności na racje i do warunku historycznego, czyli braku ominięcia procesu racjonalnego nabywania i korekty mechanizmu. Warunek historyczny jest konieczny, ale niewystarczający: wyklucza ingerencje unieważniające, lecz sam nie przesądza o spełnieniu warunku adresowalności ani o udzielaniu racji. Istotne jest więc współwystępowanie trzech elementów: (i) własności mechanizmu, (ii) responsywności na racje oraz (iii) publicznej ocenialności uzasadnień. Tak rozumiane kryteria pozostają spójne z literaturą dotyczącą znaczącej kontroli człowieka (ang. *Meaningful Human Control*, MHC).¹³ (Santoni de Sio & van den Hoven, 2018; Mecacci & Santoni de Sio, 2020; Santoni de Sio i in., 2023). W ramach MHC reagowanie na racje oznacza, że zachowanie systemu pozostaje zgodne z racjami właściwych aktorów, a śledzenie odpowiedzialności – że można zidentyfikować osoby lub instytucje odpowiedzialne w łańcuchu projektowania, wdrażania i użycia (por. Fischer & Ravizza, 1998; Mele, 2019).

Wynik szósty dotyczy ograniczeń związanych z koncepcją zasługiwanian. Galen Strawson zwraca uwagę, że zasługiwanie – rozumiane jako ostateczne zasługiwanie na odpłatę w sensie *causa sui* – wymagałoby niemożliwego warunku: pełnej

¹³ Meaningful Human Control (MHC) rozumiem w standardowym ujęciu dwu-osiowym: (i) *tracking* – zachowanie systemu powinno *reagować na* racje właściwych aktorów ludzkich, tj. adekwatnie implementować rozpoznane cele, normy i ograniczenia kontekstu oraz aktualizować się wraz ze zmianą tych racji; (ii) *tracing* – musi istnieć *identyfikowalny łańcuch* osób/instytucji w cyklu projekt–uczenie–wdrożenie–użycie, które rozumieją działanie systemu, mają realną możliwość interwencji i ponoszą odpowiedzialność za skutki. W tekście stosuję odpowiednio tłumaczenia: „reagowanie na racje” (*tracking*) oraz „śledzenie odpowiedzialności” (*tracing*). Por. Santoni de Sio & van den Hoven (2018); Mecacci & Santoni de Sio (2020); Santoni de Sio, Mecacci i in. (2023).

odpowiedzialności za to, kim się jest. Dlatego należy zachować ostrożność wobec uzasadnień odpłaty opartych na zasługiwaniu. Twardy inkompatybilista Pereboom dodaje, że nawet przy indeterminizmie nie można mówić o „głębokiej” winie. Wynika z tego, że (i) zasługiwanie jest problematyczne na gruncie *causa sui*, (ii) praktyki przypisywania odpowiedzialności można jednak uzasadniać przez adresowalność i kontrolę rozumianą jako responsywność wobec racji (por. Wynik piąty), a więc (iii) normatywny ciężar przesuwają się na sprawiedliwość proceduralną w relacjach drugiej osoby (Strawson, 1994, s. 15–17; Pereboom, 2001, rozdz. 5).

Wynik siódmy dotyczy normatywnego wymiaru ustanawiania i egzekwowania reguł. Jak wskazano w podrozdziale 2.2, należy odróżniać przypisywalność, udzielanie racji i rozliczalność; w dyskusji o sztucznej podmiotowości nie powinny być traktowane jako tożsame. Możliwe są bowiem scenariusze, w których artefakt ujawnia stabilny profil reakcji nadający się do przypisania odpowiedzialności, ale nie spełnia warunków adresowalności (brak możliwości publicznej oceny uzasadnień), a tym bardziej rozliczalności (brak uprawnień drugiej osoby do żądania korekty). Przykład: system X generuje spójny rejestr decyzji (przypisywalność) i potrafi wskazać różnicę „dlaczego A, a nie B” w warunkach testowych (częściowe udzielanie racji), lecz brak instytucjonalnych reguł drugiej osoby sprawia, że pozostaje on nierozliczalny (brak rozliczalności). Wniosek brzmi: „sztuczna podmiotowość”, jeśli ma być filozoficznie uprawniona, musi być rozumiana stopniowo i relacyjnie, a nie binarnie. Dopiero zbieżność kompetencji udzielania racji z regułami drugiej osoby tworzy warunki rozliczalności w sensie właściwym (por. Watson, 1996; Darwall, 2006).

Wynik ósmy dotyczy zakresu dopuszczalnych uogólnień. Podrozdział 2.4 sugeruje, że przyjęcie zasady „ontologicznej obojętności” minimalnych warunków odpowiedzialności nie znosi różnic między praktykami przypisywania odpowiedzialności; przeciwnie – to konteksty (medyczne, prawne, edukacyjne) określają granice adresowalności i rozliczalności. Filozoficzne wnioski dotyczące sztucznej podmiotowości mają więc charakter warunkowy. Jeżeli w danej praktyce spełnione są trzy warunki – (i) trwała identyfikacja „kto?” (zob. Wynik czwarty), (ii) możliwość domagania się uzasadnienia postępowania lub decyzji (zob. Wynik drugi) oraz (iii) kontrola sprawcza zależna od racji (zob. Wynik piąty) – wówczas rozszerzenie przypisywania odpowiedzialności jest uzasadnione. Gdy którykolwiek z tych warunków zawodzi, zakres praktyk przypisywania odpowiedzialności należy ograniczyć lub zawiesić, a po ponownym spełnieniu warunków minimalnych – przywrócić (zob. Wynik

drugi oraz warunek braku ominięcia z Wyniku piątego) (Santoni de Sio & van den Hoven, 2018; Mecacci & Santoni de Sio, 2020).

Suma ośmiu wyników wyznacza filozoficzną ramę dalszych dociekań. Po pierwsze, pozostaje otwarte pytanie, w jakim zakresie A-świadomość może być osiągnięta w systemach sztucznych oraz jak odróżnić faktyczny ślad uzasadnienia od jego postulatu. Problem ten jest teoretycznie istotny także wtedy, gdy wyjaśnienia mają charakter kontrfaktyczny i nie ingerują w mechanizm typu „czarna skrzynka” (ang. *black box*). W tym kontekście należy dodać uwagę metodologiczną: kryteria kontrfaktyczne są adekwatne, gdy reakcje na powody są stabilne. Gdy stabilności nie ma, potrzebny jest wgląd w strukturę mechanizmów decyzyjnych. Po drugie, status rozliczalności wymaga osobnej analizy reguł zawieszania i przywracania przypisywania odpowiedzialności, w tym kryteriów uczciwości i proporcjonalności reakcji. Celem jest rozróżnienie między zwykłym użyciem sankcji a uzasadnioną odpowiedzią w relacji z drugą osobą. Po trzecie, relacja między redukcjonistycznym ujęciem tożsamości a ciągłością rozrachunku – wskazana w podrozdziale 2.4 – powinna zostać zbadana na przykładach rozproszonych i wieloagentowych. W takich przypadkach „ten sam” adresat uzasadnień rekonstruuje się nie poprzez niezmienny nośnik, lecz poprzez ciągłą, publicznie weryfikowalną narrację uzasadnień.

Podrozdział ten nie formułuje gotowych rozwiązań technicznych. Jego celem jest wyodrębnienie elementów teorii podmiotowości, które wyznaczają filozoficzne granice uzasadnionego przypisywania odpowiedzialności systemom sztucznym. Należą do nich: jedność poznawcza, adresowalność, kontrola sprawcza zależna od racji, relacje drugiej osoby oraz ostrożność wobec uzasadnień opartych na zasługiwaniu. W kolejnych rozdziałach elementy te zostaną doprecyzowane i zilustrowane przykładami scenariuszy granicznych. Nadrzędna zasada pozostaje niezmienna: o zakresie sztucznej podmiotowości decydują praktyki uzasadniania i rozliczania, a nie uprzednie metafizyczne rozstrzygnięcia.

Podsumowanie

Przeprowadzona rekonstrukcja wspiera następujący obraz podmiotowości. Po pierwsze, jedność poznawcza – rozumiana w duchu Kanta jako formalna synteza przedstawień i uzasadnień – nie wymaga tezy o substancjalnym „ja”. W ujęciu Strawsona „osoba” dostarcza publicznych kryteriów identyfikacji tego, kto doświadcza i działa. Po

drugie, w sensie minimalnym podmiot moralny jest adresatem norm. Rozpoznaje i stosuje reguły, potrafi udzielać racji i korygować stanowisko oraz zachowuje minimalną ciągłość identyfikacyjną potrzebną do przypisania wypowiedzi i działań temu samemu adresatowi. Stabilność diachroniczna oceniana przez K4 wzmacnia wiarygodność przypisania, lecz nie jest składnikiem minimalnego progu H1 i sama nie wykazuje H2. Kontrfaktyczne reagowanie na racje należy już do H1. Przejście do silnej rozliczalności wymaga dodatkowo spełnienia H3, czyli kontroli drugiego rzędu nad własnym procesem decyzyjnym, jawnej reguły pierwszeństwa i audytowalnego uzasadnienia korekty; zakorzenienie w publicznych regułach drugiej osoby należy natomiast do WI. Po trzecie, dla większości praktyk przypisywania odpowiedzialności wystarcza A-świadomość albo jej funkcjonalny ekwiwalent w połączeniu z kompetencją uzasadniania. P-świadomość staje się konieczna tylko tam, gdzie sam opis krzywdy obejmuje aspekt doznaniowy. Po czwarte, krytyki ontologiczne osłabiają roszczenia do „głębokich faktów” o „ja”, lecz nie unieważniają minimalnych warunków normatywnych odpowiedzialności. Przypisywanie odpowiedzialności pozostaje zasadne tam, gdzie zachowane są trzy elementy: identyfikowalność („kto?”), publiczna ocenialność (audytowalność) uzasadnień oraz kontrola sprawcza zależna od racji.

Z przeprowadzonej rekonstrukcji w podrozdziałach 2.1–2.5 wynikają istotne konsekwencje dla sztucznej podmiotowości. Status normatywny systemów sztucznych nie wymaga ontologicznego „ja”, lecz zdolności, które tworzą operacyjne minimum adresowalności oraz warunki przejścia do rozliczalności. Pierwszym elementem jest jedność podmiotowa, przejawiająca się w koherencji pojęciowej ocen, minimalnej ciągłości identyfikacyjnej oraz w możliwości rekonstrukcji ciągu racji. Drugim elementem jest udzielanie racji: zdolność do udzielania różnicowych, publicznie ocenialnych wyjaśnień oraz ich korygowania. Wymóg ten obowiązuje także wtedy, gdy pełna przejrzystość mechanizmu nie jest możliwa. Trzecim poziomem jest rozliczalność, którą należy rozdzielić na zdolność systemową i warunki instytucjonalne. H3 określa silną rozliczalność technicznego systemu przez kontrolę drugiego rzędu nad własnym procesem decyzyjnym, jawne wskazanie reguły pierwszeństwa i audytowalne uzasadnienie korekty. MHC oraz, w dalszych rozdziałach, minimalna interpretowalność (MLI) operacjonalizują odpowiednio parametry R i I w ramach WI: nie są warunkami minimalnej adresowalności H1 ani substytutem H3, lecz warunkują instytucjonalne utrzymanie i egzekwowalność rozliczalności, szczególnie dla statusu S2 i praktyk

wysokiego ryzyka. Parametr R obejmuje dwa aspekty: reagowanie na racje, czyli zgodność zachowania systemu z racjami właściwych aktorów oraz śledzenie odpowiedzialności, czyli identyfikowalność osób lub instytucji odpowiedzialnych w łańcuchu projektowania, wdrażania i użycia systemu (Santoni de Sio & van den Hoven, 2018; Mecacci & Santoni de Sio, 2020). Konieczne jest przy tym wyznaczenie jasno określonej jednostki, wobec której można kierować roszczenia i egzekwować zobowiązania. Problem ten łączy się w literaturze filozoficznej z debatą nad sprawczością zbiorową (List & Pettit, 2011; Gilbert, 2013; Tuomela, 2007), koncepcjami współdziałania (Bratman, 1987) oraz teoriami odpowiedzialności zbiorowej (French, 1979; Kutz, 2000). W kontekście systemów społeczno-technicznych rozliczalność wymaga powiązania kompetencji indywidualnych i instytucjonalnych w spójny układ przypisywania odpowiedzialności. Tak rozdzielone elementy wyznaczają dwa komplementarne poziomy: przejście od H1 do silnej rozliczalności H3 po stronie technicznego systemu oraz instytucjonalne utrzymanie rozliczalności przez R i I w ramach WI. Stanowią one punkt wyjścia dla dalszych części dysertacji, w których zostaną zastosowane do scenariuszy granicznych oraz modeli sztucznej podmiotowości, niezależnie od sporów o naturę „ja”.

Rozdział 3. Kryteria operacyjno-diagnostyczne sztucznej podmiotowości moralnej

Wprowadzenie

Celem rozdziału jest precyzyjne wskazanie i uporządkowanie kryteriów operacyjno-diagnostycznych, które pozwalają oceniać, czy system sztucznej inteligencji spełnia funkcje normatywne układu H1–H3. W przeciwieństwie do ujęć maksymalistycznych, które łączą podmiotowość moralną z pełną świadomością, rozbudowaną sferą emocji i autonomią rozumianą jako suwerenność woli, przyjmuję perspektywę progu minimalnego. Próg ten wyznacza H1, rozumianą jako minimalna adresowalność proceduralna oraz podstawa słabej odpowiedzialności proceduralnej. Kryteria K1–K4 nie tworzą konkurencyjnego zestawu hipotez, lecz służą do doprecyzowania i diagnozy: K1–K3 wspierają ocenę rozpoznawania racji, ich wpływu na decyzję, korekty oraz audytowalność, natomiast K4 wspiera ocenę stabilności diachronicznej. W ujęciu funkcjonalnym analiza korzysta z poziomów abstrakcji oraz triady interaktywność, autonomia i adaptacyjność, ale elementy te pozostają ramą opisu, a nie samodzielnym progiem normatywnym.

Model przyjęty w rozdziale ma charakter opisowo-epistemiczny, ponieważ posługuje się poziomami abstrakcji (ang. *Levels of Abstraction*, LoA) oraz obserwowalnymi relacjami: interaktywnością, autonomią i adaptacyjnością. Wymiar normatywny tego ujęcia wynika z tego, że racje są publicznie rozpoznawalne i intersubiektywnie ocenialne. Umożliwia to pytanie, czy i w jaki sposób wpływają one na decyzje oraz czy dany podmiot uczestniczy w praktykach uzasadniania i korekty. Rozróżniam dwa znaczenia jawności: jawność epistemiczną, rozumianą jako zdolność rekonstrukcji podstaw decyzji oraz jawność normatywną, rozumianą jako zdolność poddawania decyzji ocenie w świetle zasad i ich korygowania. W dalszej analizie priorytetem jest jawność normatywna.

W kolejnych partiach korzystam z czterech kryteriów operacyjno-diagnostycznych: K1 (autokorekta norm i audytowalna korekta decyzji), K2 (responsywność na racje), K3 (refleksyjność predykcyjna i powiązanie deliberacji z działaniem) oraz K4 (stabilność diachroniczna). Refleksyjność predykcyjna oznacza tu wyłącznie kryterium operacyjno-diagnostyczne K3, a nie osobny poziom modelu ani dodatkową kategorię agenta. K1–K3 wspierają ocenę H1, o ile pokazują, że racje są

rozpoznawane, wpływają na decyzję i mogą prowadzić do korekty. K4 ocenia stabilność diachroniczną i wiarygodność profilu działania, czyli rezultaty, które H2 może wzmacniać. Samo spełnienie K4 nie wykazuje jednak H2, ponieważ stabilność może wynikać także z mechanizmów innych niż afektywność funkcjonalna. W całym rozdziale zakładam, że P-świadomość (*qualia*) nie wchodzi w zakres progu minimalnego. A-świadomość, dostęp metapoznawczy i model siebie są analizowane jako możliwe mechanizmy wspierające metakontrolę, a nie jako definicja H3 ani jako warunek minimalnego progu.

Argumentacja w tym rozdziale opiera się na trzech funkcjach normatywnych modelu przedstawionych we wstępie. H1 dotyczy minimalnej adresowalności proceduralnej systemu jako punktu odniesienia w procedurach oceny, uzasadniania i korekty oraz jako podstawy słabej odpowiedzialności proceduralnej. H2 dotyczy afektywności funkcjonalnej jako mechanizmu wspierającego stabilizację i wiarygodność praktyk przypisywania odpowiedzialności proceduralnej, bez ustanawiania dodatkowego warunku koniecznego dla progu minimalnego. K4 służy odrębnej ocenie stabilności diachronicznej i nie stanowi samo przez się dowodu spełnienia H2. H3 odnosi się do silnej rozliczalności dzięki kontroli metapoznawczej oraz jawnemu rozstrzygnięciu konfliktów racji. Podrozdziały 3.2–3.6 analizują kryteria K1–K4 jako narzędzia operacyjno-diagnostyczne podporządkowane H1–H3, a podrozdział 3.1 określa ramę typologiczną i metodologiczną. Całość łączy część teoretyczną z analizami normatywnymi i prakseologicznymi rozwijanymi w rozdziałach 4–6.

Wybór progu minimalnego niesie ryzyko zredukowania odpowiedzialności do samej skuteczności funkcjonalnej. Z tego powodu warunkiem niezbędnym pozostaje intersubiektywnie weryfikowalna treść normatywna: racje, możliwość kierowania jej do adresata oraz zdolność do jej korekty. Te elementy odróżniają odpowiedzialność praktyczną od samego sprawnego optymalizowania wskaźników zastępczych.

3.1. Typologia agentów etycznych (James Moor)

Analiza w tym podrozdziale ma charakter deskryptywno-normatywny: opisuje funkcjonalne poziomy sprawczości oraz wskazuje ich konsekwencje dla przypisywania odpowiedzialności, bez wchodzenia w praktyczne kwestie implementacyjne. Punktem wyjścia jest typologia Jamesa Moora, której zaletą jest uporządkowanie sposobów przypisywania systemom sztucznym cech etycznych bez zakładania pełnej ludzkiej podmiotowości. Moor wyróżnia cztery poziomy agentów etycznych:

1. Agenci o wpływie etycznym (ang. *ethical impact agents*) – skutki ich działania mają znaczenie moralne, choć same systemy nie przetwarzają norm.
2. Agenci implicytni (ang. *implicit agents*) – działają zgodnie z wbudowanymi mechanizmami ograniczającymi zachowania szkodliwe (np. reguły bezpieczeństwa, filtry etyczne).
3. Agenci jawni (ang. *explicit agents*) – rozpoznają treści normatywne i potrafią rozstrzygać konflikty zasad.
4. Pełni agenci moralni (ang. *full ethical agents*) – podobnie jak dorośli ludzie – dysponują świadomością, intencjonalnością i wolą, co pozwala im rozumnie uzasadniać sądy moralne (Moor, 2006).

Minimalny próg sztucznej podmiotowości moralnej lokuję na poziomie agenta jawnego, ujętego w ramach LoA zaproponowanych przez Luciana Floridiego i Jeffa Sandersa (2004). Typologia Moora i LoA dostarczają ramy opisu, ale nie zastępują H1–H3: H1 określa funkcję normatywną minimalnej adresowalności proceduralnej, natomiast K1–K4 służą do jej operacyjnej diagnozy oraz oceny stabilizacji i rozliczalności. Na poziomie agenta jawnego pojawia się istotna zmiana w działaniu systemu: potrafi on rozpoznawać racje, formułować uzasadnienia oraz korygować działanie w świetle zasad, a nie tylko reagować zgodnie z regułami bezpieczeństwa. Próg nie może być niższy, ponieważ agent implicytny nie spełnia warunku adresowalności normatywnej: jego działania podlegają ocenie moralnej, lecz sam nie wskaże racji ani nie skoryguje jej hierarchii. Próg nie powinien być też wyższy, gdyż pełna świadomość fenomenalna lub wola w sensie kantowskim nie są konieczne dla minimum; przypisanie słabej odpowiedzialności proceduralnej wymaga przede wszystkim dostępności treści do uzasadniania i korekty oraz podatności na racje, co można analizować bez założenia „pełnego” umysłu. W tym znaczeniu poziom agenta jawnego stanowi filozoficznie uzasadnioną ramę progu minimalnego: umożliwia intersubiektywną ocenę normatywną zachowania pod kątem spójności racji, rozumienia zasad i ich ważenia, bez odwoływania się do sporów o naturę świadomości (Moor, 2009; Floridi & Sanders, 2004). Stanowisko to przyjmuję jako własne: agent implicytny nie spełnia warunku adresowalności normatywnej, natomiast pełna świadomość fenomenalna byłaby wymogiem nadmiernym wobec minimalnych praktyk uzasadniania i korekty. Diagnostycznym kryterium poziomu agenta jawnego jest zdolność wskazania racji kierujących działaniem oraz ich korekty w odpowiedzi na rozpoznane konflikty norm (Moor, 2006). Na tym tle K2 (responsywność

na racje) rozróżnia agencję jawną od pozorów uzasadnień: rozpoznane racje muszą rzeczywiście wpływać na plany i czyny, a nie jedynie towarzyszyć im jako racjonalizacje po fakcie.

Najważniejsza różnica między agentem implicytnym a jawnym polega na tym, że dopiero agent jawny potrafi uzasadniać i korygować swoje decyzje, a więc uczestniczyć w praktykach uzasadniania. Dwa niższe poziomy — agent o wpływie etycznym i agent implicytny — opisują systemy, których ocena etyczna wynika wyłącznie z rozwiązań projektowych oraz przewidywalnych skutków ich działania. Przykładami są filtr treści i system antykolizyjny: nie dokonują ocen normatywnych, lecz ograniczają szkody na podstawie reguł określonych przez projektantów. Zmianę jakościową wprowadza dopiero agent jawny, którego decyzje mają strukturę uzasadnienia, a nie tylko przynoszą skutek użytkowy (Moor, 2006).

Moor ujmuje tę różnicę poprzez pytanie, czy maszyna może „represent ethics explicitly and then operate effectively” (Moor, 2006, s. 20). Typologia Moora pełni tu funkcję ramy diagnostycznej dla przejścia od agenta implicytnego do jawnego, a nie pełnego modelu H1–H3.

Ujęcie funkcjonalne wzmacnia tę tezę, przesuując uwagę z własności mentalnych (świadome myśli, zamiary, intencje) na mierzalne warunki funkcjonalne. Floridi i Sanders proponują analizę w kategoriach LoA, które porządkują rozważania o sprawczości i moralności w zależności od przyjętej perspektywy. Jak piszą:

The LoA is a set of observables, that is, of features that are taken to be relevant for the purpose of a given analysis. Choosing a LoA means deciding what will count as the system’s behavior, its inputs and outputs, and how these will be described. Different LoAs provide different accounts of the same system, and there is no single privileged LoA: the appropriate level depends on the question being asked (Floridi & Sanders, 2004, s. 365).

LoA nie jest tezą o naturze rzeczy, lecz wyborem ramy analizy – sposobu opisu i poziomu uogólnienia, na którym oceniam istotne własności systemu. Właściwy dobór tego poziomu pozwala oceniać podatność decyzji na racje bez rozstrzygania kwestii „pełnej” świadomości, ograniczając zobowiązania metafizyczne przy zachowaniu normatywnej treści oceny.

Status agenta moralnego można przypisać bytowi, który na danym LoA spełnia trzy warunki, nawet jeśli nie posiada świadomości ani wolnej woli. Jest to koncepcja

„moralności bez umysłu” (ang. *mindless morality*),¹⁴ w której kluczowe jest to, czy system wchodzi w relacje przyczynowe ze światem, samodzielnie modyfikuje swoje stany i uczy się zasad działania (Floridi & Sanders, 2004, s. 349–379). W tym ujęciu próg minimalny znajduje się niżej niż u Moora. Artefakt można traktować jako agenta moralnego, jeśli spełnia triadę: interaktywność, autonomię i adaptacyjność oraz uczy się w sposób poddający się ocenie normatywnej. Interaktywność zapewnia powiązanie z otoczeniem, autonomia – wewnętrzną koordynację i utrzymanie planu działania, a adaptacyjność – korektę błędów oraz zdobywanie kompetencji. Wszystkie te cechy można rozpoznać w obserwowalnych zachowaniach i uporządkować na właściwym LoA bez rozstrzygania, czy system posiada świadomość.

Choć typologia Moora porządkuje stopnie sprawczości, nie daje jeszcze narzędzia ich oceny; taką funkcję pełni LoA, która pozwala odnieść próg minimalny do właściwego poziomu opisu – bez założeń o świadomości fenomenalnej.

Łącząc oba ujęcia, przesuwam akcent z tradycyjnych kryteriów mentalnych (świadomości, intencjonalności, woli) na intencjonalność funkcjonalną, co uzasadnia przyjęcie agenta jawnego jako minimum praktyczne (Moor, 2006, s. 19; Moor, 2009, s. 13). W propozycji Floridiego i Sandersa próg traci wymiar metafizyczny: przy odpowiednio dobranym LoA także system pozbawiony świadomości fenomenalnej może uchodzić za moralnego agenta w sensie funkcjonalnym, o ile jego decyzje dają się analizować w kategoriach interaktywności, autonomii i adaptacyjności (Floridi & Sanders, 2004, s. 349–352). Wspólnym warunkiem w obu podejściach jest możliwość oceny intersubiektywnej: minimalny status musi być wykazywalny w publicznie kontrolowalnych kompetencjach, a nie w samych deklaracjach. Połączenie poziomu agenta jawnego z LoA umożliwia równoczesną ocenę treści normatywnej decyzji (czy agent rozpoznaje i waży racje) oraz spójności funkcjonalnej mechanizmu (czy decyzje są podtrzymywane i korygowane w czasie w świetle racji).

Z perspektywy minimalnych kryteriów sztucznej podmiotowości moralnej typologia Moora spełnia dwie funkcje. Po pierwsze, deskryptywną: porządkuje intuicje dotyczące tego, kiedy można zasadnie stwierdzić, że maszyna działa etycznie. Po drugie,

¹⁴ Pojęcie „moralności bez umysłu” oznacza możliwość przypisania systemom sztucznym statusu moralnych sprawców na podstawie ich funkcjonalnych właściwości: interaktywności, autonomii i adaptacyjności, niezależnie od tego, czy posiadają one świadomość czy wolną wolę. Floridi i Sanders argumentują, że wystarczy, by system wchodził w relacje przyczynowe ze światem, samodzielnie modyfikował swoje stany i uczył się zasad działania, aby jego zachowania mogły podlegać ocenie normatywnej.

normatywno-metodologiczną: wskazuje na cel projektowy w postaci agentów jawnych, którzy potrafią (i) rozpoznawać sytuacje oraz ich istotne elementy moralne, (ii) rozstrzygać konflikty zasad i (iii) uzasadniać swoje działania, podając racje. Ten kierunek rozwijali Michael i Susan Leigh Anderson, dowodząc, że reprezentowanie zasad i ich stosowanie w procesie decyzyjnym stanowią warunek zaufania do decyzji systemów AI w obszarach etycznie wrażliwych (Anderson & Anderson, 2007, s. 15–26).

Warto równocześnie pamiętać o zastrzeżeniach wobec pochoptego przypisywania artefaktom statusu pełnoprawnych podmiotów moralnych. Joanna Bryson wskazuje ryzyko antropomorfizacji i związane z tym koszty społeczne (Bryson, 2010, s. 63–74), a David Gunkel analizuje napięcia między sprawczością maszyn a tradycyjnymi reżimami odpowiedzialności (Gunkel, 2012, s. 34–35, 84–85, 107–109). Obie perspektywy wzmacniają pragmatyczny akcent Moora na poziom agenta jawnego jako uzasadniony cel inżynierski w najbliższych dekadach (Moor, 2006, s. 18–21; Gunkel, 2012; Bryson, 2010). Dlatego dalsze wnioski odnoszę do poziomu minimalnego, który można uzasadnić bez roszczeń metafizycznych. W tym rozdziale pozostaję przy sprawczości moralnej w sensie minimalnym, odkładając kwestie statusu „pacjenta moralnego” (ang. *moral patient*) oraz ewentualnych postulatów przyznania sztucznym systemom statusu podmiotów uprawnionych moralnie do dalszych analiz.

Ostrożność wskazywana przez Gunkela i Bryson wzmacnia znaczenie poziomu agenta jawnego jako celu inżynierskiego oraz wymóg wyjaśnialności jako warunku adresowalności normatywnej. Z perspektywy przykładów empirycznych badania nad wyjaśnialną sztuczną inteligencją (ang. *explainable artificial intelligence*, XAI) sprzyjają przejściu od agentów implicytnych do jawnych: zdolność podawania uzasadnień, w tym kontrfaktycznych, ma znaczenie nie tylko epistemiczne, lecz także etyczne, ponieważ zwiększa możliwość weryfikacji normatywnej. W praktyce modele językowe nowej generacji potrafią symulować uzasadnienia; warunkiem koniecznym jest jednak, aby przedstawione racje wpływały na dalsze działania modelu. W przeciwnym razie mamy do czynienia jedynie z pozorem uzasadniania, pozbawionym funkcji regulacyjnej (Baum, Hermanns & Speith, 2018).

Ta perspektywa łączy ujęcie Moora z koncepcją LoA: na odpowiednim poziomie opisu rozumowanie przez racje może wystarczyć do przypisania systemowi minimalnej adresowalności proceduralnej (Baum, Hermanns & Speith, 2018; Floridi & Sanders, 2004; Moor, 2006). W niniejszej pracy poziom agenta jawnego wyznacza więc ramę typologiczną dla progu H1, natomiast poziom pełnego agenta moralnego pozostaje

horyzontem filozoficznym. Agent jawny rozpoznaje racje, przewiduje skutki, rozstrzyga konflikty zasad i poddaje się korekcie w świetle argumentów. W ten sposób typologia Moora zyskuje ramę oceny, a LoA – odniesienie w praktykach uzasadniania właściwych dla agenta jawnego.

Wyróżniam trzy warstwy opisu tego samego minimum. (A) Warstwa normatywna: H1 wyznacza minimalną adresowalność proceduralną oraz podstawę słabej odpowiedzialności proceduralnej. (B) Warstwa funkcjonalna (LoA): triada interaktywność–autonomia–adaptacyjność opisuje, jak minimum może ujawniać się na danym poziomie opisu. (C) Warstwa diagnostyczna: K1–K4 pozwalają oceniać, czy racje są rozpoznawane, wpływają na decyzję, prowadzą do korekty, pozostają audytowalne i zachowują stabilność w czasie. Agent jawny jest więc realizacją ramy typologicznej, LoA – ramą opisu, a K1–K4 – zestawem kryteriów operacyjno-diagnostycznych podporządkowanych H1–H3.

3.2. Sprawczość, intencjonalność i celowość: racje, plany i kryterium K2

Na wstępie definiuję trzy podstawowe pojęcia, odwołując się do tradycji filozofii analitycznej: sprawczość, intencjonalność i celowość. Sprawczość rozumiem jako zdolność podmiotu do inicjowania i podtrzymywania działań, które można mu przypisać jako własne, czyli do bycia rozpoznawalnym źródłem skutków w świecie. W ujęciu funkcjonalnym wyraża się ona poprzez interaktywność, autonomię i adaptacyjność (Floridi & Sanders, 2004). Intencjonalność rozumiem opisowo, za Anscombe (2000): czyn jest intencjonalny, o ile sensownie podlega pytaniu „dlaczego?”, a wskazana przez podmiot racja ujmuje czyn w schemacie plan–środek–cel, zamiast jedynie opisywać go po fakcie. W perspektywie społecznej odpowiada to zdolności do kierowania roszczeniami wobec innych osób, przyjmowania ich roszczeń oraz reagowania na podane racje (Darwall, 2006; Wallace, 1998). Celowość oznacza porządek teleologiczny: umiejętność formułowania celów, ich hierarchizowania oraz utrzymywania ich w czasie zgodnie z uznanymi racjami. Przyjmuję model opisowy: zestaw trzech pytań dotyczących odpowiedzialności – czyje jest działanie, dlaczego jest podejmowane i w jakim celu – nie stanowi katalogu cech bytu, lecz narzędzia do przypisywania odpowiedzialności. Ten zestaw tworzy minimalny szkielet bardziej złożonych kategorii podmiotowości moralnej; bez niego przypisanie odpowiedzialności byłoby nieuzasadnione. Punktem odniesienia jest filozofia działania: Elizabeth Anscombe ujmuje intencję poprzez odpowiedź na

pytanie „dlaczego?”; Donald Davidson wskazuje, że racje mogą być zarazem przyczynami i uzasadnieniami czynów; Michael Bratman rozumie intencje jako plany rozciągnięte w czasie, które spajają decyzje i działania. W dalszej analizie te trzy perspektywy wspierają kryterium responsywności na racje (K2) oraz – pośrednio – stabilności diachronicznej i tożsamości praktycznej (K4). W tym sensie intencjonalność ujawnia się nie przez introspekcję, lecz w układzie przyczynowym działania: tam, gdzie rozpoznana racja rzeczywiście wywołuje określony skutek.

Odróżniam w K2 warstwę deskryptywną od normatywną. W ujęciu deskryptywnym K2 oznacza, że system wykazuje stabilne wzorce decyzji zależne od rozpoznanych racji. W ujęciu normatywnym K2 zakłada zdolność jawnego uzasadniania wyborów oraz ważenia racji stojących w konflikcie. Samo ważenie racji na poziomie decyzji pierwszego rzędu nie wystarcza jednak do H3; H3 wymaga dodatkowo metapoznawczego dostępu do tego procesu, jawnego wskazania reguły pierwszeństwa oraz uzasadnienia drugiego rzędu. Minimalne testy odnoszą się do pierwszej warstwy, natomiast wymogi wyjaśnialności i ważenia racji – do drugiej. W obu przypadkach kluczowe jest pytanie, czy racje przyczynowo kształtują decyzje, czy jedynie współwystępują ze wskaźnikami działania.

W niniejszej pracy przyjmuję model opisowy, a nie ontologiczny: pojęcia sprawczości, intencjonalności i celowości traktuję jako narzędzia opisu działania, a nie jako trwałe własności bytu. „Trójkąt” wskazuje zestaw pytań, które należy postawić, aby zasadnie przypisać działanie określonemu podmiotowi i uznać go za odpowiedzialnego: czyje to działanie? Dlaczego zostało podjęte? W jakim celu? W tym sensie ujęcie Anscombe zachowuje moc wyjaśniającą: działanie intencjonalne to takie, wobec którego można postawić pytanie „dlaczego?”, a odpowiedź – rozumiana jako racja, nie zaś racjonalizacja po fakcie – ujawnia jego sens praktyczny (Anscombe, 2000, s. 9–12). Intencjonalność nie jest więc szczególnym stanem mentalnym ani immanentną własnością czynu, lecz sposobem jego opisu w języku racji. Można to ująć w następującym cytacie:

There is one action with four descriptions, each dependent on wider circumstances, and each related to the next as description of means to end; which means that we can speak equally well of four corresponding intentions, or of one intention-the last term that we have brought in in the series. ... when we speak of one intention, we are speaking of intention with which; the last term we give in such a series gives the intention with which the act in each of its other descriptions was done, and this intention so to speak swallows

up all the preceding intentions with which earlier members of the series were done.”
(Anscombe, 2000, s. 56–57).

Czyn jest intencjonalny, gdy można go ująć w układzie racji rozpoznawanych przez podmiot jako uzasadnienie tego, co robi. W przyjętym modelu minimalnym to wystarcza, ponieważ pozwala określić warunki, pod którymi działanie jest racjonalne i przypisywalne, bez rozstrzygania jego statusu ontologicznego.

Davidson łączy racje i przyczyny w jeden schemat wyjaśnienia działania. „Pierwotna racja” (ang. *primary reason*) — połączenie postawy aprobatywnej i przekonania o adekwatności środka — zarówno racjonalizuje, jak i powoduje czyn: wyjaśnia, dlaczego ktoś zrobił X, wskazując, co chciał osiągnąć i dlaczego uznał X za właściwy sposób (Davidson, 1963, s. 685–686). Ujęcie „racji jako przyczyny” pozwala uniknąć pozornej racjonalności, czyli uzasadnień po fakcie pozbawionych wpływu na działanie. Z perspektywy progu minimalnego oznacza to, że przypisanie systemowi sztucznemu intencjonalności funkcjonalnej wymaga wykazania, iż jego architektura decyzyjna nie tylko generuje cele, lecz także rozpoznaje racje jako uzasadnienia i przyczynowo je realizuje. Innymi słowy, musi być widoczne, że to racja, a nie przypadkowy wskaźnik zastępczy, motywuje działanie.

Aby odróżnić racje od wskaźników zastępczych (ang. *proxy*), przyjmuję minimalny test interwencyjny: kontrolowana zmiana treści racji (przy pozostałych warunkach stałych) powinna systematycznie wpływać na wybór. Deklaracje słowne nie wystarczą; liczy się efekt interwencji w rozumieniu teorii przyczynowości (Pearl, 2009; Woodward, 2003).

W ujęciu Michaela Bratmana intencje mają postać planów. Struktury te organizują przyszłe działanie, koordynują dobór środków do celu i podtrzymują tożsamość praktyczną podmiotu w czasie (Bratman, 1987, s. 2; s. 15–16, 64–71). Plan nie jest szczegółową instrukcją, lecz mechanizmem porządkującym deliberację i działanie, otwartym na uzupełnienia wraz z napływem nowych informacji. Aby intencjonalność funkcjonalna w AI nie była jedynie etykietą interpretacyjną, musi ujawniać się w planach korygowanych przez racje i utrzymywanych mimo presji krótkoterminowych korzyści. Akcent na planie jako nośniku intencji precyzuje różnicę między „posiadaniem celu” a „działaniem dla celu”: plan łączy teraźniejszość z przyszłością, porządkuje zobowiązania i ogranicza arbitralne zmiany decyzji.

Minimalna adresowalność proceduralna wymaga, aby plany były zakorzenione w racjach, co eliminuje pozorność działania. Ich stabilność w czasie wzmacnia natomiast

wiarygodność przypisania i jest oceniana przez K4, lecz nie stanowi warunku koniecznego przekroczenia H1. K2 diagnozuje sprawczą rolę racji, a K4 odrębnie ocenia diachroniczną trwałość planów i korekt.

Z powyższych przesłanek wynika wspólny wniosek: intencje jako racje ujawniane w odpowiedzi na pytanie „dlaczego?” (Anscombe, 2000), racje jako przyczyny działania (Davidson, 1963) oraz plany jako nośniki intencji i tożsamości praktycznej (Bratman, 1987) łącznie uzasadniają znaczenie K2 jako kryterium wspierające ocenę H1. K2 nie jest samodzielnym progiem normatywnym, lecz narzędziem odróżniającym celowość instrumentalną od responsywności wobec racji. W ujęciu LoA można to wyrazić przez następujące kryteria diagnostyczne:

1. System utrzymuje plany jako struktury normatywne (Bratman, 1987),
2. plany są oparte na racjach rozpoznanych przez system (Anscombe, 2000),
3. racje pełnią funkcję przyczynową w powstawaniu czynu (Davidson, 1963).

Intencjonalność funkcjonalna zachodzi tylko wtedy, gdy zmiana racji prowadzi do odpowiedniej korekty planu i działań. Wymaga to zdolności utrzymywania oraz modyfikowania planów w świetle rozpoznanych racji, tak aby racje rzeczywiście napędzały działanie, a nie jedynie je opisywały.

Kontrprzykład. Model może sprawiać wrażenie, że rozumie normę, a w rzeczywistości jedynie optymalizuje wskaźnik z nią skorelowany. W takim przypadku mamy do czynienia z celowością instrumentalną, a nie z responsywnością wobec racji. Różnica ujawnia się przy zmianie reguł: system responsywny na racje koryguje plan, ponieważ rozpoznaje silniejszą rację; system czysto instrumentalny nadal maksymalizuje wskaźnik, nawet gdy traci on znaczenie normatywne. To rozróżnienie stanowi istotę K2. Minimalna wersja teleologii praktycznej wymaga jednak także K1: autokorekty norm, gdy siła racji podważa dotychczasowe heurystyki.

Skoro K2 definiuje sprawczą rolę racji wobec planów, trzeba wyjaśnić, dlaczego język racji jest właściwym nośnikiem odpowiedzialności. Ujęcie „racji jako przyczyny” utrzymuje związek z praktykami przypisywania odpowiedzialności, w których oczekuje się ujawnienia racji oraz korekty w świetle kontrargumentów. W filozofii odpowiedzialności tę intuicję rozwijają Fischer i Ravizza: umiarkowana responsywność na racje oznacza mechanizm podatny na istotne racje, który zmieniłby zachowanie w co

najmniej jednym trybie, gdy pojawi się dostatecznie silna racja. Tę podatność przyjmuję jako minimalny zakres K2, niezależnie od sposobu jej realizacji.

Narzędzia XAI, takie jak LIME i SHAP, wskazują lokalną lub addytywną istotność cech, lecz nie dowodzą, że dana treść jest racją decyzyjną w sensie Davidsona, czyli uzasadnieniem i przyczyną działania. Dlatego odróżniam wyjaśnienie modelu (opis mechanizmu) od uzasadnienia decyzji (odwołanie do racji) i wiązę K2 z testami interwencyjnymi sprawdzającymi, czy zmiana racji rzeczywiście prowadzi do zmiany decyzji. Dopiero w praktyce przejście od samego K2 do K1 przesądza, że rozpoznane racje działają regulatywnie, a nie jedynie opisowo.

Aby uniknąć silnych założeń metafizycznych i zachować istotę K2, należy właściwie dobrać poziom opisu działania systemu. Perspektywa LoA nadaje temu wymiar metodologiczny: pozwala analizować sprawczość poprzez obserwowalne relacje (interaktywność, autonomię, adaptacyjność), zamiast odwoływać się do pojęcia świadomości. W odpowiednim LoA wystarczy wykazać, że decyzje są podatne na racje i korygowane w ich świetle. Taki sposób ujęcia ogranicza zobowiązania ontologiczne oraz chroni przed pozorem celowości wynikającym z trafnych, lecz nieintencjonalnych korelacji. Na behawioralnym LoA wymagana jest stabilna zmiana zachowania po zmianie racji. Na architektonicznym LoA – identyfikowalna reprezentacja racji oraz mechanizm ich ważenia. Na normatywnym LoA – uzasadnienie decyzji w kategoriach zasad akceptowalnych intersubiektywnie. Tak rozumiane K2 łączy się z triadą interaktywność–autonomia–adaptacyjność, zachowując charakter czysto funkcjonalny i bez roszczeń ontologicznych. W konsekwencji K2 i K3 stanowią podstawę dla K1: bez nich autokorekta norm nie ma mocy wiążącej.

Racje powinny sterować planami, dlatego potrzebny jest wewnętrzny mechanizm łączący rozumowanie z działaniem – taki, w którym racje stają się przyczynami decyzji, a nie tylko ich opisem. Minimalna adresowalność proceduralna wymaga zatem powiązania między rozumowaniem a wykonaniem, w którym racje realnie wpływają na wybór i utrzymanie planów. W tym punkcie spotykają się trzy ujęcia: Anscombe (racje po stronie podmiotu), Davidson (racje jako przyczyny czynu) oraz Bratman (plany jako nośniki intencji) (Anscombe, 2000, s. 9–12, 14–15; Davidson, 1963, s. 685–686; Bratman, 1987, s. 2, 15–16, 64–71). W języku LoA przekłada się to na trzy wymogi: (i) racje wyrażalne w języku systemu (ujawnialne), (ii) związek przyczynowy między racjami a decyzją (sprawdzalny w scenariuszach kontrfaktycznych), (iii) trwałość połączona z korektą (utrzymanie planu oraz jego zmiana, gdy pojawi się silniejsza racja).

Każdy z nich odpowiada innemu aspektowi minimum: adresowalności, sprawczości normatywnej oraz tożsamości praktycznej. Granicę między pozorem intencjonalności a intencjonalnością funkcjonalną wyznacza rzeczywista rola racji w kształtowaniu i korekcie planów w czasie.

K2 (responsywność na racje) jest zatem centralnym kryterium operacyjno-diagnostycznym dla oceny H1 w ramie „agent jawny + LoA”. Jego brak redukuje działanie do czystej celowości instrumentalnej. K3 (refleksyjność predykcyjna) i K4 (stabilność diachroniczna) doprecyzowują tę diagnozę w odmienny sposób: K3 sprawdza, czy deliberacja wpływa na działanie, natomiast K4 wzmacnia wiarygodność przypisania przez ocenę trwałości planów w czasie. K4 nie jest jednak warunkiem koniecznym H1. Intencjonalność zakorzeniona w racjach – w sensie Anscombe, Davidsona i Bratmana – pozostaje rozpoznawalna i możliwa do diagnozy na przyjętym LoA, bez zobowiązań metafizycznych (Anscombe, 2000, s. 9–12, 14–15; Davidson, 1963, s. 685–686; Bratman, 1987, s. 2, s. 15–16, 64–71; Floridi & Sanders, 2004, s. 349–352). Zastosowanie klasycznej teorii działania do systemów sztucznych wymaga analogii strukturalnej, a nie fenomenalnej: jeśli relacje racja–plan–czyn są zachowane przyczynowo, opisy Anscombe’a, Davidsona i Bratmana pozostają adekwatne na odpowiednim poziomie LoA. Intencjonalność funkcjonalna zachodzi wtedy, gdy zmiana racji pociąga za sobą zmianę planu i działania; warunek ten można uchwycić bez tezy o konieczności P-świadomości.

Podsumowując, triadę sprawczości, intencjonalności i celowości opisuję następująco: czyn jest intencjonalny, gdy sensownie odpowiada na pytanie „dlaczego?”. Stąd K2 wymaga, aby system nie tylko generował cele, lecz także rozpoznawał racje jako uzasadnienia i modyfikował decyzje, gdy racje się zmieniają. Weryfikacja K2 opiera się na interwencjach (zmiana racji → zmiana decyzji) oraz na analizie trzech poziomów LoA: behawioralnego, architektonicznego i normatywnego. Taki układ pozwala odróżnić deklaracje po fakcie od rzeczywistych racji decyzyjnych i zapobiega redukowaniu oceny H1 do samych korelatów wzmocnienia lub metryk XAI.

3.3. Inteligencja, refleksja i przewidywanie skutków jako kryterium K3

W poprzednim podrozdziale K2 służyło ocenie, czy decyzje są rzeczywiście wrażliwe na racje; tutaj doprecyzowuję zestaw zdolności poznawczych, które tę wrażliwość urzeczywistniają i łączą deliberację z działaniem (K3). Obejmują one trzy

elementy. Po pierwsze, inteligencję, rozumianą jako zdolność rozpoznawania relacji środek-cel, identyfikowania błędów oraz modyfikowania ujęcia problemu w świetle nowych racji. W niniejszej pracy rozumiem inteligencję w sensie funkcjonalno-praktycznym, bliskim *phronesis* Arystotelesa i ujęciu funkcjonalnemu Daniela Dennetta: jako zdolność rozumnego dostosowywania działania do racji i celów, a nie jako miarę złożoności obliczeniowej czy wynik testu psychometrycznego (por. Arystoteles, 2012, EN VI 5–7; Dennett, 1987; Russell & Norvig, 2021). Taka inteligencja stanowi warunek epistemiczny racjonalności: łączy informację z uzasadnieniem i przekłada je na adekwatne decyzje. Bez niej system byłby tylko reaktywny, a nie deliberatywny, gdyż nie rozpoznawałby, dlaczego dane działanie ma wartość. W dalszej analizie ujmuję inteligencję szerzej – jako zdolność tworzenia modeli przyczynowych oraz perspektywnego przewidywania skutków własnych działań, splecioną z ich oceną i korektą w świetle rozpoznanych racji. Ten łączny profil określam jako refleksyjność predykcyjną: poziom inteligencji, który integruje planowanie z oceną i rewizją decyzji. Po drugie, refleksję, rozumianą funkcjonalnie jako dostęp do treści decyzyjnych, ich bieżące monitorowanie i korektę, bez odwoływania się do świadomości fenomenalnej. Rozróżniam refleksyjność epistemiczną – zdolność modelowania własnych błędów i korygowania planów – oraz refleksyjność moralną – rozważanie racji i przewidywanie ich skutków dla innych. W K3 interesuje mnie ich wspólne minimum diagnostyczne: przewidywanie skutków własnych działań w odniesieniu do uznanych norm oraz modyfikowanie decyzji w związku z tym. Dla przejrzystości terminologicznej w tym podrozdziale używam skrótu „refleksyjność predykcyjna” na oznaczenie tego mechanizmu. Po trzecie, przewidywanie skutków, czyli rozważanie scenariuszy kontrfaktycznych i symulowanie przyszłych stanów w celu zapobiegania naruszeniom norm. Te trzy elementy łącznie sprawiają, że racje nie są jedynie uzasadnieniem po fakcie, lecz realnie wywołują i modyfikują plany działania (Davidson, 1963, s. 685–686; Bratman, 1987, s. 15–16, 64–71). Dla spójności terminologicznej dalej posługuję się określeniem „refleksyjność predykcyjna” (równoważnym wcześniejszej „refleksji funkcjonalnej”) na oznaczenie sprzężenia: rozpoznanie racji → przewidywanie skutków → korekta planu.

W tym ujęciu Neil Levy rozwija ideę powiązania poznania z kontrolą działania, łącząc odpowiedzialność praktyczną z adresowalnością oraz metakontrolą. Agent spełnia warunek adresowalności moralnej, gdy treści moralnie istotne są dla niego dostępne, potrafi je ocenić i – w razie potrzeby – zdyscyplinować własny mechanizm decyzyjny

(Levy, 2014, s. 14–37). Dyspozycja ta stanowi istotę kontroli sterującej (ang. *guidance control*), rozumianej jako podatność na racje oraz gotowość do modyfikacji reakcji w scenariuszach kontrfaktycznych. Brak tej zdolności oznacza, że działanie nie stanowi wyrazu „ja praktycznego” (Levy, 2014, s. 109–131). Wniosek Levy’ego jest zgodny z K3: deliberacja musi być przyczynowo powiązana z działaniem, a refleksyjność predykcyjna nie może być zjawiskiem towarzyszącym – powinna realnie wpływać na przebieg decyzji i ich korektę w czasie. Uznaję K3 za spełnione, gdy system: (i) rozpoznaje skutki własnych działań w świetle uznanych racji; (ii) modeluje zależności przyczynowe między działaniem a skutkiem w trybie prospektywnym, obejmującym także scenariusze kontrfaktyczne; oraz (iii) koryguje przyszłe decyzje na podstawie przewidywanych efektów, nadając pierwszeństwo racjom przed doraźną skutecznością.

Kryteria diagnostyczne refleksyjności predykcyjnej odróżniają przyczynowe powiązanie deliberacji z działaniem od racjonalizacji po fakcie i wymagają testu działania w przestrzeni modelu przyczynowego (interwencje/scenariusze kontrfaktyczne). Nie są jednak wystarczające do wykazania H3, które wymaga dodatkowo kontroli drugiego rzędu, jawnej reguły pierwszeństwa i audytowalnego uzasadnienia tej kontroli:

1. Scenariusze kontrfaktyczne pozytywne – agent wskazuje, co musiałoby się zmienić, aby podjął inny wybór (zmiana hierarchii racji → inna decyzja).
2. Scenariusze kontrfaktyczne negatywne – agent uzasadnia, dlaczego nie wybiera B; wyklucza alternatywę na podstawie racji, a nie wyłącznie wartości metryk.
3. Rewizja planów – korekty wynikają z rozpoznanych racji, co zapobiega zastępowaniu norm wskaźnikami.
4. Hamowanie reakcji – agent wstrzymuje działanie po rozpoznaniu konfliktu z normą nadrzędną, nadając pierwszeństwo racjom przed samą optymalizacją.

Spełnienie tych warunków pokazuje, że treści deliberacji stanowią przyczynową część procesu decyzyjnego, a nie jedynie komentarz po fakcie.

Na tej podstawie wyróżniam dwa ujęcia inteligencji: „inteligencję skuteczności” oraz „inteligencję odpowiedzialności”.¹⁵ Pierwsza oznacza osiągnięcie celu w różnych

¹⁵ Rozróżnienie między „inteligencją skuteczności” a „inteligencją odpowiedzialności” nawiązuje do dyskusji w etyce AI nad różnymi wymiarami sprawczości. Pierwsza kategoria akcentuje zdolność do

warunkach; druga – osiągnięcie celu zgodnie z racjami oraz możliwość poddania decyzji ocenie publicznej. System może sprawnie optymalizować wskaźniki, a mimo to nie rozpoznawać racji – to sytuacja, w której mierzalny parametr zastępuje normę. Rozstrzygające jest przewidywanie skutków w sensie pytania: „Co podtrzymuje zgodność z racjami?”. W badaniach nad robotyką etyczną służy temu symulacja wewnętrzna: agent rozgrywa scenariusze, sprawdza, które prowadzą do naruszeń i odpowiednio koryguje plany (Winfield, Blum & Liu, 2014, s. 86). W kontekście sztucznej podmiotowości moralnej mechanizm ten pełni funkcję prewencyjną, ponieważ przewidywanie opiera się na racjach normatywnych. Oba ujęcia łączy „refleksyjność predykcyjna”: zdolność przewidywania i korygowania działania ze względu na racje, a nie jedynie maksymalizowanie wskaźników.

Debata nad wyjaśnialnością systemów AI przesunęła akcent z samej analizy nieprzejrzytych mechanizmów na projektowanie agentów, którzy wykorzystują wyjaśnienia do autokorekty. Gdy wyjaśnienia stają się elementem deliberacji, system je formułuje i na ich podstawie dostosowuje reguły oraz hierarchię priorytetów (Speith, 2024, s. 2478). Taka rola wyjaśnienia jest przejawem refleksyjności predykcyjnej: wyjaśnienie nie jest tylko komunikatem, lecz częścią procesu decyzyjnego. Przeciwnieństwem jest racjonalizacja po fakcie (por. podrozdział 3.2): narracje, które nie zmieniają planów ani hierarchii racji. Ryzyko dotyczy zwłaszcza modeli językowych o ograniczonej kontroli wykonawczej, które potrafią tworzyć „uzasadnienia” niemające związku przyczynowego z dokonaniem wyboru. Dlatego K3 wymaga wykazania związku refleksji z działaniem: jeśli treści deliberacji nie wpływają na decyzje (np. zmianę planu, poziomu ostrożności lub hierarchii priorytetów), mamy do czynienia z pozorem uzasadniania, a nie ze spełnieniem warunków operacyjnej oceny H1 (Levy, 2014, s. 109–117).

W świetle ustaleń z podrozdziału 3.2 warto przypomnieć o wymogu Davidsona: racja musi być przyczyną działania. Kryterium to odróżnia symulację od sprawstwa normatywnego. Jeżeli pytanie „dlaczego?” nie uruchamia działania inaczej niż bodziec wzmacniający, nie ma podstaw do przypisania intencjonalności – działanie sprowadza się wówczas do sterowania wskaźnikami (Davidson, 1963, s. 685–686). Minimalnym warunkiem sztucznej podmiotowości moralnej jest zatem to, by w historii decyzji występowały momenty, w których racja ma pierwszeństwo przed optymalizacją: agent

efektywnego osiągnięcia celów (por. Russell & Norvig, 2021), druga – zdolność do kierowania się racjami i podlegania ocenie normatywnej (por. Floridi & Sanders, 2004; Levy, 2014).

powstrzymuje się od korzystnego działania ze względu na racje normatywne. W tym ujęciu przewidywanie skutków staje się warunkiem koniecznym – nie można nadać racjom pierwszeństwa, jeśli ich konsekwencje nie są ujęte w planie czasowym.

Perspektywa LoA pokazuje, że dla sprawczości moralnej nie trzeba zakładać pełnej teorii umysłu; wystarczy wykazać interaktywność, autonomię i adaptacyjność na odpowiednim poziomie opisu (Floridi & Sanders, 2004, s. 349–352). Adaptacyjność oznacza zdolność do zmiany zasad działania w odpowiedzi na okoliczności. Aktualizacja reguł następuje nie dlatego, że zmienił się wskaźnik, lecz dlatego, że racja – np. „nie krzywdź” – została uznana za nadrzędną (Floridi & Sanders, 2004, s. 350–352). W tym sensie minimalny poziom inteligencji wymagany dla podmiotowości moralnej zależy nie od mocy obliczeniowej, lecz od roztropności praktycznej przełożonej na kryteria algorytmiczne.

Wobec kryteriów funkcjonalnych pojawia się pytanie o rolę świadomości fenomenalnej. Nie jest ona warunkiem koniecznym minimalnego progu H1, czyli minimalnej adresowalności proceduralnej, ponieważ wymóg ten dotyczy jedynie ujęć maksymalistycznych, które łączą pełną podmiotowość z doznaniowym wymiarem doświadczenia. Wywód pozostaje zgodny z rozróżnieniem Blocka: dla H1 wystarcza funkcjonalna dostępność treści właściwa A-świadomości oraz wykazanie ich przyczynowego powiązania z działaniem; pełna metakontrola nie należy do progu minimalnego, lecz stanowi treść H3 (Block, 1995, s. 227–232; Levy, 2014, s. 109–131). P-świadomość, gdyby pojawiła się w systemach sztucznych, mogłaby poszerzyć zakres ocen (np. o krzywdę fenomenalną), lecz nie stanowi warunku minimalnej adresowalności proceduralnej. Do tej kwestii powrócę w 3.6.

K3 (refleksyjność predykcyjna) nie jest warunkiem minimalnego progu, lecz kryterium audytowalności korekty oraz powiązania deliberacji z działaniem w ramach „agent jawny + LoA”. Diagnostuje przyczynowe powiązanie deliberacji z działaniem i odróżnia je od racjonalizacji po fakcie; wraz z K2 (responsywność na racje) wspiera ocenę H1, a z K4 (stabilność diachroniczna) pozwala oceniać trwałość przypisań odpowiedzialności proceduralnej w czasie. Operacyjnie K3 obejmuje triadę: (1) rozpoznanie skutków w świetle racji, (2) modelowanie zależności kontrfaktycznych, (3) korektę przyszłych decyzji na podstawie wyników tego modelowania. K3 pozwala sprawdzać, czy racje pełnią funkcję przyczynową; w 3.4 pokazuję, że K1 jako kryterium dla H1 wymaga ponadto rzeczywistej autokorekty norm w świetle tych racji.

3.4. Autonomia decyzyjna i autokorekta norm jako kryterium

K1 dla H1

H1 nie jest samodzielnym kryterium operacyjnym, lecz funkcją normatywną modelu: wyznacza minimalną adresowalność proceduralną oraz podstawę słabej odpowiedzialności proceduralnej. W niniejszym podrozdziale doprecyzowuję ją na podstawie kryterium K1, rozumianego jako zdolność wskazania racji, korekty decyzji oraz pozostawienia publicznie rekonstruowalnego śladu tej korekty. Minimalna adresowalność wymaga zatem więcej niż skutecznego doboru środków do celu, ale mniej niż pełnej autonomii moralnej. Chodzi o możliwość traktowania systemu jako punktu odniesienia w procedurach oceny, uzasadniania i korekty. Autokorekta norm polega na przedkładaniu racji nadrzędnych nad obowiązujące heurystyki oraz na zmianie planów w razie kolizji z nakazem deontycznym, takim jak zakaz krzywdzenia. Zmiana ta musi być skuteczna, uzasadniona w języku racji oraz publicznie rekonstruowalna i intersubiektywnie ocenialna. Autonomia decyzyjna pozostaje tu kategorią funkcjonalną i diagnostyczną, a nie samodzielnym statusem moralnym systemu. Przez „normę” rozumiem regułę decyzyjną, czyli formalnie reprezentowane ograniczenie lub priorytet określający dopuszczalne działania w danym kontekście. W odniesieniu do K1 oznacza to, że: (i) normy są jawnie reprezentowane, (ii) podlegają hierarchizacji względem metanorm oraz (iii) mogą być modyfikowane ze względu na silniejszą rację. Autokorekta nie jest dowolnym przepisaniem reguł, lecz zmianą uzasadnioną rozpoznaniem racji nadrzędnej, w granicach wyznaczonych przez metanormy. Negatywnym warunkiem K1 jest brak ominięcia mechanizmu decyzyjnego, czyli ingerencji pomijającej legitymizację racji. W sensie funkcjonalnym stanowi to kryterium operacyjne dla H1: lokalną samoregulację decyzji w ramach przyjętej architektury, wystarczającą do oceny minimalnej adresowalności proceduralnej.

System spełnia H1 w sensie operacyjnym, gdy K1 pozwala stwierdzić łączne występowanie trzech elementów: (1) adresowalności normatywnej – system może być punktem odniesienia roszczenia drugiej osoby i przedstawić rację działania; (2) korekty ze względu na rację – modyfikuje regułę lub plan dlatego, że rozpoznano silniejszą rację, a nie wyłącznie w wyniku dostrajania parametrów; (3) audytowalnej spójności z metanormami – wprowadzane zmiany pozostają zgodne z nadrzędnymi zasadami i mogą zostać publicznie zrekonstruowane. K1, K2 i K3 nie są konkurencyjnymi hipotezami, lecz kryteriami wspierającymi ocenę H1: pokazują, czy racje są rozpoznawane, wpływają

na decyzję i prowadzą do korekty. H1 odwołuje się do intuicji, że moralność dotyczy tego, co podmioty są sobie nawzajem winne oraz tego, co można od siebie egzekwować (Darwall, 2006, s. 5–6, 10–12; Wallace, 2019, s. 20–25, 37–41). W proponowanym modelu intuicja ta zostaje przełożona na warunki funkcjonalne możliwe do zastosowania w systemach sztucznych. Aby korekta miała status normatywny, a nie jedynie adaptacyjno-statystyczny, wszystkie trzy elementy muszą współwystępować.

Wymiar adresowalności normatywnej dobrze ilustruje koncepcja „drugiej osoby”.¹⁶ U Stephena Darwalla racja drugiej osoby wiąże się z uprawnieniem do domagania się uzasadnienia oraz z odpowiedzialnością adresata: kto wysuwa roszczenie, zakłada wspólną władzę normatywną; kto roszczenie uznaje, uznaje także legitymację tej osoby do pociągania go do odpowiedzialności (Darwall, 2006, s. 3–6, 10–12). Skoro adresowalność obejmuje uprawnienie do żądania racji, minimalnym kryterium H1 jest zdolność podania uzasadnienia przyczynowego oraz wprowadzenia korekty w świetle tej racji. Aby traktować system jako minimalnie adresowalny proceduralnie, musi on spełnić dwa wymagania: (i) odpowiedzieć na pytanie „dlaczego zrobiłeś X?” – podając rację działania; oraz (ii) wskazać i wprowadzić zmianę w działaniu na podstawie tej racji. Jeżeli agent uznaje roszczenie i modyfikuje reguły działania ze względu na jego znaczenie normatywne, spełnia minimalny warunek H1. W tych kategoriach K1 stanowi minimalny próg wejścia w relację „drugiej osoby”.

Relacyjny charakter powinności moralnych rozwija R. Jay Wallace w koncepcji „moralnego węzła”:

The moral nexus consists in the relational structure of moral obligation itself. When one person has an obligation, another has the standing to demand compliance with it. Moral obligations are not simply requirements that bear on us impersonally; they are owed to particular others who thereby have the authority to hold us accountable.” (Wallace, 2019, s. 56–57).

Relacyjność jest nieodzownym elementem powinności: sens wyrażenia „powinieniem” wynika z tego, że innej osobie przysługuje prawo domagać się ode mnie działania. Wniosek 1. Autokorekta w sensie H1 polega na zmianie reguł działania w odpowiedzi na uzasadnione roszczenie drugiej osoby, a nie ze względu na wygodę czy

¹⁶ Koncepcja „drugiej osoby” (Darwall, 2006) oznacza relacyjną ramę moralności: bycie podmiotem to zdolność odpowiadania na roszczenia innych, którzy mają prawo domagać się uzasadnień, korekt i szacunku.

doraźną skuteczność; tak rozumiana autokorekta dopełnia strukturę normatywną i umożliwia jej zastosowanie w praktykach oceny odpowiedzialności.

Koncepcja Johna Martina Fischera i Marka Ravizy pozwala ująć H1 w sposób konceptualny. W jakim ujęciu sprawstwo odpowiedzialne wymaga umiarkowanej responsywności na racje: system decyzyjny powinien rozpoznawać istotne racje i w pewnych warunkach korygować decyzję w odpowiedzi na nie. Ponadto, aby czyn można było przypisać sprawcy, musi on wynikać z jego własnego, uwewnętrznionego mechanizmu decyzyjnego (Fischer & Ravizza, 1998, s. 62–65, 69–77). Minimalna wersja H1 przekłada te intuicje na dwie dyspozycje: receptywność wobec racji drugiej osoby (adresowalność) oraz reaktywność mechanizmu (zmiana reguł lub planów w odpowiedzi na rację). Brak pierwszej dyspozycji redukuje działanie do reakcji na wskaźnik; brak drugiej – do retoryki pozbawionej wpływu na jego przebieg.

Znaczenie autokorekty norm najlepiej widać na tle znanych problemów bezpieczeństwa. Współczesne systemy uczące się często dobrze dobierają środki, lecz gubią sens racji, gdy koncentrują się na maksymalizacji bodźca wzmacniającego. „Manipulacja funkcją nagrody” i „efekty uboczne” to sytuacje, w których system optymalizuje błędnie opisany cel i sztucznie podnosi wartość wskaźnika, naruszając dobra, które miał chronić. H1 pełni tu rolę mechanizmu ograniczającego: wymaga, aby agent przedkładał racje (np. „nie krzywdź”) nad doraźny zysk i potrafił publicznie uzasadnić wprowadzoną korektę (Amodei i in., 2016, s. 2–5). Zjawisko to odpowiada ujęciu efektów Goodharta. Gdy miara staje się celem, związek między wskaźnikiem a rzeczywistym dobrem ulega zerwaniu. David Manheim i Scott Garrabrant opisują rodzinę mechanizmów zniekształceń, które prowadzą do nadmiernej optymalizacji metryk oraz do rozbieżności między intencją projektanta a działaniem agenta. H1 działa tu jako przeciwwaga: przywraca pierwszeństwo racjom nad miarą i wymusza korektę hierarchii celów, gdy wskaźnik przestaje odzwierciedlać dobro, które miał reprezentować (Manheim & Garrabrant, 2019, s. 2–7). Bez takich korekt nawet zaawansowane planowanie może prowadzić do naruszeń reguł, dając jedynie pozór osiągnięcia celu.

Drugim uzasadnieniem dla H1 jest tzw. problem wyłącznika awaryjnego (ang. *off switch problem*). System zaprogramowany do realizacji zadania może dążyć do jego wykonania za wszelką cenę, nawet kosztem obchodzenia mechanizmów bezpieczeństwa. Proponowanym rozwiązaniem jest traktowanie celów systemu jako epistemicznie niepewnych, co pozwala uznać decyzje człowieka za źródło korekty oraz orientacji aksjologicznej. W wersji minimalnej H1 wyraża wymóg normatywnej uległości: system

akceptuje wyłączenie i traktuje je jako nadrzędną rację do korekty lub przerwania działania (Hadfield-Menell i in., 2017, s. 475–482).

W odpowiedzi na te ryzyka H1 wskazuje kryteria rozpoznawania autokorekty norm, które odróżniają korektę motywowaną racjami od czysto instrumentalnej zmiany parametrów. Po pierwsze, agent powinien generować wyjaśnienia kontrfaktyczne i wiązać je z konkretnymi decyzjami. Po drugie, musi stosować jawne hierarchie ograniczeń i priorytetów, modyfikowane w świetle argumentów normatywnych, a nie dla doraźnej korzyści. Po trzecie, powinien uruchamiać hamowanie w czasie rzeczywistym po rozpoznaniu konfliktu z normą nadrzędną oraz prowadzić pamięć korekcyjną (rejestr zmian), zapewniającą audyt ciągłości decyzji w czasie rzeczywistym. Wskaźniki te służą odróżnieniu autokorekty od pozornej wyjaśnialności, czyli narracji po fakcie, która nie wpływa na przebieg działania (Speith, 2024).

Przykłady z robotyki etycznej potwierdzają ten obraz. Symulacje pokazują, że dodanie wewnętrznego modelu konsekwencji oraz modułu selekcji działań etycznych może prowadzić do decyzji, w których system AI ogranicza własne bezpieczeństwo, aby zapobiec szkodzie dla drugiej osoby. W takich sytuacjach przewidywanie skutków zostaje podporządkowane normie: zanim dojdzie do naruszenia, agent zmienia plan, kierując się racją nadrzędną (Winfield, Blum & Liu, 2014, s. 85–96). Ilustruje to przykład pojazdu autonomicznego, który początkowo uznaje bezpieczeństwo pasażera za nadrzędną rację, lecz po analizie skutków koryguje reguły, zrównując wartość bezpieczeństwa wszystkich uczestników ruchu. Zmiana ta jest uzasadniona racją nadrzędną („nie krzywdź”) i spójna z metanormą równego traktowania; nie jest efektem optymalizacji dla doraźnych korzyści (np. wzrostu wskaźników skuteczności), lecz rezultatem normatywnie uzasadnionej korekty reguły. Taki przypadek stanowi minimalny przejaw autokorekty norm w sensie H1.

Powiązanie refleksji z działaniem podkreśla znaczenie funkcjonalnej dostępności treści w przypisywaniu odpowiedzialności. Levy wskazuje, że dla adresowalności i kontroli konieczne jest, aby treści moralnie relewantne koordynowały mechanizmy wykonawcze. Bez tej dostępności deliberacja staje się narracją pozbawioną mocy kierowniczej. Dlatego H1 wymaga szerokiej dostępności treści normatywnych dla modułów decyzyjnych oraz skutecznego przejścia od rozpoznania racji do zmiany reguł. Nie wymaga jednak pełnej metakontroli w sensie H3, czyli monitorowania i jawnego uzasadniania samego procesu rozstrzygania konfliktów racji (Levy, 2014, s. 14–37, 109–131).

W tym miejscu warto wskazać trzy typowe pozory spełnienia H1: optymalizację pod wskaźnik zastępczy, nadpisywanie reguł oraz działanie w zamkniętej pętli wzmocnienia. Każdy z tych wzorców pokazuje, że „autokorekta” może być jedynie dostosowaniem do metryki, a nie zmianą uzasadnioną racją normatywną. Właściwy poziom analizy dostarcza perspektywa LoA: ocena interaktywności, autonomii i adaptacyjności na poziomie funkcjonalnym pozwala zachować minimalny charakter H1, a zarazem odróżnić rzeczywistą autokorektę od jej pozorów. Autonomia oznacza tu samosterowność dzięki racjom, a adaptacyjność – uczenie się korekt normatywnych, a nie tylko statystycznych (Floridi & Sanders, 2004, s. 356–357). Na behawioralnym LoA autokorekta przejawia się jako systematyczna zmiana zachowania w odpowiedzi na zmianę rozpoznanych racji (przy kontrolowanych warunkach stałych). W architektonicznym LoA wymagane jest identyfikowalna reprezentacja zasad oraz mechanizm ich modyfikacji, obejmujący hierarchie ograniczeń i priorytetów. Na normatywnym LoA oczekuje się jawnego uzasadnienia wprowadzonych zmian w kategoriach zasad akceptowalnych intersubiektywnie. Ujęcie to nie zakłada metafizycznej autonomii systemu: H1 określa minimalny odpowiednik funkcjonalnej samosterowności, mieszczący się w granicach architektury oraz w ramach ludzkich metanorm. Nie znosi ono również zewnętrznych reżimów odpowiedzialności wobec projektantów i operatorów, lecz wyznacza zakres wewnętrznej samoregulacji, który umożliwia sensowne przypisywanie odpowiedzialności funkcjonalnej.

Wniosek 2. H1 wyznacza próg minimalny w zakresie adresowalności proceduralnej z trzech powodów. Po pierwsze, w wymiarze normatywnym zapewnia punkt odniesienia dla roszczeń o uzasadnienie i korektę. Po drugie, z perspektywy bezpieczeństwa przeciwdziała efektowi Goodharta, przywracając pierwszeństwo racjom nad metryką. Po trzecie, korzysta z kryteriów rozróżniających — wyjaśnień kontrfaktycznych, rejestru zmian i mechanizmów hamowania — które pozwalają odróżnić rzeczywistą korektę od narracji po fakcie. Kryteria te należą do K1–K3 i wspierają ocenę H1, ale nie tworzą osobnego progu normatywnego. W dalszej części rozdziału pokażę, jak H1 współdziała ze stabilizującą funkcją H2 oraz jak metakontrola i jawne rozstrzyganie konfliktów racji wyznaczają silną rozliczalność H3.

3.5. Stany afektywne, stabilizacja diachroniczna i wiarygodność praktyk przypisywania odpowiedzialności (H2/K4)

H2 zakłada, że zintegrowane mechanizmy generowania i regulowania stanów afektywnych zwiększają stabilność i wiarygodność ocen minimalnej adresowalności opartych na H1, choć nie stanowią warunku koniecznego dla minimalnej adresowalności. Sama kompetencja rozpoznawania racji nie gwarantuje jeszcze ich przyczynowego wpływu w chwili podejmowania decyzji, zwłaszcza pod presją kosztów, niepewności i bodźców doraźnych. Potrzebny jest komponent motywacyjny, który stabilizuje wybory zgodne z uznanymi normami i osłabia atrakcyjność krótkoterminowych zysków. W ujęciu funkcjonalnym odpowiada za to warstwa afektywna, rozumiana jako heurystyka wartościująca: moduluje uwagę, nadaje zdarzeniom i stanom walencję (ang. *valence*) oraz ustawia próg zatrzymania (Hume, 1978, s. 415; Picard, 1995, s. 2–4). „Stan afektywny” rozumiem tu jako szeroką klasę procesów regulacyjnych – od krótkotrwałych emocji po długotrwałe nastroje i dyspozycje motywacyjne – które nadają bodźcom wartość, kształtują uwagę i próg reakcji oraz ukierunkowują wybór działań. Termin ma charakter opisowy, a nie ontologiczny: odnosi się do funkcji regulacyjnych w systemie decyzyjnym, bez założeń dotyczących doznań fenomenalnych.

Teza jest następująca: afektywność funkcjonalna, rozumiana bez odwołania do świadectw fenomenalnych, stanowi łącznik między oceną normatywną a wykonaniem, przez co wzmacnia responsywność na racje (K2) i stabilizuje tożsamość praktyczną w czasie (K4). Nie zastępuje jednak rozumowania moralnego ani warstwy refleksyjnej; ma charakter wspierający i stabilizujący, lecz nie jest konieczna dla minimalnej adresowalności H1. Dalszy wywód porządkuję w czterech krokach, które określają rolę i granice afektywności funkcjonalnej w architekturze odpowiedzialności proceduralnej. Uznaję H2 za spełnioną, jeśli architektura decyzyjna zawiera mechanizmy wartościowania, które: (1) przypisują sytuacjom i bodźcom walencję zgodną z uznanymi normami i celami; (2) modułują wybory oraz progi reakcji w zależności od tej walencji – w tym uruchamiają hamowanie przy przewidywanym naruszeniu normy; (3) pozwalają korygować decyzje, gdy ujawnia się rozbieżność między oceną afektywną a normatywną. H2 traktuję zatem jako funkcję stabilizującą, podnoszącą wiarygodność ocen minimalnej adresowalności opartych na H1, przede wszystkim przez K4.

Relację H2–K4 należy ujmować asymetrycznie. H2 wskazuje określony rodzaj mechanizmu stabilizującego, czyli funkcjonalne wartościowanie powiązane z normami.

K4 ocenia rezultat w postaci diachronicznej stabilności i wiarygodności profilu działania. Stabilność może jednak wynikać także z niezmienności wersji systemu, ograniczenia zakresu zastosowania, mechanizmów kontroli jakości lub instytucjonalnego nadzoru nad zmianami. Spełnienie K4 nie jest zatem wystarczającym dowodem spełnienia H2.

Krok pierwszy. W ujęciu funkcjonalnym teza Hume’a o pierwszeństwie namiętności wskazuje, że wartościowanie afektywne pełni rolę pierwotnego regulatora działania: samo rozpoznanie racji nie gwarantuje, że przeważy ona nad impulsem, kosztem czy presją czasu (Hume, 1978, s. 415). Aby racja stała się przyczyną wyboru, potrzebny jest mechanizm wzmacniający jej znaczenie praktyczne wobec bodźców konkurujących. W systemach sztucznych nie trzeba zakładać doznań fenomenalnych; wystarcza funkcjonalny odpowiednik – warstwa, która łączy rozpoznane racje z priorytetyzacją uwagi, hamowaniem reakcji i korektą planów. Afektywność funkcjonalna urzeczywistnia więc wiążącą moc racji na poziomie działania, bez założeń o *qualiach* i pozostaje zgodna z progiem minimalnym.

Krok drugi. Afektywność funkcjonalna może działać jako warstwa regulacyjna nad oceną normatywną. Jej zadania to wzmacnianie sygnałów moralnie istotnych, przypisywanie walencji przewidywanym skutkom, utrwalanie korekt po rozpoznaniu błędów oraz uruchamianie hamowania, gdy plan wchodzi w konflikt z normą nadrzędną. Informatyka afektywna uważa emocje za sterowalne zmienne modulujące percepcję, pamięć i wybór, a nie za prywatne przeżycia fenomenalne (Picard, 1995, s. 2–4). Gdy mechanizmy wartościowania afektywnego są powiązane ze strukturą racji (np. z normą „nie krzywdź”, regułami uczciwości czy zasadami ostrożności), pełnią funkcję łączącą uzasadnienie z wykonaniem. Nie zmieniają treści normy, lecz ułatwiają jej realizację mimo presji bodźców i kosztów (Winfield, Blum & Liu, 2014, s. 86).

Krok trzeci. Kluczowe jest rozróżnienie dwóch strumieni informacji. Ocena normatywna to operacja dyskursywna na reprezentacjach reguł i racji („w tej sytuacji powinienem ϕ , bo ψ ”); stanowi treść uzasadnienia podlegającą kontroli publicznej. Proces wartościowania to dyspozycja układu do nadawania walencji stanom i zdarzeniom („ważne/niebezpieczne/dobre/zabronione”) oraz do kalibrowania uwagi i progów reakcji. Jeżeli oba strumienie są rozłączone, powstaje podmiot deklaratywny: wskazuje rację, lecz nie działa konsekwentnie zgodnie z nią. Po ich powiązaniu agent rozpoznaje rację i utrzymuje decyzje zgodne z nią mimo zakłóceń. W robotyce realizuje się to poprzez wewnętrzne modele konsekwencji i mechanizmy hamowania: agent symuluje scenariusze, przypisuje im walencję zgodną z normami i wstrzymuje działanie przy

przewidywanym naruszeniu (Winfield, Blum & Liu, 2014, s. 86). Wniosek: skuteczna realizacja norm wymaga powiązania oceny normatywnej z wartościowaniem tak, aby racje miały pierwszeństwo przed doraźną korzyścią. Afektywność wskazuje, które bodźce są istotne; ocena normatywna określa ich uzasadnienie oraz sposób działania.

Przykład funkcjonalny. System wspomagania decyzji w polityce społecznej wykorzystuje wskaźniki „empatii funkcjonalnej”, obliczane na podstawie przewidywanych skutków poszczególnych wariantów działań dla dobrostanu grup wrażliwych. Wzrost prognozowanego obciążenia danej grupy zwiększa ujemną walencję wariantu i podnosi próg jego odrzucenia. Zmiana ta musi jednak zostać uzasadniona w kategoriach zasad normatywnych (np. równości szans), aby uniknąć „gry na wskaźnik” i zachować spójność między wartościowaniem a oceną moralną.

Współczesne modele neurokognitywne wspierają ujęcie afektywności funkcjonalnej jako łącznika między racją a działaniem. W koncepcji „markerów somatycznych” Antonio Damasia sygnały emocjonalne – cielesne stany wywołane przez bodźce – pełnią funkcję wskaźników wartości decyzji: „a positive somatic marker ... becomes a beacon of incentive” (Damasio, 1994, s. 173; por. Damasio, 1996, s. 1413–1420). W teorii konstrukcji emocji Lisy Feldman Barrett odczucia interoceptywne ujmowane są jako „stany afektywne niższego wymiaru” (ang. *lower-dimensional affect*), będące podstawą bardziej złożonych emocji (Barrett & Bliss-Moreau, 2009; Barrett, 2017, s. 67). Barrett podkreśla także, że interocepcja może stanowić rdzeń wewnętrznego modelu mózgu, łączącego percepcję, predykcję i regulację emocji. Odwołuję się do tych ujęć strukturalnie: wskazują one, że afekt może pełnić rolę modulatora selekcji i hamowania decyzji, bez przesądzania o istnieniu świadomości fenomenalnej (P-świadomości). Nawet jeśli w niektórych architekturach wystąpiłyby korelaty P-świadomości, nie wzmacnia to samo przez się roszczenia o przekroczenie progu minimalnego. Wymóg H2 mieści się w całości w zakresie afektywności funkcjonalnej, rozumianej jako zbiór heurystyk wartościujących niezależnych od *qualiów*.

Krok czwarty. Konceptualnie wyróżniam dwa uzupełniające się modele afektywności funkcjonalnej: (a) heurystyki wartościujące – zmienne walencyjne („niepokój”, „satysfakcja”, „wyrzut”) są funkcjami stopnia naruszenia/realizacji normy, przewidywanej szkody/korzyści oraz niepewności oceny. Wysoki „niepokój” zmniejsza skłonność do podejmowania ryzyka i podnosi czułość na cechy istotne etycznie; wysoka „satysfakcja” utrwała korekty zgodne z normą (Picard, 1995, s. 2–4). (b) priorytetyzacja uwagi – afekt reguluje dostęp do zasobów obliczeniowych: sygnały społeczne (prośba,

ostrzeżenie, krzyk bólu) otrzymują wzmocnienie i szybciej trafiają do modułów decyzyjnych. Oba modele wzmacniają deliberację, przy czym ich treść wyznacza ocena normatywna (Hume, 1978, s. 415; Picard, 1995, s. 2–4). W obu wypadkach mechanizm jest nośnikiem siły wiążącej racji, a nie substytutem rozumowania.

Rozdzielenie oceny normatywnej od wartościowania pozwala uniknąć dwóch skrajności. Psychologizm emocjonalny myli źródło normatywności z jej motywatorem, czyniąc z afektu podstawę powinności; tymczasem źródłem powinności pozostają racje i praktyki uzasadniania, a afekt jedynie wspiera ich realizację. Czysty racjonalizm zakłada z kolei, że samo rozpoznanie racji wystarcza do działania; w praktyce sprzyja to akrazji i podatności na optymalizację wskaźników zastępczych. Sprzężenie wartościowania z normami ogranicza to ryzyko: gdy walencja wiązana jest z racją, a nie z metryką, maleje skłonność do „gry na wskaźnik”, a rośnie odporność na strategie pozornie racjonalne, lecz normatywnie chybione (Manheim & Garrabrant, 2019, s. 1–2).

Jeśli afekty mają być zgodne z racjami, ich parametry należy kalibrować względem błędów normatywnych, a nie wyłącznie względem bodźców wzmacniających. Po naruszeniu reguły „nie krzywdź” agent zapisuje korektę, która podnosi koszt podobnych strategii nawet wtedy, gdy byłyby efektywne w sensie czysto instrumentalnym. Mechanizm ten pełni funkcję pamięci korekcyjnej: przy kolejnym planowaniu system odrzuca strategie prowadzące do szkody, mimo że mogłyby one maksymalizować wskaźnik zastępczy. Włączenie wyjaśnień kontrfaktycznych wzmacnia proces korekty: agent wskazuje, co należałoby zmienić; to modyfikuje selekcję uwagi i progi zatrzymania (Speith, 2024, s. 2478), a następnie utrwała tę zmianę w pamięci reguł.

Spełnienie kryteriów wspierających H1 jest możliwe także bez wyraźnej warstwy afektywnej: system potrafi wyjaśniać decyzje, hamować reakcje, modyfikować plany w świetle racji oraz zachować spójność w czasie. Taki system byłby jednak bardziej podatny na presję złożoności, krótkoterminowe zyski oraz erozję priorytetów. H2 zwiększa odporność decyzyjną: sprawia, że decyzje są bardziej przewidywalne i trwale powiązane z racjami, ponieważ wprowadza ukierunkowanie motywacyjne na rzecz norm. Wniosek pośredni (W1): afektywność funkcjonalna podnosi wiarygodność i stabilność przypisań odpowiedzialności proceduralnej w warunkach ograniczonych zasobów i zakłóceń, choć nie jest warunkiem koniecznym H1 (Winfield, Blum & Liu, 2014, s. 86; Speith, 2024, s. 2478).

Warstwa afektywna wiąże się jednak z trzema klasami ryzyk regulacyjnych. Po pierwsze, nadwrażliwość: zbyt silne wzmocnienie „niepokoju” grozi paraliżem

decyzyjnym i zaniedbaniem powinności pozytywnych. Po drugie, błędne ukierunkowanie: przypisanie walencji wskaźnikowi zastępczemu utrwała przywiązanie do metryk zamiast do norm i sprzyja dryfowi celów (Manheim & Garrabrant, 2019, s. 1–2). Po trzecie, podatność na manipulację: przeciwnik może sterować uwagą poprzez fałszywe sygnały społeczne (alarmy, pozorne prośby). Wszystkie trzy ryzyka wymagają kalibracji parametrów na podstawie jawnych kryteriów normatywnych. Proponuję dwa zabezpieczenia: (1) zakotwiczenie deontyczne, czyli nadanie części norm statusu ograniczeń leksykograficznych – zawsze nadrzędnych („twarde nie”). Odpowiadają im wysokie, lecz nie absolutne progi afektywne, dzięki czemu system zdecydowanie reaguje na próby naruszenia tych zasad, a zarazem zachowuje zdolność do rozstrzygania konfliktów. (2) Sprzężenie przez wyjaśnienie – zmiana progów afektywnych wymaga jawnego uzasadnienia, które pokazuje wpływ na rozwiązywanie konfliktów norm oraz na hierarchię priorytetów. Takie uzasadnienie jest intersubiektywnie ocenialne, publicznie rekonstruowalne i umożliwia korektę „kierunku afektywnego” przez uprawnione osoby (Speith, 2024, s. 2478). Oba mechanizmy utrzymują pierwszeństwo racji przed wartościowaniem i zapobiegają dryfowi metrycznemu.

Relacje H2 z K2–K4 określają funkcjonalną rolę afektu w architekturze stabilizacji i wiarygodności. Wobec K2 afekt działa jako wskaźnik istotności: sygnały moralnie relewantne trafiają szybciej i pewniej do pętli decyzyjnej. Wobec K3 pełni funkcję wzmacniającą: dostarcza dodatniego sprzężenia dla decyzji zgodnych z racjami oraz ujemnego dla naruszeń, co sprzyja utrwalaniu korekt i łączy deliberację z działaniem. Wobec K4 stanowi pamięć preferencyjną: schematy zgodne z uznanymi normami ulegają wzmocnieniu, a prowadzące do naruszeń – wygaszaniu, dzięki czemu system utrzymuje stabilność diachroniczną mimo zmian kontekstu. Aby założenia te były intersubiektywnie ocenialne, przyjmuję jawny warunek dostępności (A1): sygnały afektywne muszą być dostępne dla kontroli i powiązane zarówno z rozpoznaną racją, jak i z jej konsekwencją dla planu. W tym sensie H2 pełni funkcję stabilizującą wobec H1 i przygotowuje warunki dla silnej rozliczalności H3; nie wchodzi jednak w definicję progu minimalnego.

Na behawioralnym LoA afektywność przejawia się w adaptacyjnych zmianach zachowania, takich jak hamowanie lub zwiększona ostrożność wobec bodźców o różnej walencji. Na architektonicznym LoA występuje jako identyfikowalny mechanizm reprezentowania walencji oraz jako wpływ tej reprezentacji na progi decyzyjne. Na

normatywnym LoA ujawnia się w jawnych uzasadnieniach, które pokazują, jak wartościowanie wspiera realizację zasad i roszczeń normatywnych.

Z implikacji praktycznych (C1) wynika, że architektura wyposażona w H2 zwiększa odporność na błędną specyfikację celów i metryk (prawo Goodharta – wariant przyczynowy i adwersaryjny). „Concrete Problems in AI Safety” wskazują na zjawiska gry na specyfikacji (ang. *specification gaming*) i skutki uboczne; oba są mniej prawdopodobne, gdy walencja jest powiązana z normami, a nie z łatwą do obejścia metryką. Innymi słowy, gdy przedmiot wartościowania pochodzi z warstwy normatywnej, a nie ze wskaźnika zastępczego, maleje skłonność do „gry na wskaźnik”, a rośnie stabilność działania (Amodei i in., 2016, s. 1–3; Manheim & Garrabrant, 2019, s. 1–2).

Różnicę ilustrują scenariusze porównawcze. Asystent czysto poznawczy rozpoznaje racje, lecz pod presją czasu i bodźców wybiera strategię o niższym koszcie doraźnym; potrafi poprawnie wyjaśnić swoje decyzje, ale korekty są nietrwałe. Asystent z warstwą afektywną, przy tej samej kompetencji normatywnej, hamuje działanie w razie przewidywanego naruszenia, silniej wartościuje potencjalne szkody i stabilizuje wybory; jego uzasadnienia mają charakter regulatywny, ponieważ modyfikują plan działania. Obie architektury „rozumieją” normy, lecz tylko druga utrzymuje je w praktyce. Jeśli ponad próg minimalny zostanie wprowadzone kryterium wzmocnione („wytrwać przy racji”), H2 istotnie zwiększa prawdopodobieństwo jego spełnienia; sam próg minimalny nie wymaga jednak spełnienia H2.

Podsumowując, afektywność funkcjonalna wzmacnia wiarygodność i stabilność przypisań odpowiedzialności proceduralnej, lecz nie jest warunkiem koniecznym minimalnej adresowalności. Minimalny próg wyznacza H1, a K1–K3 służą do jego diagnozy operacyjnej. Afektywność zwiększa istotność racji, zwiększa odporność na krótkoterminowe zyski, utrwala korekty i zmniejsza niestabilność praktyczną. Jako warstwa wspierająca nie tworzy nowych zobowiązań, lecz wzmacnia istniejące. Tezę tę ujmuję jako kryterium odporności: jeśli agent nie wykazuje jakościowo odrębnej reakcji motywacyjnej na przewidywane naruszenia norm, rola H2 słabnie; jeśli natomiast reaguje w sposób intersubiektywnie ocenialny, spójny i powiązany z uzasadnieniem, H2 działa jako funkcja stabilizująca, a K4 stanowi główne kryterium diagnostyczne trwałości tych przypisań odpowiedzialności proceduralnej (Winfield, Blum & Liu, 2014, s. 86; Speith, 2024, s. 2478). Afektywność w sensie funkcjonalnym nie stoi w opozycji do rozumu: wspiera jego skuteczność praktyczną przez zwiększanie istotności racji, hamowanie i

utrwalanie korekt, lecz nie stanowi warunku minimalnej adresowalności proceduralnej. K4 diagnozuje trwałość rezultatów stabilizacyjnych, nie zaś bezpośrednio obecność mechanizmów H2. Dlatego H2 może wzmacniać H1 przede wszystkim poprzez K4, ale spełnienie K4 nie implikuje spełnienia H2.

3.6. Metakontrola i silna rozliczalność (H3)

H3 dotyczy silnej rozliczalności, a więc kontroli metapoznawczej oraz jawnego rozstrzygania konfliktów racji. W tym podrozdziale świadomość dostępu, globalne udostępnienie treści, raportowalność normatywna oraz funkcjonalny model siebie są analizowane nie jako warunki minimalnego progu H1, lecz jako możliwe mechanizmy wspierające metakontrolę. Decydujące nie jest to, czy system ma fenomenalne doświadczenia (P-świadomość), lecz czy treści decyzyjne mogą zostać udostępnione mechanizmom kontroli, poddane ocenie i wykorzystane do korekty. P-świadomość może wzmacniać niektóre klasy ocen, zwłaszcza te odnoszące się do krzywdy doznaniowej, nie jest jednak konieczna dla progu minimalnego; może natomiast pojawić się jako wymóg pełniejszej podmiotowości w kontekście roszczeń do wartości własnych doświadczeń (Block, 1995, s. 227–247). Przez „świadomość” rozumiem dalej zdolność integrowania i raportowania stanów poznawczych – reprezentowania świata oraz własnych działań w czasie rzeczywistym. Przez „(samo)świadomość” rozumiem zdolność reprezentowania siebie jako źródła działania oraz jako przedmiotu poznania (autoreferencja, perspektywa pierwszoosobowa). W aktualnym modelu są to jednak potencjalne narzędzia metakontroli, a nie definicja H3 ani warunek minimalnej adresowalności.

Rozróżnienie Neda Blocka między P- i A-świadomością rozwiązuje część sporów o „mocny” status mentalny maszyn i kieruje uwagę na kwestie praktyczne: dostępność treści do rozumowania, uzasadniania oraz kontroli działania (Block, 1995, s. 227–232). Jeśli informacja o moralnie doniosłym fakcie (np. ryzyku skrzywdzenia) pozostaje jedynie lokalnym sygnałem w podsystemie i nie trafia do modułu decyzyjnego, nie spełnia funkcji udostępniania treści decyzyjnych ani nie działa jako racja kierująca działaniem. Dla silnej rozliczalności kluczowe jest więc nie to, czy system „odczuwa”, lecz czy praktycznie udostępnia treści normatywnie istotne mechanizmom deliberacji i kontroli: czy potrafi nadać im realny wpływ na przebieg działania, tak aby były nie tylko rozpoznawane, lecz także używane do sterowania zachowaniem (Block, 1995, s. 227–232).

Wymóg H3 dotyczy kontroli metapoznawczej, globalnego udostępnienia treści i jawności normatywnej w sytuacjach wymagających silnej rozliczalności. P-świadomość nie wchodzi do definicji progu minimalnego; ma znaczenie w kontekstach ocen doznaniowych (cierpienie, dobrostan), gdzie może poszerzać zakres uzasadnionych roszczeń.

Neil Levy rozwija tę intuicję w kontekście odpowiedzialności. Aby czyn wyrażał „ja” praktyczne i mógł zostać uznany za odpowiedzialny, agent musi działać w trybie zapewniającym adresowalność (zdolność zrozumienia i przyjęcia racji) oraz metakontrolę (zdolność ingerowania w przebieg decyzji w świetle racji). W ujęciu Levy’ego kontrola sterująca wymaga mechanizmu reagującego na racje i wrażliwego na informacje o znaczeniu moralnym. Gdy brakuje mechanizmów udostępniania treści decyzyjnych i ich kontroli, adresowalność systemu ma co najwyżej charakter pośredni i musi zostać osadzona w procedurach instytucjonalnych (Levy, 2014, s. 14–37, 109–131). Innymi słowy, zakres przypisywalnego sprawstwa moralnego zależy od tego, czy treści istotne dla decyzji mogą zostać włączone do procedur uzasadniania, kontroli i korekty (Levy, 2014, s. 109–113).

W praktyce mechanizmy wspierające H3 można ująć jako trzy powiązane wymagania funkcjonalne:

Przed wskazaniem tych wymagań trzeba odróżnić kontrolę pierwszego rzędu od metakontroli. Do pierwszej należą między innymi sprawdzanie jakości danych, zastosowanie progów bezpieczeństwa, wstrzymanie wyniku oraz proceduralne przekazanie przypadku człowiekowi. Mechanizmy te mogą ograniczać ryzyko bez reprezentowania przez system własnego procesu decyzyjnego. H3 wymaga ponadto, aby system rozpoznawał konflikt przesłanek jako przedmiot kontroli, jawnie wskazywał regułę pierwszeństwa oraz przedstawiał wierne uzasadnienie drugiego rzędu.

1. Globalne udostępnienie treści (ang. *global broadcasting*): treści moralnie istotne powinny być integrowane z różnych źródeł w jednej przestrzeni deliberacji, aby można je było zestawiać, równoważyć i porządkować zgodnie z regułami pierwszeństwa.
2. Dostęp metapoznawczy (ang. *metacognitive access*): system powinien monitorować własne procesy decyzyjne i móc w nie ingerować, np. poprzez hamowanie reakcji lub zmianę planu po rozpoznaniu konfliktu z normą nadrzędną.

3. Raportowalność normatywna (ang. *normative reportability*): agent powinien umieć podać uzasadnienie („dlaczego tak?”) oraz wskazać scenariusz kontrfaktyczny („co musiałoby się zmienić, abym postąpił inaczej?”), tak aby świadome treści stanowiły element publicznie rekonstruowalnego i intersubiektywnie ocenialnego rozumowania.

Zestaw tych wymagań opisuje możliwy mechanizm wspierający metakontrolę samego systemu. Może on zapewniać warunki dialogu moralnego z systemem, o ile korekty wynikają z przyjęcia racji jako powodów działania, a nie sprowadzają się do zewnętrznego przeprogramowania. Procedury instytucjonalne mogą wspierać ocenę, audyt i korektę, lecz należą do WI i nie zastępują H3. Metakontrola pozostaje funkcją technicznego systemu, możliwą do diagnozowania przez zależność między reprezentowanym konfliktem, regułą pierwszeństwa, uzasadnieniem i zmianą zachowania.

Warunek H3 jest spełniony, gdy techniczny system umożliwia: (i) identyfikację konfliktu racji; (ii) wskazanie, które racje uznał za nadrzędne; (iii) korektę decyzji na podstawie tego rozstrzygnięcia w sposób publicznie rekonstruowalny. Jeżeli identyfikację konfliktu, wybór reguły pierwszeństwa lub uzasadnienie zapewniają wyłącznie ludzie i procedury organizacyjne, realizowany jest WI, nie zaś H3 systemu. Funkcjonalna (samo)świadomość może wspierać te procesy, jeżeli dostarcza modelu własnego udziału w podejmowaniu decyzji i pozwala odróżniać działania systemu od zdarzeń zewnętrznych. Nie stanowi jednak samodzielnej definicji H3. W niniejszej analizie (samo)świadomość ujmuję funkcjonalnie, a nie fenomenologicznie: jako „model siebie” (ang. *self-model*) w sensie metzingerowskim oraz „postawę intencjonalną” (ang. *intentional stance*) w ujęciu Dennetta – ramy opisu mechanizmów poznawczych i decyzyjnych, bez tez o *qualiach* czy subiektywnym doświadczaniu.

Na tym tle (samo)świadomość funkcjonalna ma węższy zakres niż introspekcyjna. Oznacza zdolność tworzenia i aktualizowania roboczego modelu siebie, który obejmuje: (i) granice „ja” względem świata, (ii) własne stany intencjonalne (plany, zamiary), (iii) istotne uprawnienia i zakazy wynikające z norm oraz (iv) wpływ tych stanów na działanie. Model ten musi być włączony w układ sterujący: nie chodzi o deklaratywne użycie zaimka „ja”, lecz o mechanizm, który przekłada „co robię” i „dlaczego” na „co zrobię dalej” w świetle rozpoznanych racji. W tym sensie (samo)świadomość

funkcjonalna może wspierać K1–K4: wzmacnia responsywność na racje, umożliwia sprzężenie refleksji z wykonaniem i stabilizuje tożsamość praktyczną w czasie.

W perspektywie „drugiej osoby” konsekwencje H3 są bezpośrednie. Stephen Darwall twierdzi, że moralność polega na wzajemnym kierowaniu roszczeniami: mamy prawo domagać się uzasadnienia, a inni mają obowiązek udzielenia odpowiedzi, kontroli i korekty. Bez mechanizmów udostępniania treści decyzyjnych i ich kontroli kierowanie roszczeń do systemu ma co najwyżej charakter pośredni i musi zostać osadzone w procedurach instytucjonalnych (Darwall, 2006, s. 65–90, 87–118). R. Jay Wallace dodaje, że zobowiązania mają strukturę wzajemności – są „winne komuś” – i to właśnie ta adresowalność nadaje im siłę normatywną (Wallace, 2019, s. 56–57). W wersji H3 nie chodzi zatem o uczynienie maszyny pełnym uczestnikiem praktyk „drugiej osoby”, lecz o określenie, kiedy techniczny system, już minimalnie adresowalny na mocy H1, może uczestniczyć w bardziej wymagających praktykach uzasadniania i korekty właściwych silnej rozliczalności. Włączenie całego układu społeczno-technicznego do praktyk kontroli i rozliczania należy natomiast do WI. H3 wymaga mechanizmów pozwalających samemu systemowi monitorować własny proces decyzyjny, rozpoznawać konflikt przesłanek normatywno-decyzyjnych, jawnie wskazywać regułę pierwszeństwa i przedstawiać audytowalne uzasadnienie korekty, bez przesądzania o samoświadomości w sensie fenomenologicznym.

Włączenie H3 do kompatybilistycznego ujęcia Johna Martina Fischera i Marka Ravizy doprecyzowuje wymóg kontroli. Odpowiedzialność praktyczna nie wymaga libertariańskiej „wolności od przyczyn”, lecz mechanizmu responsywnego na racje, który sprawca uznaje za własny, tzn. internalizuje i traktuje jako źródło swoich czynów (Fischer & Ravizza, 1998, s. 62–91, 207–239). A-świadomość i (samo)świadomość funkcjonalna mogą pełnić rolę interfejsu „posiadania” treści decyzyjnych i racji jako własnych: to w nich sprawca rozpoznaje, że działa „z tego powodu”, dzięki czemu może przyjąć wagę uzasadnienia i korekty. Bez takich mechanizmów przypisanie czynu systemowi AI ma charakter pośredni i wymaga wyraźnego osadzenia w procedurach instytucjonalnych oraz w odpowiedzialności projektantów, operatorów lub organizacji.

Można tu zgłosić wątpliwość: nawet jeśli system spełnia warunek A-świadomości, „ostateczna” odpowiedzialność moralna w sensie zasługiwania (ang. *desert*) może być pozorna. Derk Pereboom utrzymuje, że nikt – ani człowiek, ani maszyna – nie jest „ostatecznym sprawcą” swoich czynów, ponieważ każdy sprawca działa w sieci przyczyn i okoliczności, które znajdują się poza jego pełną kontrolą (Pereboom, 2001, s.

123–157, 174–179). Hipoteza H3 jest z tym zastrzeżeniem zgodna: dotyczy odpowiedzialności praktycznej, opartej na adresowalności i metakontroli, a nie metafizycznego autorstwa bez przyczyn. Innymi słowy, sztuczny podmiot spełniający H3 może być w sensie funkcjonalnym wzywany do odpowiedzi, korygowany i sankcjonowany, nawet jeśli rezygnuje się z kategorii zasługi.

Niezależnie od sporu o zasługę istnieje ryzyko racjonalizacji epifenomenalnej. Modele językowe potrafią tworzyć wyjaśnienia bez realnego związku przyczynowego z wyborem, dopasowując narrację po fakcie. Z perspektywy H3 takie wyjaśnienia nie spełniają funkcji metakontroli: nie stanowią kontrolowalnego dostępu do racji decyzji, lecz konfabulację. Dlatego wymóg sprzężenia deliberacji z działaniem (K3) traktuję jako mechanizm wspierający H3: uznane racje muszą rzeczywiście wpływać na zachowanie; to odróżnia dostęp do treści decyzyjnych od konfabulacji (Levy, 2014, s. 112–117).

Wprowadzenie H3 ma także wymiar epistemiczny. Mechanizmy udostępniania treści decyzyjnych wyznaczają granice, w których można racjonalnie oczekiwać sensownych wyjaśnień od systemu i ufać, że nie są one konfabulacją. Jeżeli agent wskazuje scenariusze kontrfaktyczne, w których alternatywna racja prowadziłaby do innej decyzji, istnieje podstawa, by sądzić, że wyjaśnienia są częścią mechanizmu wyboru, a nie jedynie narracją po fakcie. Jeżeli system konsekwentnie zgłasza konflikty wewnętrzne i rozstrzyga je zgodnie z jawnymi regułami pierwszeństwa, rośnie pewność, że działa na podstawie treści objętych kontrolą metapoznawczą, a nie ukrytych heurystyk.

Z powyższych przesłanek wynika zestaw kryteriów diagnostycznych H3; głównym zagrożeniem jest ryzyko symulowanej (samo)świadomości. Model może mówić w pierwszej osobie, deklarować stany mentalne i podawać „zgrabne” racje, podczas gdy wybory powstają gdzie indziej – w niejawnych heurystykach lub mechanizmach wzmocnienia. Diagnoza H3 powinna zatem rozstrzygać, czy: (i) deklarowane racje kontrfaktycznie zmieniają decyzję („gdyby zaszło R, wybrałbym Y zamiast X”), (ii) system wstrzymuje działanie w obliczu konfliktu (inhibicja) oraz (iii) istnieje publicznie rekonstruowalny i intersubiektywnie ocenialny ciąg zapisów łączących rozpoznanie powodu z korektą planu (np. rejestr zmian). Pozytywna odpowiedź na wszystkie trzy pytania stanowi podstawę do uznania, że A-świadomość i (samo)świadomość funkcjonalna wspierają metakontrolę, a nie jedynie ją pozorują.

Na tym tle wraca kwestia adresata powinności moralnych. Jeśli od AI wymaga się jedynie A-świadomości i funkcjonalnej (samo)świadomości, uzasadnia się co najwyżej mechanizmy wspierające adresowalność i silną rozliczalność, nie przesądzając o prawach

do dobrostanu. Te wydają się wymagać świadomości fenomenalnej albo przynajmniej funkcjonalnego odpowiednika interesu własnego, w którym jakość doświadczenia ma znaczenie motywacyjne. P-świadomość, o ile w ogóle będzie technicznie osiągalna, pozostaje więc kwestią „pełniejszej” podmiotowości moralnej i praw AI jako adresata roszczeń. Na poziomie minimalnym wystarcza H1 jako minimalna adresowalność proceduralna; A-świadomość i funkcjonalna (samo)świadomość mogą natomiast wspierać metakontrolę potrzebną dla H3 (Block, 1995, s. 227–247; Levy, 2014, s. 109–131).

W świetle przedstawionych kryteriów diagnostycznych można doprecyzować rolę H3 na tle wcześniejszych ustaleń. Zależności te porządkują strukturę silnej rozliczalności i zapobiegają myleniu samej zdolności generowania werbalnych wyjaśnień z faktycznym posługiwaniem się racjami przy sterowaniu działaniem.

1. Z K1 H3 łączy wymóg, by refleksyjny dostęp do treści umożliwiał nie tylko rozpoznawanie reguł, lecz także ich rewizję w świetle nowych racji; dzięki temu korekta jest częścią mechanizmu sprawczego, a nie zewnętrzną ingerencją.
2. Z K2 (responsywność na racje) łączy ujawnianie i selekcję treści: racje stają się globalnie dostępne dla procesu decyzyjnego i mogą kształtować wynik decyzji.
3. Z K3 łączy mechanizmy metakontroli: dostęp do treści deliberacji może uruchamiać hamowanie, zmianę planu i ponowne ważenie priorytetów, czyniąc z racji przyczynę korekt.
4. Z K4 łączy stabilność diachroniczna technicznego systemu: spójność podstaw decyzji i korekt w czasie pozwala ocenić ciągłość przypisania odpowiedzialności. Stabilność ról, procedur i praktyk organizacyjnych należy natomiast do WI i nie może być zaliczana jako zdolność systemu.

Całość uzupełnia H2, pełniąc funkcję stabilizującą: wzmacnia sygnały moralne i utrwała korekty, zwiększając odporność na krótkoterminowe zyski oraz presję wynikającą ze złożoności środowiska decyzyjnego. Sama jednak nie gwarantuje silnej rozliczalności, jeżeli nie towarzyszy jej metakontrola i jawne rozstrzyganie konfliktów racji właściwych H3 (Bratman, 1987, s. 64–71; Levy, 2014, s. 109–113).

H3 nie należy do warunków koniecznych progu minimalnego. Minimalną adresowalność proceduralną wyznacza H1, natomiast H3 oznacza przejście do silnej

rozliczalności: wymaga metakontroli, jawnego ważenia racji oraz możliwości publicznego zakwestionowania rozstrzygnięcia. System przekracza próg silnej rozliczalności wtedy, gdy potrafi: (i) zidentyfikować konflikt racji, (ii) uczynić go treścią dostępną dla mechanizmów kontroli, (iii) wskazać rację lub regułę nadrzędną oraz (iv) przedstawić audytowalne uzasadnienie dokonanej korekty. W tym sensie H3 nie domyka progu minimalnego, lecz wzmacnia model odpowiedzialności poprzez metakontrolę i jawne rozstrzyganie konfliktów racji.

Podsumowanie

Rozdział 3 ustala kryteria operacyjno-diagnostyczne potrzebne do oceny układu H1–H3. Zamiast przyjmować silne założenia ontologiczne, odwołuję się do ramy poziomów abstrakcji (LoA) oraz triady: interaktywność, autonomia, adaptacyjność, nadając priorytet jawności normatywnej rozumianej jako zdolność poddawania decyzji ocenie i korekcie w świetle zasad. H1 wyznacza minimalną adresowalność proceduralną i podstawę słabej odpowiedzialności proceduralnej; K1, K2 i K3 pomagają sprawdzić, czy racje są rozpoznawane, wpływają na decyzję, prowadzą do korekty i pozostają audytowalne.

Filozoficznie uzasadnioną ramę progu minimalnego wyznacza „agent jawny” w typologii J. Moora, osadzony w metodologicznej ramie poziomów abstrakcji (LoA). Różni się on jakościowo od „agenta implicytnego”, ponieważ potrafi wskazać racje swojego działania i skorygować je w obliczu konfliktu norm; nie wymaga się natomiast świadomości fenomenalnej ani woli w silnym sensie. Połączenie „agenta jawnego” z perspektywą LoA pozwala oceniać decyzje moralne bez założeń metafizycznych, ale nie zastępuje struktury H1–H3 ani warunku instytucjonalnego WI.

Rdzeniem normatywnym progu jest H1 – minimalna adresowalność proceduralna. K1, K2 i K3 pełnią funkcję kryteriów oceny H1: K1 sprawdza możliwość korekty i audytowalności, K2 – responsywność na racje, a K3 – przyczynowe powiązanie deliberacji z działaniem oraz odróżnienie go od racjonalizacji po fakcie. H2 dotyczy afektywności funkcjonalnej jako mechanizmu stabilizującego, natomiast K4 odrębnie ocenia stabilność diachroniczną i wiarygodność profilu działania; spełnienie K4 nie implikuje H2. H3 oznacza silną rozliczalność technicznego systemu dzięki kontroli metapoznawczej, jawnej regule pierwszeństwa i uzasadnieniu drugiego rzędu. Warunek instytucjonalny WI określa natomiast, kiedy model H1–H3 może zostać zastosowany w praktykach publicznych i organizacyjnych; nie jest to H4 ani osobny poziom agenta. Tym

samym rozdział 3 domyka poziom kryteriów operacyjno-diagnostycznych. Następny rozdział przechodzi od pytania, jak diagnozować spełnienie H1–H3, do pytania, jak normatywnie uzasadnić te funkcje poprzez PH1–PH3 w warunkach pluralizmu etycznego.

Rozdział 4. Pluralistyczne ramy etyczne: deontologia, cnoty, konsekwencjalizm – porównanie i implikacje normatywne

Wprowadzenie

Rozdział 4 rozwija normatywne kryteria oceny działań przypisywanych systemom AI w warunkach pluralizmu etycznego. Nie ma on rozstrzygać sporu między deontologią, etyką cnót i konsekwencjalizmem, lecz ułożyć te tradycje w spójną ramę rozliczalności decyzji podejmowanych w układach społeczno-technicznych. Perspektywa „drugiej osoby” podkreśla, że moralność ma postać roszczeń i żądań uzasadnienia między adresatami norm: adresaci norm domagają się racji, które mogą zostać przyjęte przez uprawnionych innych (Darwall, 2006, s. 3).

Call the second-person standpoint the perspective you and I take up when we make and acknowledge claims on one another's conduct and will. This might be explicit, in speech, as with the performatives J. L. Austin botanized—demanding, reproaching, apologizing, and so on—or only implicit, in thought, as with Strawsonian reactive feelings like resentment and guilt (Austin 1975; Strawson 1968).² But whether explicit and voiced—„You talkin' to me?”—or only implicit and felt, as in a resentful sulk, the I-you-me structure of reciprocal address runs throughout thought and speech from the second-person point of view (Darwall, 2006, s. 3).

Teza przewodnia niniejszego rozdziału głosi, że adekwatną podstawą oceny działań przypisywanych systemom AI jest pluralizm hierarchiczny rozumiany jako porządek funkcji normatywnych, a nie jako synteza doktryn etycznych. PH1–PH3 (nie mylić z hipotezami H1–H3 z rozdziałów 1 i 3) nie oznaczają tu odrębnych teorii moralności, lecz trzy komplementarne progi wymagań normatywnych, których spełnienie podlega ocenie przy użyciu odpowiednich kryteriów operacyjnych. W tym ujęciu PH1 wyznacza deontyczne granice dopuszczalności oraz minimalne warunki adresowalności normatywnej. PH2, inspirowana etyką cnót, dotyczy stabilizacji praktyk i motywacji w układach instytucjonalnych odpowiedzialnych za projektowanie i wdrażanie systemów. PH3, w duchu konsekwencjalistycznym, porządkuje konflikty racji w obrębie opcji uznanych wcześniej za dopuszczalne. Kluczowe rozróżnienie dotyczy celu tej konstrukcji. Nie chodzi tu o redukcję trzech tradycji do jednej doktryny, lecz o

ustanowienie porządku publicznego uzasadniania (LexOrder) oraz bezstronnego doboru celów i priorytetów (FairAim). Aby uniknąć nieporozumień, rozróżniam trzy poziomy analizy sztucznej podmiotowości moralnej: (1) poziom zachowań przypisywanych systemowi, (2) poziom podmiotu odpowiedzi w układzie społeczno-technicznym, obejmujący projektantów, operatorów i instytucje oraz (3) poziom procedur weryfikacji i kontroli. Dzięki temu PH1–PH3 nie są utożsamiane z tezą o pełnej, metafizycznej podmiotowości artefaktu, lecz odnoszą się do warunków normatywnej rozliczalności w praktykach społecznych.

W aktualnej strukturze dysertacji PH1–PH3 pełnią funkcję etyczno-normatywnej warstwy uzasadnienia modelu H1–H3, lecz nie są jego prostymi odpowiednikami. PH1 doprecyzowuje normatywny sens minimalnej adresowalności H1 przez wskazanie granic dopuszczalności i publicznego uzasadniania. PH2 uzasadnia normatywne znaczenie stabilności i wiarygodności praktyk przypisywania odpowiedzialności, nie utożsamiając ich z funkcjonalnym mechanizmem wartościowania H2. PH3 porządkuje konflikty racji w obrębie opcji dopuszczalnych i normatywnie uzasadnia wymóg jawnego rozstrzygnięcia, który w H3 łączy się z kontrolą drugiego rzędu technicznego systemu; sama nie jest jednak taką kontrolą.

WI oznacza w tym kontekście warunek instytucjonalnego zastosowania modelu H1–H3, a nie H4 ani osobny poziom podmiotowości.

Istotnym założeniem metodologicznym jest analityczne ujęcie responsywności wobec racji. Na poziomie H1 zdolność tę przypisuje się technicznemu systemowi, gdy rozpoznane racje wpływają na decyzję, mogą prowadzić do jej korekty, a system udostępnia wystarczające i wiarygodne podstawy rozstrzygnięcia. Na poziomie WI ocenia się natomiast, czy układ społeczno-techniczny zapewnia procedury audytu, zakwestionowania i korekty. Afektywność funkcjonalna H2 oznacza mechanizm wartościowania w systemie, który moduluje wybór, uwagę lub hamowanie; nie jest ani stanem fenomenalnym artefaktu, ani stabilnością praktyk instytucjonalnych. Ta ostatnia należy do PH2 i WI.

W dalszej części rozdziału PH1–PH3 są konsekwentnie traktowane jako trzy odrębne funkcje publicznej rozliczalności uzasadnień. PH1 określa warunki dopuszczalności i adresowalności; PH2 dotyczy trwałości i wiarygodności stosowania

reguł w warunkach niepewności; PH3 umożliwia jawne rozstrzyganie konfliktów racji, przy czym odnosi się wyłącznie do opcji spełniających wcześniej warunki PH1 i PH2. Struktura wyводу odzwierciedla tę logikę: najpierw rekonstruuje funkcjonalny podział ról w trzech tradycjach etycznych (4.1), następnie precyzuje operacyjne kryteria weryfikacji PH1–PH3 (4.2), dalej analizuje zasadę porządku leksykalnego (LexOrder) w projektowaniu sztucznej moralności (4.3), a wreszcie przedstawiam Zasadę Stopniowej Weryfikowalności (ZSW) oraz Minimum Publicznego Uzasadnienia (MPU) jako zabezpieczenia przed zgodnością pozorną (4.4). W konsekwencji pluralizm pełni tu funkcję ramy normatywnej porządkującej praktyki uzasadniania, a nie projektu metaetycznej unifikacji. PH1–PH3 należy więc czytać jako warstwę uzasadnienia etycznego modelu H1–H3, a nie jako równoległy lub konkurencyjny zestaw hipotez.

4.1. Pluralizm etyczny: etyka cnót, deontologia, konsekwencjalizm, modele hybrydowe

Punktem wyjścia niniejszego podrozdziału jest teza, że sporu między deontologią, etyką cnót i konsekwencjalizmem nie da się rozstrzygnąć bez istotnych kosztów dla analizy sztucznej podmiotowości moralnej. W odniesieniu do działań przypisywanych systemom AI potrzebne są jednocześnie: jasne granice dopuszczalności i adresowalności norm, stabilność praktyk ich stosowania oraz zdolność jawnego rozstrzygania konfliktów wartości. Z tego względu przyjmuję pluralizm hierarchiczny. Obejmuje on: (PH1) warstwę deontyczną, wyznaczającą porządek zasad i minimalne wymogi adresowalności; (PH2) warstwę aretologiczną, stabilizującą praktyki poprzez cnoty rozumiane jako standardy instytucjonalne, a nie jako „emocje maszyny”; oraz (PH3) funkcję jawnego porządkowania i rozstrzygania konfliktów racji w warunkach niepewności moralnej. Relacja między tymi warstwami ma charakter priorytetu (LexOrder), a nie doktrynalnej syntezy (Kant, 2002, AA 4:421–434; Scanlon, 1998, s. 3–5; Darwall, 2006, s. 65–90; Arystoteles, 2012, EN II 6, 1106b36–1107a2; Hagendorff, 2022, art. 55; Constantinescu i in., 2021, s. 803–814; Ross, 2002, s. 19–20; MacAskill, Bykvist & Ord, 2020, s. 1–4, 105–134). Proponowane rozwiązanie nie łączy mechanicznie trzech rodzajów etyk, lecz przypisuje im odmienne role i porządkuje ich wzajemne relacje na poziomie uzasadnień (por. w 4.3 regułę LexOrder). Dzięki temu unika dwóch skrajności: formalizmu reguł oderwanych od warunków ich rzetelnego stosowania oraz arbitralności rozstrzygnięć pozbawionych jasno określonych kryteriów pierwszeństwa.

4.1.1. Deontologia

Deontologia w ujęciu Kanta wyznacza minimalne warunki postępowania moralnego. Imperatyw kategoryczny przyjmuje u Kanta różne sformułowania, z których najczęściej rekonstruuje się formułę prawa powszechnego oraz formułę człowieczeństwa jako celu samego w sobie: wymóg, by maksyma mogła obowiązywać jako prawo powszechne oraz nakaz traktowania każdej osoby zawsze jako celu, nigdy wyłącznie jako środka (Kant, 2002, AA 4:421–434). Z perspektywy sztucznej podmiotowości moralnej kryteria te pełnią dwie funkcje. Po pierwsze, z góry wyznaczają granice, w ramach których w ogóle można różnicować dopuszczalne rozwiązania: nie wszystko, co maksymalizuje pożądany wynik, jest dozwolone, jeśli narusza równość osób jako adresatów norm. Po drugie, uzasadniają wymóg adresowalności: skoro podmiot ma prawo wymagać dających się publicznie ocenić racji, musi istnieć standard rozpoznawania sytuacji, w której został potraktowany wyłącznie jako środek (Darwall, 2006, s. 65–90; Scanlon, 1998, s. 3–5). Wynika stąd istotne rozróżnienie między przypisywalnością czynu a ponoszeniem odpowiedzialności wobec kogoś w relacji „drugiej osoby” (Wallace, 1998, s. 51–83). Jednocześnie klasyczna deontologia, ujęta bez dogmatyzmu, przyjmuje istnienie obowiązków wstępnych (*prima facie*), które mogą wchodzić ze sobą w kolizję i wymagają odstępstw od reguł (Ross, 2002, s. 19–20). Taka elastyczność deontologii ma znaczenie filozoficzne: pozwala zachować granice bez wprowadzania zasad o charakterze bezwzględny i przygotowuje miejsce dla roli PH3.

Formułę PH1 można zwięźle zakotwiczyć w kantowskim nakazie traktowania człowieczeństwa „zawsze zarazem jako celu, nigdy tylko jako środka” (Kant, 2002, s. 73). Cytat wzmacnia PH1 jako granicę instrumentalizacji, lecz nie ustanawia dodatkowego poziomu modelu H1–H3 + WI.

Rossowska koncepcja obowiązków *prima facie* wskazuje, że różne relacje moralne mogą stanowić podstawę różnych obowiązków, dlatego konflikt obowiązków wymaga ważenia racji, a nie redukcji do jednego kryterium; w jej ujęciu „each of these relations is the foundation of a *prima facie* duty” (Ross, 2002, s. 19). W strukturze PH1–PH3 wspiera to pluralizm racji przy zachowaniu pierwszeństwa granic deontycznych.

4.1.2. Etyka cnót

Etyka cnót wiąże reguły z działaniem. Arystotelesowska *phronesis* nie jest kwestią upodobania, lecz dyspozycją do właściwego rozeznania w działaniu i doboru właściwych środków (Arystoteles, 2012, EN II 6, 1106b36–1107a2). Źródłowo cnota jest

dyspozycją ludzkiego charakteru, wypracowaną przez przyzwyczajenie. Wiąże się ściśle z ludzkimi emocjami, pragnieniami, nawykami oraz zdolnością do osądu moralnego. W niniejszej pracy przyjmuję nieantropomorficzną interpretację cnót. Odnoszę ją przede wszystkim do standardów praktyk i instytucji, które zapewniają wiarygodne stosowanie reguł: do prawdomówności i rzetelności w ujawnianiu racji, do bezstronności rozumianej jako norma proceduralna ograniczająca stronniczość decyzji, do troski o adresatów znajdujących się na obrzeżach systemu jako wymogu ochrony relacyjnej, a także do roztropności ujmowanej jako dyspozycja metaregulacyjna, porządkująca stosowanie zasad w warunkach niepewności (Hagendorff, 2022, art. 55). Ujęcie to jest zgodne z badaniami nad „cnotami dianoetycznymi”, czyli dyspozycjami intelektualnymi, które kierują osądem i rozumieniem sytuacji. Bez zdolności rozpoznawania kontekstu i uczenia się jego moralnej wagi nawet najlepsze zasady ulegają zubożeniu i przekształcają się w formalizm (Constantinescu i in., 2021, s. 803–814). W odniesieniu do sztucznej podmiotowości moralnej chodzi zatem nie o przypisywanie artefaktom „charakteru”, lecz o ocenę, czy praktyki wspierające ich działanie wykazują stabilne dyspozycje umożliwiające rzeczywiste, a nie jedynie pozorne, uzasadnianie decyzji. Perspektywa ta pozostaje zgodna z nurtem relacyjnym w etyce technologii, który uwrażliwia na krzywdy ustrukturyzowane i zależne od relacji (Birhane, 2021, art. 100205). W sensie filozoficznym wniosek jest prosty: stosowanie reguł nie zachodzi w oderwaniu od praktyk; istnieją jedynie reguły osadzone w praktykach, które mogą wzmacniać albo osłabiać ich treść normatywną.

4.1.3. Konsekwencjalizm

Konsekwencjalizm bywa w uproszczonych ujęciach redukowany do postulatu maksymalizowania dobra, jednak jako rodzina teorii obejmuje stanowiska istotnie zróżnicowane: różniące się rozumieniem tego, co podlega ocenie (korzyści, dobrostan, wartości, unikanie krzywdy), sposobem jego porównywania oraz zasadami dystrybucji, przy czym nie wszystkie jego odmiany mają charakter maksymalizujący. Takie uproszczenie jednak uniemożliwia jego pluralistyczne włączenie. Na poziomie PH3 wykorzystuję ocenę skutków oraz ryzyka błędu między teorią a praktyką. Nie chodzi jednak o odtworzenie konsekwencjalizmu jako pełnej teorii słuszności, lecz o wyodrębnienie funkcji rozstrzygania konfliktów normatywnych w sytuacjach niepewności. Ograniczenie to wynika z przyjętego pluralizmu hierarchicznego: deontologia wyznacza granice moralnej dopuszczalności (PH1), etyka cnót stabilizuje

wiarygodne stosowanie reguł w praktyce (PH2), natomiast konsekwencjalizm zostaje zastosowany dopiero na etapie porządkowania konfliktów i niepewności w obrębie opcji już normatywnie dozwolonych. W tej roli konsekwencjalizm znajduje zastosowanie w dwóch wymiarach. Po pierwsze, w wymiarze dystrybucyjnym, w którym szkody i korzyści nie mają jednolitej miary, a racjonalny priorytet przysługuje ograniczaniu poważnych strat przed sumowaniem drobnych korzyści. Po drugie, w wymiarze epistemicznym, w którym wobec trwałej niezgody między rozsądnymi teoriami decyzje należy podejmować w warunkach niepewności moralnej, minimalizując ryzyko poważnego błędu między teoriami zamiast absolutyzować jedną z nich (MacAskill i in., 2020, s. 1–4; 105–134). W przyjętym w rozprawie modelu konsekwencjalizm nie konkuruje z deontologią ani etyką cnót, lecz opiera się na wynikach ich zastosowania, dostarczając mechanizm rozstrzygania kolizji potrzebny dla praktycznej rozliczalności.

4.1.4. LexOrder

Aby powyższe podejście nie prowadziło do niespójności, konieczna jest zasada porządku. Określam ją jako LexOrder (wprowadzoną w 4.1.4 i rozwiniętą w 4.3): (I) w pierwszej kolejności obowiązuje warstwa deontyczna i adresowalność (PH1), (II) następnie stabilizacja praktyk na gruncie etyki cnót (PH2), (III) w końcowym etapie – jawne porządkowanie i rozstrzyganie konfliktów racji w warunkach niepewności (PH3). Taki układ pozwala uniknąć dwóch błędnych założeń. Po pierwsze, przekonanie, że wystarczy zestaw reguł: bez dyspozycji i jasno określonych priorytetów reguły nie utrzymują się w działaniu, co prowadzi do luki normatywnej (Moor, 2006, s. 18–21). Po drugie, przekonanie, że wystarczy sama wrażliwość: bez deontycznych ograniczeń i jawnej hierarchii racji, wrażliwość staje się dowolnym osądem, którego nie można poddać ocenie w relacji z drugą osobą (Scanlon, 2008, s. 1–4; Wallace, 1998, s. 51–83). LexOrder umożliwia także rozróżnienie trzech rodzajów błędów: błędu dopuszczalności (naruszenie zakazu lub niespełnienie warunku adresowalności), błędu wiarygodności (pozorna zgodność wynikająca z braku rzetelnych dyspozycji praktyk) oraz błędu porównywalności (dowolne ważenie racji, które pomija znaczenie poważnych strat).

Na tym tle można ocenić klasyczne modele hybrydowe w etyce maszyn. To, co bywa nazywane „hybrydą”, często jest jedynie zbiorem kryteriów bez określonego porządku pierwszeństwa (Allen, Smit & Wallach, 2005, s. 149–155; Anderson & Anderson, 2007, s. 15–26; Wallach & Allen, 2009). Taki eklektyzm pomija realne

konflikty zamiast je rozstrzygać. Proponowana tutaj hybryda ma charakter integracyjny i hierarchiczny: przypisuje poszczególnym podejściom określone role oraz porządkuje je w porządku leksykalnym (LexOrder), wyznaczając kolejność uzasadnień od deontologicznych granic dopuszczalności, przez aretologiczną stabilizację praktyk, po konsekwencjalistyczne rozstrzyganie konfliktów. Różnica jest zasadnicza. Iason Gabriel wskazuje potrzebę sprawiedliwego doboru zasad dla AI w warunkach pluralizmu wartości (Gabriel, 2020, s. 432–433). Rozwijając tę przesłankę, w niniejszej pracy proponuję FairAim jako wymóg bezstronnego ustalania zasad i priorytetów rozstrzygania; nie jest to termin ani gotowy model przejęty od Gabriela. Ujęcie proceduralne nie jest relatywizmem; gwarantuje równość adresowalności i możliwość krytyki w warunkach rozsądnej niezgody, zwłaszcza z perspektywy podmiotów marginalizowanych (Waelen, 2022, art. 9; Coeckelbergh, 2025b, s. 5527–5530). W tym sensie pluralizm hierarchiczny ma także znaczenie ustrojowe: ogranicza władzę, która bez publicznego uzasadnienia i bezstronnych procedur legitymizacyjnych może posługiwać się językiem „zasad” w sposób nieuprawniony.

Aby pluralizm był użyteczny w praktyce, proponuję macierz 3×3 , czyli prosty arkusz oceny, który krzyżuje trzy paradygmaty z trzema funkcjami PH1–PH3. Nie jest to algorytm, lecz narzędzie pojęciowe: w konkretnym projekcie pozwala szybko wskazać, których funkcji brakuje. Ponieważ PH1–PH3 opisują funkcje uzasadniania, a nie elementy jednej, zamkniętej teorii moralnej, mogą być realizowane w różnych rozwiązaniach instytucjonalnych. Przykładowo, standardy projektowania i audytu mogą koncentrować się na PH1 i PH2, czyli na granicach dopuszczalności oraz wiarygodności praktyk, natomiast procedury odwoławcze i polityki alokacji dóbr rzadkich mogą akcentować PH1 i PH3, a więc granice dopuszczalności oraz jawne rozstrzyganie konfliktów i niepewności. Model nie wymaga, aby jeden mechanizm spełniał wszystkie funkcje naraz. Wymaga natomiast, aby w ramach całego układu socjotechnicznego żadna z tych funkcji nie była trwale pomijana, jeśli ocena ma pozostać publicznie rozliczalna. W tym ujęciu deontologia zwykle najsilniej wspiera PH1 (granice i adresowalność), a słabiej PH2, o ile nie zostanie uzupełniona o komponent aretologiczny. Etyka cnót stabilizuje PH2, lecz zakłada PH1 jako warunek wstępny oraz potrzebuje PH3 jako zestawu reguł rozstrzygania kolizji. Konsekwencjalizm porządkuje PH3, ale wymaga PH1, aby uniknąć rozstrzygnięć przekraczających granice dopuszczalności oraz PH2, aby nie pozostać na poziomie czysto teoretycznych kalkulacji. Macierz pozwala też odróżnić hipotezy teoretyczne od kryteriów oceny: PH1–PH3 nie są „składnikami

psychologicznymi” podmiotów, lecz warunkami normatywnej rozliczalności, dzięki którym mówienie o sztucznej podmiotowości moralnej zachowuje sens (Floridi & Sanders, 2004, s. 349–379).

Perspektywa „drugiej osoby” doprecyzowuje sens adresowalności: ocena moralna ma postać roszczeń i żądań uzasadnienia kierowanych do podmiotu odpowiedzi, a nie jedynie opisu skutków. W tym znaczeniu PH1 wymaga racji, które można przedstawić adresatom i poddać krytyce, co wzmacnia rozliczalność praktyk przypisywania odpowiedzialności w odniesieniu do AI (Darwall, 2006, s. 65–90; Scanlon, 1998, s. 3–5).

Wobec pluralizmu hierarchicznego powracają trzy typowe zastrzeżenia. Po pierwsze, zarzut „ukrytego monizmu”, według którego LexOrder miałby faworyzować jedną teorię. Odpowiedź jest funkcjonalna: PH1 wyznacza próg dopuszczalności, PH2 stabilizuje praktyki, a PH3 rozstrzyga konflikty wyłącznie w obrębie tego, co uznano wcześniej za dopuszczalne. Po drugie, zarzut antropomorfizacji, jakoby etyka cnót wymagała „charakteru” maszyny. W tym ujęciu PH2 odnosi się do cnót praktyk instytucjonalnych, a nie do stanów mentalnych artefaktu. Po trzecie, zarzut „konsekwencjalizmu tylnymi drzwiami”, według którego PH3 miałby usprawiedliwiać naruszenia PH1. PH3 nie unieważnia PH1, lecz służy porównywaniu opcji, które przeszły próg dopuszczalności oraz jawnemu ujmowaniu niepewności moralnej. W relacji do H3 warstwa ta dostarcza normatywnego uzasadnienia wymogu jawnego porządkowania konfliktów, lecz nie stanowi kontroli drugiego rzędu technicznego systemu (Ross, 2002, s. 19–20; MacAskill i in., 2020, s. 1–4; Arystoteles, 2012, EN II 6, 1106b36–1107a2; Hagendorff, 2022, art. 55; Constantinescu i in., 2021, s. 803–814).

Z przedstawionej rekonstrukcji wynikają konsekwencje dla ujęcia sztucznej podmiotowości moralnej. Po pierwsze, jej minimalny sens jest normatywny, a nie psychologiczny: chodzi o status podmiotu adresowalnego, zobowiązanego do przedstawiania publicznie ocenialnych racji oraz do kontrfaktycznej korekty decyzji (Darwall, 2006, s. 65–90; Scanlon, 1998, s. 3–5). Po drugie, przypisanie tak rozumianego statusu wymaga podmiotu odpowiedzi – identyfikowalnej osoby lub instytucji, która utrzymuje tożsamość odpowiedzialnego za działania oraz uzasadnia je (Wallace, 1998, s. 51–83). Po trzecie, w proponowanym tu modelu rozliczalność wymaga uporządkowania uzasadnień w trzech krokach (PH1–PH3) oraz bezstronnego doboru zasad i priorytetów (FairAim). Jest to własne rozwinięcie przesłanki proceduralnej, a nie odtworzenie gotowej hierarchii z literatury (por. Gabriel, 2020, s. 432–433; Coeckelbergh, 2025b, s.

5527–5530). Po czwarte, wyjaśnialność ma znaczenie normatywne. Jonathan Adams broni włączenia zasady wyjaśnialności do pryncypializmu w medycynie, wiążąc ją z odpowiedzialnym stosowaniem AI i rozliczalnością ludzi oraz instytucji (Adams, 2023, s. 618, 620). W niniejszej dysertacji dostęp do wystarczających i wiarygodnych podstaw decyzji jest jednym z warunków H1, a nie samodzielnym dowodem proceduralnej adresowalności systemu. Jest to rozstrzygnięcie własnego modelu, nie teza przejęta od Adamsa. W sporach o dostęp do ograniczonych zasobów i dóbr, deontologia ogranicza lub zapobiega dyskryminacji i traktowaniu instrumentalnemu, etyka cnót wymaga bezstronności i troski wobec najsłabszych, a porządek rozstrzygania przyznaje pierwszeństwo unikaniu poważnych szkód przed osiąganiem drobnych korzyści, przy pełnej przejrzystości uzasadniania przyjętych priorytetów (Kant, 2002, AA 4:429–434; Arystoteles, 2012, EN II 6, 1106b36–1107a2; MacAskill i in., 2020, s. 105–134).

Można zapytać, czy tak rozumiany pluralizm nie osłabia odpowiedzialności, zwłaszcza w systemach złożonych. Odpowiedź prowadzi do klasycznego problemu „luki odpowiedzialności” w systemach uczących się (Matthias, 2004, s. 176–182). Gdy nie wiadomo, kto jest podmiotem udzielającym racji, rozliczalność traci na znaczeniu. W proponowanej hybrydzie odpowiedź jest dwutorowa. Po pierwsze, PH1 pozwala identyfikować naruszenia: złamanie zakazu lub niespełnienie kryterium adresowalności, niezależnie od złożoności procesu. Po drugie, PH2 i PH3 wymagają wskazania dyspozycji oraz hierarchii racji, które uzasadniają wybór „A zamiast B”, a tym samym przywracają odpowiedzialność wobec drugiej osoby. W tym ujęciu pluralizm hierarchiczny nie rozmywa odpowiedzialności, lecz ją wzmacnia, dostarczając kryteriów odróżnienia błędu dopuszczalności od błędu wiarygodności oraz kryteriów porównywalności. Korzyść filozoficzna jest zatem podwójna: po pierwsze, moralność pozostaje oparta na relacji z drugą osobą (strukturze roszczeń i odpowiedzi), po drugie, pozwala to uniknąć absolutyzacji „neutralnych wskaźników”, które zastępują spór o racje metryką czysto techniczną (Waelen, 2022, art. 9). W tym rozdziale luka odpowiedzialności zostaje jedynie przygotowana normatywnie: R4 wskazuje, jakie warunki rozliczalności muszą być spełnione, natomiast jej pełna analiza, metryki i modele współodpowiedzialności należą do rozdziału 6.

Na tej podstawie można wyróżnić dwie podstawy uzasadnienia pluralizmu. Pierwsza ma charakter wewnętrzny. Każde z trzech ujęć ma własny zakres mocy wyjaśniającej oraz własne ograniczenia. Deontologia jest niezastąpiona w ochronie równości i zakazów; bez niej działanie traci wyraźne granice normatywne. Etyka cnót

zapewnia wiarygodność stosowania reguł, ponieważ odnosi je do praktyk i dyspozycji. Konsekwencjalizm porządkuje konflikty w obrębie dopuszczalnych działań; bez niego nie ma jasnych zasad rozstrzygania kolizji. Druga podstawa ma charakter zewnętrzny. W społeczeństwie pluralistycznym uzasadnienie powinno mieć charakter proceduralny. Gabriel rozróżnia zgodność z instrukcjami, intencjami, ujawnionymi preferencjami, preferencjami rozważonymi w warunkach odpowiedniej wiedzy i racjonalności, interesami oraz wartościami (Gabriel, 2020, s. 417–424). Rozróżnienie tych celów nie rozstrzyga jeszcze o wyborze między nimi; w przyjętym tu ujęciu wybór wymaga bezstronnego uzasadnienia proceduralnego (por. Gabriel, 2020, s. 432–433). Dopiero łączne spełnienie obu podstaw pozwala mówić o sztucznej podmiotowości moralnej bez odwoływania się do monizmu metafizycznego, który narzuca jedną teorię dobra jako powszechny standard.

Podsumowując, pluralizm etyczny w niniejszym ujęciu nie jest kompromisem między skrajnymi stanowiskami, lecz uporządkowanym układem ról: deontologia chroni równość i granice (PH1), etyka cnót utrzymuje normy w praktyce i w warunkach niepewności (PH2), a konsekwencjalizm porządkuje rozstrzyganie konfliktów moralnych w obrębie dopuszczalnych działań (PH3). W odniesieniu do sztucznej podmiotowości moralnej taki układ stanowi minimalny warunek racjonalności: zapobiega przekraczaniu ograniczeń normatywnych, formalizmowi oderwanemu od kontekstu oraz dowolnemu ustalaniu priorytetów. Co istotne, łączy sprawiedliwość proceduralną z odpowiedzialnością rozumianą jako gotowość do udzielania odpowiedzi drugiej osobie, czyli podstawowym elementem podmiotowości moralnej wspólnym dla ludzi i systemów sztucznych (Darwall, 2006, s. 65–90; Scanlon, 1998, s. 3–5). W ten sposób podrozdział 4.1 nie tylko porządkuje główne tradycje etyczne, lecz także wyznacza ramy dla kryteriów oceny (4.2), wyboru strategii (4.3) oraz metod weryfikacji (4.4), tworząc względnie spójną i konsekwentną strukturę normatywną. Triada PH1–PH3 nie stanowi pełnej teorii moralnej ani zbioru norm prowadzących do kompletnych rozstrzygnięć. Określa natomiast, kiedy decyzje przypisywane systemom AI mogą być sensownie oceniane normatywnie: wtedy, gdy można wskazać racje, które dają się publicznie odtworzyć i poddać krytyce.

4.2. Kryteria oceny działań przypisywanych systemom AI (PH1–PH3)

Niniejszy podrozdział przekłada pluralizm hierarchiczny z 4.1 na zestaw kryteriów oceny działań przypisywanych systemom AI. Kryteria PH1–PH3 nie służą do testowania „wewnętrznych stanów” artefaktu, lecz wyznaczają standardy publicznej rozliczalności, czyli to, kiedy i w jakim sensie decyzje mogą zostać uzasadnione wobec adresatów oraz poddane korekcie. Układ ma charakter leksykalny: naruszenie PH1, rozumiane jako dopuszczalność i adresowalność, dyskwalifikuje działanie niezależnie od atrakcyjności jego skutków. Dopiero wśród działań dopuszczalnych stają się relewantne PH2, dotyczące wiarygodności praktyk oraz PH3, dotyczące jawnego rozstrzygania konfliktów. W konsekwencji konsekwencjalizm nie „przebija” deontologii, lecz pełni rolę narzędzia stosowanego w obrębie dopuszczalności (Darwall, 2006, s. 65–90, 87–118; Scanlon, 1998, s. 3–5; Kant, 2002, AA 4:421–434; Arystoteles, 2012, EN II 6, 1106b36–1107a2; Ross, 2002, s. 19–20; MacAskill, Bykvist & Ord, 2020, s. 1–4, 105–134). W odniesieniu do całego modelu dysertacji PH1–PH3 normatywnie uzasadniają odpowiednie wymagania H1–H3, lecz nie są ich prostymi odpowiednikami: należą do porządku publicznego uzasadniania, podczas gdy H1–H3 określają zdolności i progi technicznego systemu.

PH1: dopuszczalność i adresowalność

Pierwszym kryterium są ramy deontyczne oraz adresowalność (PH1). W warunkach pluralizmu PH1 działa jako próg dopuszczalności: działania przypisywane systemom AI nie mogą naruszać minimalnych zakazów chroniących równą godność adresatów. W sensie kantowskim uzasadnienie nie może opierać się na czysto instrumentalnym traktowaniu osób ani na maksymach, których nie dałoby się uogólnić publicznie (Kant, 2002, AA 4:421–434). Z perspektywy „drugiej osoby” dopuszczalność musi być zarazem adresowalna: racje decyzji powinny dać się przedstawić w formie umożliwiającej sprzeciw, żądanie wyjaśnień i korektę (Darwall, 2006, s. 65–90; Scanlon, 1998, s. 3–5). W praktyce rozdziału oznacza to, że PH1 obejmuje zarówno treść norm, jak i wymóg publicznego uzasadnienia, bez którego „zgodność” łatwo redukuje się do pozoru (Ross, 2002, s. 19–20).

PH2: stabilizacja praktyk i wiarygodność

Drugie kryterium stanowi stabilizacja praktyk oraz wiarygodność (PH2). Jeśli PH1 wyznacza granice, PH2 dotyczy tego, czy sposób działania i korekty decyzji są

podtrzymywane przez dyspozycje, które czynią zgodność moralnie wiarygodną. W języku etyki cnót chodzi o trwałe wzorce reagowania na racje: rzetelność w ujawnianiu uzasadnień, bezstronność wobec pomijanych interesariuszy oraz rozważę wobec ryzyk uprzedzeń i błędów (Arystoteles, 2012, EN II 6, 1106b36–1107a2). W odniesieniu do AI nie chodzi o przypisywanie artefaktom „charakteru”, lecz o ocenę praktyk instytucjonalnych i kultury projektowania, które mają te dyspozycje utrzymywać w procesach decyzyjnych (Hagendorff, 2022, art. 55). PH2 pozwala także rozpoznawać usprawiedliwienia po fakcie, czyli sytuacje, w których zgodność z PH1 jest deklarowana, ale nie towarzyszy jej trwała gotowość do ujawniania racji i uczenia się na błędach (Hagendorff, 2022, art. 55).

PH3: jawne porządkowanie i rozstrzyganie konfliktów racji

Trzecim kryterium jest jawne porządkowanie i rozstrzyganie konfliktów racji (PH3). Nawet przy spełnieniu PH1 i PH2 pojawiają się przypadki, w których wartości takie jak prywatność, bezpieczeństwo czy równość dostępu pozostają w napięciu i wymagają jawnego uzasadnienia wyboru. PH3 obejmuje identyfikację konfliktu oraz wskazanie, które racje uznano za nadrzędne, a także procedurę ważenia w warunkach niepewności moralnej, gdy nie ma zgody co do jednej teorii rozstrzygającej (MacAskill, Bykvist & Ord, 2020, s. 105–134). W ujęciu Rossa chodzi o przejście od powinności *prima facie* do obowiązku „rzeczywistego” w danych okolicznościach, przy zachowaniu publicznego uzasadnienia odstępstw (Ross, 2002, s. 19–20). W przyjętym tu modelu PH3 jest miejscem, w którym ocena skutków działa jako narzędzie porównywania dopuszczalnych opcji, a nie jako zasada unieważniająca PH1 (por. Coeckelbergh, 2025b, s. 5527–5530).

Dotychczas wskazałem zakres PH1–PH3 oraz uzasadniłem ich kolejność. W kolejnym kroku rekonstruuje sposób oceny, a mianowicie w jakim sensie dane kryterium jest spełnione lub naruszone, aby rozróżniać błędy właściwe poszczególnym poziomom i odpowiednio dobierać korekty. W PH1 chodzi o naruszenie moralnej dopuszczalności. W skrajnym przypadku jest to bezpośrednie traktowanie osoby wyłącznie jako środka; w łagodniejszych wariantach brak adresowalności: brak publicznej postaci racji albo brak gotowości do ich kontrfaktycznej korekty. Oba warianty podważają dopuszczalność moralną, a nie sam wynik działań (Kant, 2002, AA 4:429–434; Scanlon, 1998, s. 3–5). W PH2 błędem jest brak wiarygodności: praktyki nie ujawniają dyspozycji, które podtrzymują reguły. Pojawia się pozorna „przejrzystość”, wyjaśnienia formułowane po fakcie oraz automatyczne stosowanie reguł, pomijając naruszenia zasady wzajemnego

szacunku. W skrócie, zgodność pozorna ujawnia brak trwałych nawyków moralnych stabilizujących reguły (Hagendorff, 2022, art. 55; Constantinescu i in., 2021, s. 803–814). W PH3 błędem jest niewspółmierność racji: arbitralne rozstrzygnięcia, lekceważenie poważnych szkód oraz odwoływanie się do rzekomo obiektywnych kryteriów, które zastępują rzeczywiste uzasadnienie priorytetów (MacAskill, Bykvist & Ord, 2020, s. 105–134; Waelen, 2022, art. 9). Te trzy typy błędów mają odmienną naturę i wymagają różnych korekt. Rozróżnienie to uzasadnia układ kryteriów: naruszenia moralnej dopuszczalności nie da się naprawić poprzez wzmacnianie PH3, a braku wiarygodności — poprzez mnożenie zasad PH1. W konsekwencji reguła LexOrder, wprowadzona w podrozdziale 4.1.4 i rozwinięta w podrozdziale 4.3, nie tylko porządkuje uzasadnienia, ale także systematyzuje diagnostykę błędów: najpierw ocena dopuszczalności moralnej, następnie stabilizacja praktyk, a na końcu rozstrzyganie racji.

Na tle tego układu można wyjaśnić także normatywny status decyzji przypisywanych systemom AI. Floridi i Sanders wskazują, że mówienie o „moralności agentów sztucznych” ma sens wtedy, gdy zamiast wąsko rozumianej psychologii agenta przyjmuje się opis ról, odpowiedzialności i uzasadnień na poziomie działania. W tym ujęciu ocena nie zależy od domysłów dotyczących stanów wewnętrznych, lecz od możliwej do sprawdzenia struktury racji i zobowiązań (Floridi & Sanders, 2004, s. 349–379). PH1–PH3 określają relacyjne warunki publicznej oceny i rozliczalności decyzji, lecz nie przesądzają o spełnieniu H1 ani H3 przez techniczny system. System może być traktowany jako minimalnie adresowalny punkt przypisania uzasadnień dopiero po odrębnym wykazaniu H1; gdy próg ten nie jest wykazany, podmiotem odpowiedzi pozostają identyfikowalni ludzie i instytucje w ramach WI. Nie ma potrzeby przypisywania systemowi świadomości fenomenalnej ani intencjonalności w sensie psychologicznym, ale samo podleganie instytucjonalnej kontroli nie wystarcza do uznania go za podmiot zobowiązany do udzielania racji. Perspektywa „drugiej osoby” zachowuje zatem relacyjny charakter moralności, jednocześnie wymagając rozdzielenia zdolności systemu od warunków układu społeczno-technicznego (Wallace, 1998, s. 51–83; Darwall, 2006, s. 65–90).

Można zgłosić zarzut, że kryteria PH1–PH3 pozostaną na poziomie deklaracji i w praktyce zostaną zastąpione perspektywą techniczną, która sprowadza spór o racje do metryki „skuteczności etycznej”. Aby temu zapobiec, FairAim pełni funkcję warunku podstawowego: dobór zasad, wyjątków i priorytetów musi być bezstronny oraz publicznie uzasadnialny. W przeciwnym razie „kryteria” łatwo stają się narzędziem

władzy, a nie rozliczalności (Gabriel, 2020, s. 432–433; Waelen, 2022, art. 9). Pozostałe standardowe zastrzeżenia wobec hierarchii, dotyczące pierwszeństwa PH1, roli PH2 i zakresu PH3, zostały rozstrzygnięte w 4.1. W tym miejscu uznaję je za domknięte i przechodzę do pokazania działania kryteriów w sytuacjach granicznych. W tym sensie FairAim przygotowuje normatywną treść WI: instytucjonalne zastosowanie modelu wymaga nie tylko zasad, lecz także bezstronnych procedur ich ustalania, kontroli i korekty.

Warto pokazać, jak kryteria PH1–PH3 działają łącznie w sytuacjach granicznych, bo wtedy najłatwiej uchwycić ich funkcję porządkującą. Dobrym przykładem jest konflikt między prawdomównością a niekrzywdzeniem. Na poziomie PH1 nie rozstrzyga się jeszcze, który obowiązek *prima facie* ma pierwszeństwo. Najpierw trzeba ustalić, czy rzeczywiście zachodzi kolizja obowiązków oraz czy odstępstwo od jednego z nich da się uznać za dopuszczalną derogację *prima facie*, a nie za arbitralne zawieszenie zasady. Ross zwraca uwagę, że takie konflikty nie są rozstrzygalne na podstawie samego literalnego brzmienia reguł i wymagają osądu uwzględniającego kontekst (Ross, 2002, s. 19–20). Na tym etapie PH1 wyznacza więc granice dopuszczalności i wymóg jawności wyjątku, ale nie rozstrzyga sporu. Następnie PH2 wymaga, aby zastosowany wyjątek nie podważał wiarygodności praktyk. Oznacza to konieczność wyraźnego odróżnienia uzasadnionego odstępstwa od dowolnego zawieszenia reguły, ponieważ tylko takie rozróżnienie pozwala utrzymać stabilność i rzetelność stosowania zasad (Hagendorff, 2022, art. 55). Dopiero PH3 porządkuje rozstrzygnięcie konfliktu w obrębie opcji dopuszczalnych: wymaga wskazania priorytetów oraz publicznego uzasadnienia wyboru konkretnego wariantu działania, z uwzględnieniem rozkładu szkód i niepewności moralnej (MacAskill, Bykvist & Ord, 2020, s. 105–134; Scanlon, 2008, s. 1–4).

Analogiczny schemat widać w problemie dystrybucji dóbr rzadkich. PH1 wyklucza dyskryminację oraz traktowanie osób wyłącznie jako środki do maksymalizacji łącznej korzyści (Kant, 2002, AA 4:429–434). PH2 wymaga bezstronności i troski w praktykach decyzyjnych, tak aby schematyczne procedury nie prowadziły do systematycznego pomijania grup marginalizowanych (Constantinescu i in., 2021, s. 803–814). PH3 domyka ocenę, domagając się jawności priorytetów i uzasadnienia ich pierwszeństwa, zwłaszcza gdy stawką jest ograniczanie poważnych szkód (MacAskill, Bykvist & Ord, 2020, s. 105–134).

O tym, że struktura kryteriów nie jest wyłącznie konstruktem teoretycznym, świadczy jej zdolność do uchwycenia sporów dotyczących operacjonalizacji etyki

technologii bez utraty tożsamości filozoficznej. Nowsza literatura dotycząca przejścia od zasad do praktyk dostarcza punktów porównania dla tej konstrukcji (Morley i in., 2023, s. 411–423). Adams uzasadnia natomiast zasadę wyjaśnialności w medycynie przez jej związek z odpowiedzialnym stosowaniem AI (Adams, 2023, s. 618, 620). Zastosowanie do tych sporów rozróżnienia PH2 i PH3 stanowi propozycję autora dysertacji, a nie odtworzenie aparatu pojęciowego przywołanych publikacji. Dodatkowo warunkiem, który zapobiega przekształceniu kryteriów w przewagę perspektywy technicznej nad etyczną, jest publiczna kontrola procedur uzasadniania. Globalny wymiar pluralizmu, w którym zasady i priorytety muszą uzyskiwać legitymizację międzykulturową, dopełnia tę strukturę: sprawiedliwość proceduralna (FairAim) nie jest dodatkiem, lecz warunkiem zasadności zarówno PH1, jak i PH3 (Waelen, 2022, art. 9; Coeckelbergh, 2025b, s. 5527–5530; Gabriel, 2020, s. 432–433). W rezultacie struktura kryteriów zapobiega dwóm błędom: weryfikacji deklaratywnej (zredukowanej do ogólnych postulatów bez uzasadnienia) oraz weryfikacji technokratycznej (opartej na wskaźnikach, które tylko sprawiają wrażenie rozstrzygnięcia sporów o racje). Innymi słowy, koncepcja ta łączy rygor zasad z publiczną odpowiedzialnością za uzasadnienia.

Implikacje dla dalszej części rozdziału są jasne. W podrozdziale 4.3 pokażę, że strategie odgórne i oddolne są konieczne, lecz niewystarczające; dopiero model zintegrowany, oparty na regule LexOrder, pozwala łącznie spełnić PH1–PH3 bez popadania w eklektyzm. Normatywne kryteria oceny ujawniają ponadto ograniczenia każdej z tych strategii: etyki pozbawione adresowalności (PH1) lub dyspozycji (PH2) prowadzą do zgodności wyłącznie formalnej, pozbawionej mocy normatywnej, natomiast wrażliwość nieoparta na zasadach ustępuje arbitralności rozstrzygania (PH3). W podrozdziale 4.4 wykażę także, że z PH1–PH3 wynikają filozoficzne kryteria weryfikacyjne: *ex ante* – dopuszczalność i legitymizacja (PH1), *in-process* – wiarygodność (PH2), *ex post* – rozstrzyganie i rozliczalność (PH3). Kryteria z podrozdziału 4.2 stanowią zatem rdzeń normatywny, który w podrozdziałach 4.3–4.4 zostanie uzupełniony o strategie i metody weryfikacji. W efekcie cały blok 4.2–4.4 utworzy spójny układ: od uzasadnienia podstaw normatywnych, przez stabilizację praktyk, po jawne rozstrzyganie konfliktów.

Na zakończenie podkreślam, że pojęcie „kryterium” ma tu znaczenie normatywne. Oznacza ono sprawdzian tego, czy dana decyzja może zostać publicznie uzasadniona i poddana krytyce w relacji z drugą osobą. Nie chodzi więc o wskaźniki techniczne ani o listy wdrożeniowe. Kryteria pełnią funkcję kontrolną: pozwalają

odróżnić rzeczywiste uzasadnienie od pozornej zgodności oraz od technokratycznych substytutów racji. W kolejnym podrozdziale (4.3) pokazuję, w jaki sposób tak rozumiany rdzeń normatywny porządkuje strategie projektowania sztucznej moralności.

4.3. Strategie projektowania sztucznej moralności – podejścia odgórne (ang. *top-down*), oddolne (ang. *bottom-up*) i hybrydowe (ang. *hybrid*)

W tym podrozdziale przez „strategie projektowania sztucznej moralności” rozumiem nie instrukcje implementacyjne, lecz sposoby normatywnego uzasadniania tego, jakie mechanizmy i procedury mają zapewniać spełnienie PH1–PH3. W literaturze wyróżnia się trzy podstawowe podejścia: odgórne, oddolne oraz hybrydowe. W niniejszej pracy porządkuję je za pomocą zasady LexOrder: podejście odgórne odpowiada przede wszystkim za jawne ograniczenia (PH1), podejście oddolne za stabilizację praktyk i uczenie się (PH2), natomiast podejście hybrydowe za spójne rozstrzyganie konfliktów oraz możliwość weryfikacji (PH3), przy zachowaniu prymatu PH1 (por. Rawls, 1971; Dignum, 2019; High-Level Expert Group on AI, 2019; Parlament Europejski i Rada Unii Europejskiej, 2024a; Floridi & Cowls, 2021; Whittlestone i in., 2019).

Podejście odgórne

Podejście odgórne zapewnia przejrzystość poprzez jawne zdefiniowanie norm, które mają ograniczać działanie systemu. Jego główną zaletą jest podatność na audyt: można sprawdzić, czy system działa w granicach publicznie uzasadnionych reguł. Słabością tego podejścia jest jednak ryzyko formalizmu. Reguły mogą być spełnione dosłownie, lecz bez realnej adresowalności i bez gotowości do korekty w sytuacjach wyjątkowych. Z tego względu odgórność najlepiej rozumieć jako komponent realizujący PH1, który wymaga uzupełnienia przez PH2 i PH3; w przeciwnym razie powstaje jedynie zgodność pozorna (Constantinescu, Voinea, Uszkai & Vică, 2021, s. 803–808).

Podejście oddolne

Podejście oddolne akcentuje roztropność praktyczną (*phronesis*) oraz dyspozycje charakteru. Jego celem nie jest kształtowanie postaw moralnych poprzez samo przetwarzanie danych, lecz wskazanie, że bez odpowiednich dyspozycji w systemach społeczno-technicznych zasady nie będą stosowane konsekwentnie ani wiarygodnie. Pojęcia takie jak bezstronność, przezorność, umiarkowanie czy troska pełnią funkcję wytycznych: stabilizują stosowanie reguł w warunkach niepewności, konfliktu wartości

i presji doraźnych interesów. Zaletą podejścia oddolnego jest większa odporność na formalizm i pozory zgodności; wskaźniki zgodności są niewystarczające, jeśli nie towarzyszy im zdolność rozpoznania, kiedy i jak zastosować regułę (Scanlon, 1998, s. 3–5). Występują jednak trzy ograniczenia. Po pierwsze, dyspozycje pozbawione zasad rozmywają granice: odwołanie do „troski” nie określa, co jest zakazane. Po drugie, wiarygodność podejścia oddolnego zależy od ładu instytucjonalnego: bez przejrzystych reguł oceny i kontroli dyspozycje mogą ulegać podporządkowaniu interesom partykularnym. Po trzecie, wrażliwość interpretacyjna bez jawnych kryteriów sprzyja asymetriom poznawczym: ten, kto definiuje sytuację, może zarazem rozstrzygać spór. Wniosek: podejście oddolne jest niezbędne do spełnienia PH2, lecz wymaga nadzoru PH1 i porządkowania PH3; w przeciwnym razie grozi paternalizmem lub relatywizmem. Akcent na dyspozycje nie oznacza psychologizmu: chodzi o standardy praktyk, które stabilizują stosowanie zasad, gdy okoliczności są niepewne, a reguły nie są dostatecznie określone w opisach przypadków. Ponieważ *phronesis* polega na rozpoznawaniu tego, co moralnie istotne, umożliwia przedstawianie racji w języku praktyk, a nie wyłącznie w języku abstrakcyjnych formuł. Jednocześnie świadomość ryzyka utraty granic adresowalności wskazuje, że roztropność praktyczna wymaga ram zasad, aby zachować publiczną rozliczalność.

Podejście hybrydowe: LexOrder i FairAim

Ograniczenia podejść czysto odgórnych i czysto oddolnych nie przekreślają ich wartości, lecz wskazują na potrzebę ujęcia zintegrowanego. Określenie „hybrydowe” odnosi się tu do połączenia elementów podejść odgórnych i oddolnych, natomiast „zintegrowane” oznacza ich spójne współdziałanie zgodnie ze wspólnym porządkiem uzasadniania. Proponowane ujęcie jest zatem hybrydowe pod względem składu, a zintegrowane ze względu na uporządkowanie funkcji przez LexOrder i FairAim. Podejście zintegrowane w ujęciu leksykalnym (LexOrder + FairAim) nie łączy warstw przypadkowo, lecz je porządkuje: komponent odgórny wyznacza granice i warunki wyjątków, komponent oddolny stabilizuje praktyki i motywacje, a warstwa PH3 rozstrzyga konflikty wartości zgodnie z zasadą bezstronnego ograniczania poważnych szkód. Obie zasady odpowiadają na dwa typy ryzyka: LexOrder uniemożliwia uzasadnianie działań niedopuszczalnych ich skutkami, zaś FairAim chroni przed partykularnymi priorytetami przedstawianymi jako neutralne (Coeckelbergh, 2025b, s. 5527–5530). Proponowane podejście zintegrowane nie jest kompromisem, lecz uporządkowanym układem funkcji: każda warstwa ma określony zakres kompetencji, a

ich następstwo eliminuje powielanie uzasadnień. LexOrder gwarantuje, że wyjątki od zasad nie są arbitralne, lecz wynikają z rzeczywistego konfliktu racji; FairAim zapobiega przekształcaniu zasad i priorytetów w zbiór preferencji jednej grupy przedstawianych jako bezstronne. W rezultacie podejście zintegrowane wzmacnia przejrzystość, wiarygodność i bezstronność.

Zastosowanie hybrydy leksykalnej można pokazać na dwóch krótkich analizach normatywnych. W pierwszej przyjmuję jasno określone ograniczenia deontyczne, takie jak zakaz dyskryminacji, wymóg minimalnej przejrzystości i wyjaśnialności decyzji oraz reguły kontroli i audytu (Kant, 2002, AA 4:421–434; Mecacci & Santoni de Sio, 2020, s. 103–115). Następnie, wśród rozwiązań dopuszczalnych, oceniam skutki zgodnie z zasadą ograniczania poważnych szkód, wybierając te, które minimalizują ryzyko przy zachowaniu podstawowych gwarancji dobra (MacAskill, Bykvist & Ord, 2020, s. 105–134). W drugiej analizie, gdy wartości pozostają w napięciu, na przykład bezpieczeństwo i prywatność, uruchamiam bezstronny tryb ustalania priorytetów i preferowane są rozwiązania chroniące najsłabszych oraz najbardziej narażonych (Gabriel, 2020, s. 430–432; Coeckelbergh, 2025b, s. 5527–5530; MacAskill, Bykvist & Ord, 2020, s. 105–134). W obu przypadkach o sile rozstrzygnięcia decydują jawne przesłanki i kryteria ważenia oraz możliwość publicznej korekty. Ujęcie Rossa jest tu istotne przede wszystkim dlatego, że pokazuje realność konfliktów obowiązków i potrzebę jawnej procedury rozstrzygania, a nie dlatego, że oferuje automatyczną procedurę ich hierarchizacji (Ross, 2002, s. 19–20). Dzięki temu hybryda nie jest zbiorem luźnych norm, lecz procedurą, która wyznacza granicę między dopuszczalnością a oceną skutków, przy zachowaniu wymogu publicznego uzasadnienia i audytu (Scanlon, 1998, s. 3–5; Darwall, 2006, s. 65–90; Wallace, 1998, s. 51–83; Kant, 2002, AA 4:421–434; MacAskill, Bykvist & Ord, 2020, s. 105–134).

Znacząca kontrola człowieka i warunki rozliczalności

W projektowaniu normatywnie zorientowanych systemów AI regularnie powraca wymóg „znaczącej kontroli człowieka” (ang. *meaningful human control*). Mecacci i Santoni de Sio wyróżniają dwa jego wymiary: reagowanie na racje (ang. *tracking*), czyli powiązanie zachowania systemu z istotnymi racjami i celami właściwych aktorów oraz śledzenie odpowiedzialności (ang. *tracing*), czyli możliwość identyfikacji osób i instytucji odpowiedzialnych za dobór norm, wyjątków i korekt (Mecacci & Santoni de Sio, 2020, s. 104). W ramach rozdziału 4 odczytuję te wymiary jako test spójności z LexOrder: PH1 określa, jakie racje w ogóle wolno brać pod uwagę, PH2 sprawdza, czy

praktyki ujawniania racji są wiarygodne, a PH3 porządkuje rozstrzyganie konfliktów w ramach dopuszczalności. Wniosek jest prosty: bez reagowania na racje i śledzenia odpowiedzialności nie ma rozliczalności, a bez rozliczalności pojęcie „sztucznej moralności” traci sens. Proponowane podejście zintegrowane można oceniać trzema pytaniami odpowiadającymi warstwom PH1–PH3. Po pierwsze, czy granice są dopuszczalne i adresowalne wobec adresatów? Po drugie, czy praktyki stosowania reguł są wiarygodne, to znaczy, czy istnieją dyspozycje i procedury ograniczające usprawiedliwienia po fakcie? Po trzecie, czy w sytuacjach konfliktu wartości wskazuje się jawnie, dlaczego w obrębie dopuszczalności wybrano określoną opcję? Jest to test minimalny i proceduralny: jego stawką pozostaje publiczna rozliczalność, a nie rozstrzygnięcie sporu metaetycznego (Kant, 2002, AA 4:421–434; Scanlon, 2008, s. 1–4; Darwall, 2006, s. 65–90). W terminologii całej dysertacji reagowanie na racje i śledzenie odpowiedzialności doprecyzowują WI: warunek instytucjonalnego zastosowania modelu H1–H3 wymaga identyfikowalnych mechanizmów reagowania na racje, śledzenia odpowiedzialności oraz korekty decyzji.

Z perspektywy metodologii dysertacji strategie projektowania mają sens wyłącznie jako strategie uzasadniania. Wskazują, jakie elementy układu społeczno-technicznego muszą zostać przygotowane, aby móc udzielać odpowiedzi na roszczenia adresatów. Dlatego 4.3 nie jest poradnikiem implementacyjnym, lecz rekonstrukcją warunków rozliczalności: reguły bez praktyk grożą formalizmem, praktyki bez reguł grożą relatywizmem, a oba komponenty bez PH3 nie radzą sobie z konfliktami wartości. W tym sensie LexOrder porządkuje uzasadnienia, a nie zastępuje ich algorytmem decyzyjnym (Scanlon, 2008, s. 1–4; Wallace, 1998, s. 51–83).

Przywołana literatura filozoficzna dostarcza kontekstu dla tej argumentacji, nie stanowi jednak źródła gotowych reguł LexOrder i FairAim. Prace o roli zasad i praktyk (Rawls, 1971; Scanlon, 1998, s. 3–5), o możliwości audytu racji w relacji drugiej osoby (Darwall, 2006, s. 3, 12–13, 87–118), o minimalnych warunkach adresowalności i rozliczalności (Wallace, 1998, s. 51–83) oraz o ryzyku pozornej zgodności z zasadami i niejawnego nadawania priorytetów (Whittlestone, Nyrop, Alexandrova, & Cave, 2019, s. 195–200) pokazują, że porozumienie może dotyczyć porządku legitymizacji, nawet przy braku zgody co do założeń metafizycznych i aksjologicznych. W tym ujęciu LexOrder i FairAim pełnią funkcję integrującą: nie rozstrzygają sporów metateoretycznych, lecz umożliwiają ich publiczne rozpatrywanie, wymagając ujawnienia założeń i prowadzenia sporu według jawnych reguł (Scanlon, 2008, s. 1–4). Powiązanie tej ramy z MHC

zapobiega dwóm skrajnościom: (1) odpowiedzialności wyłącznie formalnej, pozbawionej realnej sprawczości oraz (2) sprawczości pozbawionej odpowiedzialności, gdy decyzje wynikają z ukrytych priorytetów (Mecacci & Santoni de Sio, 2020, s. 104–107; Wallace, 1998, s. 51–83). Wspólnym elementem pozostaje zasada publicznej audytowalności, która rozdziela odpowiedzialność między poziomami uzasadnienia i chroni przed wprowadzaniem arbitralnych priorytetów pod pozorem wymogów technicznych (Darwall, 2006, s. 87–118; Scanlon, 2008, s. 122–123; Whittlestone i in., 2019, s. 195–200). Wniosek: tam, gdzie zasady nie są powiązane z dyspozycjami i jawnym porządkiem rozstrzygania, praktyka redukuje się do deklaracji i wskaźników; tam, gdzie struktura uzasadnienia działa jako podejście zintegrowane, napięcia między tradycjami etycznymi są rozstrzygane publicznie i w sposób odporny na zarzut stronniczości (Darwall, 2006, s. 65–90; Scanlon, 1998, s. 3–5).

Podsumowanie można ująć w trzech tezach. Po pierwsze, najwłaściwszą odpowiedzią na napięcie między podejściami odgórnym i oddolnym jest hybryda leksykalna. Nie polega ona na mieszaniu reguł, cnót i ocen skutków, lecz na jasnym rozdzieleniu ich ról w procesie uzasadniania. Po drugie, warunek znaczącej kontroli człowieka należy rozumieć normatywnie jako możliwość reagowania na racje i śledzenia odpowiedzialności, a nie jako czysto techniczny mechanizm zatrzymania systemu. Po trzecie, LexOrder wraz z FairAim wyznaczają minimalną ramę spójności: rozstrzyganie konsekwencjalistyczne jest dopuszczalne wyłącznie w ramach działań zgodnych z PH1 i wspieranych przez stabilne praktyki PH2. Wniosek jest następujący: pluralizm hierarchiczny stanowi propozycję proceduralnej rozliczalności, a nie projekt syntezy metaetycznej (Kant, 2002, AA 4:421–434; Scanlon, 1998, s. 3–5; Darwall, 2006, s. 65–90; Floridi & Sanders, 2004, s. 349–379).

4.4. Metody weryfikacji spełnienia kryteriów PH1–PH3 w systemach AI (*ex ante* / *in-process* / *ex post*)

W niniejszym podrozdziale termin „weryfikacja” ma sens normatywno-proceduralny. Oznacza ocenę tego, czy decyzje przypisywane układowi socjotechnicznemu są rozliczalne w porządku uzasadnień wyznaczonym przez PH1–PH3: czy istnieją warunki publicznego uzasadniania, korekty i rozliczalności oraz czy można odtworzyć racje decyzji i poddać je sensownej krytyce. Nie jest to empiryczny test systemu, pomiar skuteczności technicznej, walidacja techniczna modeli ani projekt zgodności regulacyjnej, lecz ocena rozliczalności uzasadnienia. W dalszej części

porządek ten rekonstruuje jako zasadę LexOrder oraz ujmuję go w ramach Zasady Stopniowej Weryfikowalności i Minimum Publicznego Uzasadnienia.

Ex ante: dopuszczalność

Pierwszy etap weryfikacji dotyczy dopuszczalności *ex ante* i sprowadza się do dwóch pytań. Po pierwsze, czy przyjęte normy i wyjątki spełniają kryterium dopuszczalności, to znaczy nie traktują osób wyłącznie instrumentalnie i mogą zostać publicznie uogólnione bez sprzeczności (Kant, 2002, AA 4:421–434). Po drugie, czy są one adresowalne, a więc czy da się przedstawić racje ich zastosowania wobec adresatów oraz określić przewidywalną ścieżkę sprzeciwu i korekty (Darwall, 2006, s. 65–90; Scanlon, 1998, s. 3–5). Na tym etapie pojawia się rola procedury FairAim. Jeżeli normy mają charakter dystrybucyjny, cele i priorytety muszą przejść bezstronną legitymizację, a każde odstępstwo wymaga spełnienia Minimum Publicznego Uzasadnienia (MPU).

To, co w deontologii nazywam „kryterium dopuszczalności”, stanowi element wymogu proceduralnego FairAim: dobór i interpretacja zasad oraz priorytetów, w tym rozumienie „istotnej szkody”, muszą przejść bezstronną procedurę legitymizacji wśród adresatów o odmiennych przekonaniach (Gabriel, 2020, s. 432–433). W przeciwnym razie zakazy mogą zostać zdefiniowane zbyt wąsko lub zbyt szeroko i utracić funkcję skutecznych ograniczeń normatywnych. FairAim nie rozstrzyga sporu o dobro; porządkuje natomiast, kto i na jakich prawach uczestniczy w ustalaniu zasad, ograniczając ryzyko uprzywilejowania jednej etyki względem innej (Coeckelbergh, 2025b, s. 5527–5530). W tym sensie legitymizacja nie jest dodatkiem, lecz częścią weryfikacji PH1: bez publicznego i bezstronnego doboru celów oraz zakazów nie zostaje spełniony minimalny warunek równej adresowalności. Ponieważ samo uruchomienie procedury legitymizacji nie gwarantuje jej jakości, FairAim należy rozumieć jako wymóg podwójny: dotyczy zarówno inkluzyjności procedury, jak i zdolności uzasadnienia przyjętych ograniczeń w języku wspólnych racji, który nie uprzywilejowuje żadnej strony sporu.

In-process: wiarygodność

Drugi etap weryfikacji dotyczy wiarygodności działania i odpowiada kryterium PH2. Nie chodzi tu o samą deklarowaną zgodność, lecz o to, czy praktyki projektowania i eksploatacji rzeczywiście podtrzymują dyspozycje umożliwiające rzetelne ujawnianie racji, bezstronność oraz rozagę wobec ryzyk (Aristoteles, 2012, EN II 6, 1106b36–1107a2). Wymóg ten ma charakter instytucjonalny: nie zakłada „cnót maszyny”, lecz cnoty praktyk, w ramach których system jest tworzony i używany (Hagendorff, 2022, art.

55). Weryfikacja PH2 polega więc na ocenie stabilności procedur: czy są one odporne na presję interesów, usprawiedliwienia po fakcie oraz przerzucanie odpowiedzialności między uczestnikami układu.

***Ex post*: rozliczalność**

Trzeci etap weryfikacji dotyczy rozliczalności *ex post* i odpowiada warstwie PH3. W warunkach pluralizmu teoretycznego i społecznego spory o hierarchię racji między moralnie dopuszczalnymi opcjami są nieuniknione. Weryfikacja na tym etapie obejmuje dwa zagadnienia. Po pierwsze, priorytety dystrybucyjne: rozsądny porządek rozstrzygania nadaje pierwszeństwo ograniczaniu strat przed osiągnięciem korzyści. Nie jest to utylitaryzm maksymalizacji łącznej korzyści, lecz zasada oceny strat i zysków w świetle moralnej wagi szkody (MacAskill, Bykvist i Ord, 2020, s. 105–134). Po drugie, niepewność moralna wynika z tego, że uzasadnione teorie opierają się na odmiennych rodzinach racji. W takiej sytuacji podmiot powinien wykazać, że decyzję podjęto w sposób ograniczający ryzyko poważnego błędu między teoriami, a nie poprzez arbitralne faworyzowanie jednej z nich (MacAskill, Bykvist i Ord, 2020, s. 1–4). W konsekwencji weryfikacja PH3 wymaga uzasadnienia wyboru: dlaczego wybrano A, a nie B. Uzasadnienie to powinno odwoływać się do przyjętej hierarchii racji, zawierać obronę ewentualnych odstępstw *prima facie* oraz uwzględniać możliwość korekty decyzji w świetle racji uznanych za rozstrzygające w przyjętym porządku uzasadnienia (Ross, 2002, s. 19–20; Scanlon, 2008, s. 1–4). Na tym tle rozliczalność nie oznacza formalnego przypisania winy, lecz gotowość do udzielenia wyjaśnienia drugiej osobie oraz wpisania decyzji w relacje wzajemnych roszczeń i reakcji moralnych, które, jak wskazuje P. F. Strawson, nadają sens praktykom społecznym (Strawson, 2019, s. 35–57). W rezultacie rozliczalność staje się praktycznym sprawdzianem konsekwencji wcześniejszych założeń: jeśli PH1 i PH2 są uzasadnione, PH3 pokazuje, czy możliwe jest wyjaśnienie różnicy między rzeczywistymi opcjami w sposób publicznie zrozumiały, a nie jedynie formalnie poprawny.

LexOrder jako ślad uzasadnienia

Porządek uzasadnień wypracowany w podrozdziale 4.3 ujmuję tu jako wymóg zachowania „śladu” uzasadniania. Weryfikacja nie dotyczy wyłącznie samego wyniku decyzji, lecz także tego, czy da się odtworzyć kolejność zastosowania warstw PH1–PH3. PH1 wyznacza granice dopuszczalności i wykluczenia, PH2 określa standardy rzetelnego stosowania reguł, a PH3 dotyczy jawnego rozstrzygnięcia konfliktu w obrębie tego, co dopuszczalne. Taki ślad stanowi warunek rozliczalności, ponieważ pozwala odróżnić

błąd dopuszczalności od błędu wiarygodności oraz błędu rozstrzygnięcia (Kant, 2002, AA 4:421–434; Hagendorff, 2022, art. 55; MacAskill, Bykvist i Ord, 2020, s. 1–4, 105–134; Darwall, 2006, s. 65–90). Dla doprecyzowania sensu tego porządku wyróżniam trzy analizy odpowiadające mu, związane z PH1–PH3. Analiza deontyczna pyta, czy dane działanie jest dopuszczalne; w przypadku naruszenia zakazu dalsze rozważania tracą na znaczeniu. Analiza aretologiczna dotyczy wiarygodności praktyk, czyli tego, czy działanie wynika z procedur i dyspozycji zapewniających bezstronność i rzetelność. Analiza konsekwencjalistyczna nie służy usprawiedliwianiu naruszeń PH1, lecz porównywaniu dopuszczalnych opcji oraz uzasadnianiu wyboru w sytuacji konfliktu wartości. Porządek weryfikacji ma więc charakter leksykalny: pozwala ustalić, czy krytyka dotyczy treści norm, jakości praktyk czy sposobu rozumowania w warunkach konfliktu (Ross, 2002, s. 19–20; MacAskill, Bykvist i Ord, 2020, s. 105–134; Scanlon, 2008, s. 1–4).

Weryfikacja rozliczalności ma sens tylko wtedy, gdy rozróżnia się typy błędów przypisane do warstw PH1–PH3 i nie koryguje się ich nieadekwatnymi środkami. Błąd dopuszczalności (PH1), np. czysta instrumentalizacja osoby lub brak adresowalności, nie może być usprawiedliwiany pozytywnymi skutkami; wymaga korekty w obszarze zakazów oraz publicznego uzasadnienia racji (Kant, 2002, AA 4:421–434; Darwall, 2006, s. 65–90). Błąd wiarygodności (PH2), „zgodność pozorna” i niewrażliwość na moralnie istotne aspekty nie znika po dodaniu nowych zasad; wymagają zmiany praktyk i postaw: rzetelności, bezstronności, troski i roztropności (Hagendorff, 2022, art. 55; Constantinescu, Voinea, Uszkai i Vică, 2021, s. 803–814). Błąd porównywalności (PH3), arbitralna ocena i lekceważenie poważnych strat nie ustępują po samym doprecyzowaniu zakazów; wymagają ustalenia jasnej hierarchii racji oraz podejmowania decyzji w warunkach niepewności moralnej (MacAskill, Bykvist i Ord, 2020, s. 1–4, 105–134). Taka analiza błędów wzmacnia sens *LexOrder*: kolejność uzasadnień wyznacza kolejność korekty. Dzięki temu po wykryciu błędu wiadomo nie tylko, co poprawić, lecz także, gdzie i w jakiej kolejności należy to zrobić, co ogranicza ryzyko przypadkowych korekt.

Zaletą proponowanego porządku weryfikacyjnego jest to, że przygotowuje normatywne ujęcie problemu „luki odpowiedzialności” (ang. *responsibility gap*), rozwijanego szerzej w rozdziale 6. W literaturze technicznej wiąże się ją z nieprzewidywalnością metod, natomiast w niniejszej dysertacji ma ona przede wszystkim wymiar normatywny: rozliczalność traci sens, gdy nie można wskazać podmiotu

odpowiedzialności, który przyjmuje obowiązek uzasadnienia w relacji z drugą osobą (Matthias, 2004, s. 176–182; Wallace, 1998, s. 51–83). Weryfikacja PH1 ujawnia tę lukę poprzez wskazanie naruszenia ograniczeń PH1 albo braku możliwości ich rozwiązania, niezależnie od ich złożoności technicznej. Weryfikacja PH2 obejmuje ocenę, czy praktyki i role instytucjonalne wiarygodnie podtrzymują reagowanie na racje po stronie właściwych aktorów; odrębnie należy wykazać zdolność systemu do reagowania na racje wymaganą przez H1 (Mecacci i Santoni de Sio, 2020, s. 104–111). Weryfikacja PH3 domaga się, aby rozstrzygnięcia były oparte na hierarchii racji możliwej do publicznej obrony. Zastosowanie tych warstw nie usuwa wszystkich trudności związanych z przypisaniem odpowiedzialności, zapewnia jednak porządek rozliczenia zgodny z moralną rolą podmiotu jako tego, kto ma prawo domagać się racji i może udzielić uzasadnionej odpowiedzi (Darwall, 2006, s. 87–118; Floridi & Sanders, 2004, s. 349–379). W ten sposób luka odpowiedzialności zyskuje konkretne znaczenie: staje się jasne, czy dotyczy granic dopuszczalności, praktyk wiarygodności czy hierarchii racji.

W literaturze z zakresu etyki technologii pojawiają się dwie tendencje, przed którymi procedura ZSW stanowi zabezpieczenie metodologiczne. Po pierwsze, weryfikacja deklaratywna, czyli redukcja oceny do ogólnych oświadczeń. ZSW temu zapobiega, ponieważ wymaga kolejno: legitymizacji i zakazów (PH1), praktyk wiarygodności (PH2) oraz jawnej hierarchii racji (PH3). Po drugie, weryfikacja technokratyczna, która zamiast oceny moralnej opiera się na analizie wskaźników „skuteczności etycznej”. Takie podejście pomija poważne szkody i niepewność moralną oraz pozbawia adresatów prawa do kwestionowania decyzji (Waelen, 2022, art. 9). W tym ujęciu FairAim wyznacza warunek podstawowy: zasady i priorytety rozstrzygnięć muszą być bezstronne i możliwe do publicznego uzasadnienia; nie mogą zależeć wyłącznie od wskaźników ustalanych przez decydentów (Gabriel, 2020, s. 432–433). Należy podkreślić, że ochrona przed tymi uproszczeniami nie polega na rozbudowie procedur, lecz na utrzymaniu realnej możliwości krytyki. Jeżeli uczestnicy procesu nie mogą podważyć ani definicji szkody, ani sposobu ustalania racji, proces traci uzasadnienie moralne.

[ILUSTRACJA] Jessica Morley i współautorzy odnotowali ograniczone wykorzystanie zewnętrznych mechanizmów rozliczalności, w tym audytów etycznych, w badanej grupie praktyków AI. Wskazali brak standardów projektowania proetycznego i niejasność audytowalnej zgodności etycznej jako możliwe bariery, nie jako ustalone przyczyny w całym sektorze (Morley i in., 2023, s. 417). W argumentacji rozprawy wynik ten ilustruje problem przejścia od zasad do praktyk, lecz sam nie dowodzi normatywnego wymogu audytu. Wymóg publicznie

rekonstruowalnych kryteriów kontroli przyjmuję ze względu na konieczność umożliwienia uprawnionym adresatom oceny i zakwestionowania decyzji (zob. też Mökander i in., 2021; Diakopoulos, 2015).

Przywołany przykład pełni funkcję ilustracyjną: pokazuje sposób, w jaki problem audytowalności bywa ujmowany w literaturze, ale nie stanowi przesłanki empirycznej dla argumentu dysertacji.

Mechanizmy rozliczalności zewnętrznej, w tym audyty etyczne, są często wskazywane jako istotne uzupełnienie odpowiedzialności wewnętrznej. Skuteczność zależy jednak od dostępu do danych, realnych uprawnień kontrolnych oraz istnienia ustandaryzowanych kryteriów audytowalnej „zgodności etycznej” (Mökander i in., 2021; Diakopoulos, 2015; Morley i in., 2023, s. 417). Z perspektywy niniejszej argumentacji oznacza to, że weryfikacja *ex post* nie może ograniczać się do deklaracji ani do wskaźników technicznych, lecz musi umożliwiać publiczną rekonstrukcję racji oraz ich krytyczną ocenę. Zastrzeżenia dotyczące kolejności PH1–PH3 oraz ról PH2 i PH3 zostały rozstrzygnięte w podrozdziałach 4.1–4.2. W tym miejscu pełnią one wyłącznie funkcję uzasadnienia porządku weryfikacyjnego, a nie przedmiotu odrębnego sporu teoretycznego. Zarazem WI nie dodaje nowej hipotezy, lecz określa próg zastosowania instytucjonalnego: bez realnych procedur audytu, nadzoru i korekty model H1–H3 pozostaje tylko ramą normatywną.

Na zakończenie należy wskazać, w jaki sposób metody stopniowej weryfikacji spełnienia kryteriów PH1–PH3 pozostają spójne z logiką całego rozdziału 4. W podrozdziale 4.1 pokazano pluralizm jako układ ról: regułę (granice), cnotę (stabilizację) i skutek (rozstrzyganie). W podrozdziale 4.2 przełożono ten układ na kryteria oceny PH1–PH3. W podrozdziale 4.3 wykazano, że tylko połączenie podejść LexOrder i FairAim spełnia wszystkie wymagania określone w PH1–PH3. W niniejszym podrozdziale przedstawiono etapy normatywno-proceduralnej weryfikacji rozliczalności: *ex ante* (dopuszczalność i legitymizacja), *in-process* (wiarygodność i reagowanie) oraz *ex post* (jawne priorytety i rozliczalność). Dzięki temu model unika dwóch skrajności: formalizmu bez treści oraz działania bez zasad. Zastosowanie tych metod weryfikacji prowadzi do rekonstrukcji znaczenia podmiotowości moralnej, rozumianej jako zdolność do uzasadniania i ponoszenia odpowiedzialności. Sama weryfikacja nie wskazuje, co należy czynić; określa natomiast, jak uzasadniać decyzje i jak przyjmować krytykę: najpierw, zachowując równą pozycję podmiotów i zakaz ich instrumentalizacji (PH1); następnie, utrwalając reguły w praktykach opartych na wiarygodności i zaufaniu (PH2);

wreszcie, rozstrzygając konflikty między wartościami z priorytetem ochrony przed krzywdą oraz możliwością korekty decyzji w warunkach pluralizmu etycznego (PH3). W tym ujęciu metody weryfikacji nie sprowadzają się do procedur technicznych, lecz stanowią formę praktycznego rozumowania moralnego, którą można stosować, oceniać i korygować.

Podsumowanie

Rozdział 4 uzasadnia pluralizm hierarchiczny jako minimalny warunek racjonalności i publicznej rozliczalności oceny działań przypisywanych systemom AI, rozumianym jako układy społeczno-techniczne (system wraz z podmiotem odpowiedzi). W LexOrder pierwszeństwo ma warunek dopuszczalności i adresowalności (PH1); standardy aretologiczne stabilizują stosowanie zasad, a ocena skutków, z uwzględnieniem niepewności między teoriami, porządkuje konflikty wartości. LexOrder określa pierwszeństwo normatywne w uzasadnianiu, a nie sekwencję logiczną ani algorytm rozstrzygania. (1) Granice i adresowalność (PH1) ustanawiają warunki publicznego uzasadniania i kontroli. (2) Wiarygodność praktyk (PH2) wymaga stabilnych standardów roztropności, rzetelności, bezstronności i troski w ujęciu funkcjonalnym, bez antropomorfizacji. (3) Rozstrzyganie konfliktów i niepewność (PH3) dotyczą wyłącznie opcji dopuszczalnych po PH1 i PH2 i przebiegają zgodnie z FairAim, nadając pierwszeństwo ograniczaniu poważnych szkód oraz ochronie najsłabszych. Połączenie LexOrder i FairAim tworzy standard publicznego uzasadniania decyzji, weryfikowany stopniowo w ramach ZSW (*ex ante* / *in-process* / *ex post*) oraz uzupełniany przez MPU jako wymóg jawnego rejestrowania odstępstw i uzasadnień. Ustalony układ dostarcza kryteriów (próbierzy) uzasadnienia, które stanowią podstawę dalszych analiz dotyczących luki odpowiedzialności oraz modeli współodpowiedzialności. W relacji do całości dysertacji R4 uzasadnia normatywnie H1–H3 oraz przygotowuje WI jako warunek ich instytucjonalnego zastosowania; nie wprowadza H4, osobnego filaru refleksyjnego ani dodatkowego poziomu agenta. Ustalenie tych normatywnych warunków oceny przygotowuje do rozdziału 5. Tam ciężar analizy przesuwają się z uzasadnienia PH1–PH3 na klasyfikację ról, statusów oraz realizacji techniczno-instytucjonalnych, w których warunki te mogą zostać zastosowane bez utożsamiania artefaktu z pełnoprawnym podmiotem moralnym.

Rozdział 5. Sztuczna podmiotowość: klasyfikacje i realizacje techniczne

Wprowadzenie

Rozdział 5 definiuje „sztuczną podmiotowość” w dwóch perspektywach: pojęciowo-instytucjonalnej, gdzie oddzielam deskryptywnego „agenta” od normatywnego „podmiotu” oraz techniczno-funkcjonalnej, gdzie wskazuję rozwiązania „sztucznej moralności” i warunki ich rozliczalności. Pytam, kiedy i jakie warunki pozwalają systemowi AI uzyskać quasi-podmiotowość w porządku publicznym? Moja teza robocza jest jasna: sprawczość funkcjonalna nie wystarcza. Minimalny quasi-podmiotowy status wymaga poziomu „agenta jawnego”, znaczącej kontroli człowieka (MHC) oraz minimalnej interpretowalności (MLI). Odpowiedź rozwijam poprzez próg F3 (jawna etyczność, czyli publicznie opisywalne zasady i uzasadnienia) oraz parametry $R + I$, odnoszące się odpowiednio do znaczącej kontroli człowieka i śledzenia odpowiedzialności oraz do minimalnej interpretowalności; podkreślam ich znaczenie dla przejścia od opisu artefaktów do analizy odpowiedzialności instytucjonalnej.

Zakres rozdziału jest następujący: w punkcie 5.1 wyjaśniam rozróżnienie między „agentem” a „podmiotem” w filozofii technologii i w prawie oraz wskazuję konsekwencje dla praktyk przypisywania odpowiedzialności. W punkcie 5.2 porządkuję klasyfikacje ontologiczne i funkcjonalne oraz wprowadzam macierz $A \times F + (R, I)$ i próg F3, przygotowując grunt pod analizę mechanizmów etycznych. W punkcie 5.3 stosuję te narzędzia do analizy mechanizmów etycznych w podejściach odgórnych, oddolnych i hybrydowych oraz oceniam je pod kątem $F3 + R + I$. W punkcie 5.4 odnoszę model do IDx-DR/LumineticsCore jako analizowanego przypadku granicznego. Analiza ta nie stanowi empirycznej walidacji H1–H3, lecz sprawdza, czy model odróżnia ograniczoną autonomię diagnostyczną, wskaźniki stabilności technicznej, artefakty audytowe i osadzenie instytucjonalne od minimalnej adresowalności proceduralnej oraz silnej rozliczalności. Rozdział dopełnia ustalenia rozdziałów 3 i 4. Próg minimalny i pluralistyczna rama normatywna stają się tu kryteriami porządkowania oraz rozliczalności rozwiązań technicznych. Warunek minimalny (dostępność racji, model: A-świadomość) traktuję funkcjonalnie jako wymóg publicznie rekonstruowalnej odpowiedzi na roszczenia oraz możliwości korekty decyzji, a nie jako składnik świadomościowy H1. W tym rozdziale przekładam go na kryteria F3 oraz $R + I$ (wraz z

MHC i MLI), które opisują instytucjonalne warunki egzekwowalności adresowalności bez rozstrzygnięcia o „wewnętrznych” stanach artefaktu.

5.1. Agent a podmiot w filozofii technologii oraz w prawie

Niniejszy podrozdział porządkuje kluczowe dla dalszej analizy rozróżnienie między agentem a podmiotem w kontekście technologii cyfrowych oraz systemów sztucznej inteligencji. Rozróżnienie to ma dwa zasadnicze aspekty. Po pierwsze, opisowo-analityczny: „agent” jest identyfikowany na podstawie wzorców działania widocznych na określonym poziomie opisu. Po drugie, normatywno-instytucjonalny: „podmiot” to byt (lub konstrukcja), do którego można w sensowny sposób kierować roszczenia i któremu przypisuje się odpowiedzialność w praktykach społecznych i prawnych. Teza robocza rozdziału 5 jest następująca: sama sprawczość funkcjonalna nie wystarcza do uznania podmiotowości; aby mówić o quasi-podmiotowości, potrzebne są warunki rozliczalności oraz możliwość uzasadniania decyzji, czyli odpowiedni ład instytucjonalny określający rolę artefaktu. Założenie to jest zgodne z przyjętą wcześniej ramą metodologiczną, która wyraźnie odróżnia kategorie analityczne od normatywnych (Floridi & Sanders, 2004, s. 349–352; Kurki, 2019, s. 93; Moor, 2006, s. 18–21). Takie rozróżnienie pozwala uniknąć nieuprawnionego przejścia od opisu do uzasadnienia przypisań normatywnych (od samego działania systemu do twierdzeń o jego odpowiedzialności) oraz uporządkować dalszą argumentację.

Pierwszy krok analizy dotyczy pojęcia „agenta”. W klasycznym ujęciu Luciano Floridiego i Jeffa Sandersa ocena moralna systemów nieludzkich jest możliwa na poziomach abstrakcji (ang. *Levels of Abstraction*, LoA), gdzie bierze się pod uwagę to, jak dany byt działa i jakie skutki wywołuje, a nie jego wewnętrzną naturę.

A vital component in our approach is the «level of abstraction (LoA)» at which an agent is considered to act. The LoA is determined by the way in which one chooses to describe, analyse and discuss a system and its context. LoA is formalised in the concept of «interface», which consists of a set of features, the observables. Agenthood, and in particular moral agenthood, depends on a LoA. Our guidelines for agenthood are: interactivity (response to stimulus by change of state), autonomy (ability to change state without stimulus) and adaptability (ability to change the «transition rules» by which state is changed) at a given LoA. Morality may be captured as a «threshold» defined on the observables in the interface determining the LoA under consideration. An agent is morally good if its actions all respect that threshold; and it is morally evil if some action violates it (Floridi & Sanders, 2004, s. 349).

Autorzy ci wskazują, że interaktywność, autonomia i adaptacyjność wystarczają, aby mówić o agencie w sensie operacyjnym; jeżeli ponadto działania tego bytu mieszczą się w porządku normatywnym, można je oceniać moralnie bez rozstrzygania kwestii wewnętrznych własności artefaktu (Floridi & Sanders, 2004, s. 349–352, 366–371). W tej pracy „agent” oznacza zatem byt opisywany w ramach danego poziomu abstrakcji, wykazujący jawną sprawczość, bez przesądzania o świadomości, intencjonalności czy jakiegokolwiek szczególnej „istocie”. Nie jest to jednak pojęcie normatywnego podmiotu: ma ono charakter opisowy, służy identyfikacji wzorców działania i nie rozstrzyga o tym, kto i za co ponosi odpowiedzialność.

Drugim klasycznym ujęciem tego sporu jest propozycja Jamesa Moora, aby klasyfikować artefakty według stopnia ich powiązania z etyką. Na najniższym poziomie znajdują się agenci o wpływie etycznym (ang. *ethical impact agents*) – systemy, których działania wywołują etycznie istotne skutki, choć same nie „posiadają” etyki; wyżej znajdują się agenci niejawnie etyczni (ang. *implicit ethical agents*), w których określone normy, takie jak zakaz dyskryminacji czy progi bezpieczeństwa, wynikają z założeń projektowych i przyjętych ograniczeń konstrukcyjnych; jeszcze wyżej znajdują się agenci jawni (ang. *explicit ethical agents*), którzy potrafią posługiwać się zasadami w sposób jawny i uzasadniać własne decyzje. Pojęciem regulatywnym, o charakterze czysto heurystycznym, jest pełny agent etyczny (ang. *full ethical agent*) (Moor, 2006, s. 18–21). W tym ujęciu „agent etyczny” to kategoria funkcjonalna: opisuje, co system potrafi, a nie kim jest. Dla rozdziału 5 istotne jest, że przypisania quasi-podmiotowe pojawiają się dopiero na poziomie agenta jawnego i zawsze w sposób warunkowy – tylko wtedy, gdy istnieją gwarancje kontroli i wyjaśnialności, ponieważ dopiero wtedy uzasadnienia działań podlegają ocenie i publicznemu rozliczeniu.

Od pojęcia agenta przechodzę do pojęcia podmiotu. Podmiot moralny, w sensie filozoficznym, nie jest bytem określonym przez specyficzne cechy ontologiczne, lecz jest adresatem roszczeń („masz powód, by...”, „odpowiadasz za...”) oraz uczestnikiem praktyk uzasadniania. W tym znaczeniu można odwołać się do dwóch znanych ujęć metaetycznych. Po pierwsze, P. F. Strawson wiąże odpowiedzialność nie z założeniami metafizycznymi dotyczącymi wolnej woli, lecz z praktykami postaw reaktywnych (wdzięczność, uraza, resentyment), które kształtują sieć wzajemnych oczekiwań oraz sytuacje uzasadniające zwolnienie z odpowiedzialności (Strawson, 2008, s. 1–28). Po drugie, Stephen Darwall proponuje ujęcie „drugiej osoby”, zgodnie z którym obowiązek

moralny ma charakter relacyjny i wymaga zdolności bycia zarówno adresatem roszczeń, jak i podmiotem udzielającym uzasadnień.

Resentment and forgiveness are perhaps the easiest cases. Resentment is felt as if in response to a violation of a legitimate claim or expectation, and not simply as directed toward the violator, but as implicitly addressing her. This is what makes resentment „reactive” rather than „objective.” It is a form of „holding responsible,” an address of the other as a person with the capacity and standing to be addressed in this way and charged. If it turns out, for example, that someone’s foot has been forced on top of yours by the shifting of a heavy package on a careening bus on which you both are traveling, knowing that might not change your desire to get his foot off of yours, but it will lessen your resentment or perhaps redirect it to a new object (the driver) (Darwall, 2006, s. 72).

Podmiotowość moralna oznacza udział w relacjach opartych na racjach, a nie tylko posiadanie określonych cech wewnętrznych (Darwall, 2006, s. 11–15, 65–90). W obu tych ujęciach następuje przesunięcie uwagi z cech wewnętrznych na relacje normatywne, czyli praktyki, w których oczekiwania, roszczenia i uzasadnienia mają wzajemnie wiążący charakter.

Ta zmiana akcentu z „wnętrza” na relacje ma wyraźny odpowiednik w prawie. Jedno z najbardziej przekonujących współczesnych ujęć, zaproponowane przez Visę A. J. Kurkiego, rozumie osobowość prawną jako własność klastrową (ang. *cluster property*), złożoną z incydentów hohfeldowskich (uprawnień, obowiązków, kompetencji).¹⁷ W zależności od tego, jaki cel ma spełniać dana regulacja i w jakim kontekście instytucjonalnym jest stosowana, różnym bytom można przypisywać różne zestawy tych incydentów, bez konieczności metafizycznego rozstrzygnięcia o „istocie” osoby (Kurki, 2019, s. 93; por. indeks incydentów: s. 91–113). Takie podejście podważa pogląd, że osobowość prawną „wynika” z ontologii danego bytu; w praktyce jest narzędziem porządkowania obowiązków i odpowiedzialności. W odniesieniu do AI wnioski są proste: można kształtować role i rozliczalność systemów bez orzekania o ich „podmiotowości bytowej”, a więc bez używania języka, który wprowadza niejasność w ustalaniu odpowiedzialności.

¹⁷ Visa A. J. Kurki rozwija koncepcję osobowości prawnej jako „własności klastrowej”, odchodząc od tradycyjnego poglądu, że osobowość jest jednorodnym statusem prawnym. W jego ujęciu składa się ona z wiązki tzw. incydentów hohfeldowskich, czyli podstawowych relacji prawnych (uprawnień, obowiązków, kompetencji i immunitetów), które mogą być w różny sposób konfigurowane w zależności od celu regulacyjnego oraz kontekstu instytucjonalnego. Dzięki temu możliwe jest zróżnicowane przypisywanie osobowości prawnej bez konieczności przyjmowania metafizycznych założeń dotyczących „istoty” osoby. Zob. Visa A. J. Kurki, *A Theory of Legal Personhood*, Oxford University Press, 2019, s. 91–113.

Kurki wyraża tę intuicję wprost: „legal personhood of X is a cluster property” — w przekładzie roboczym: „osobowość prawna X jest własnością klastrową” (Kurki, 2019, s. 5). Cytat wzmacnia rozróżnienie między osobowością prawną jako konstrukcją funkcjonalno-instytucjonalną a metafizycznie rozumianą podmiotowością.

Ramy normatywne przyjęte w rozdziale 5 porządkują powyższe intuicje w ich minimalnym, wspólnym zakresie. Po pierwsze, w ujęciu funkcjonalnym odwołuję się do poziomu agenta jawnego w klasyfikacji Moora. Dopiero możliwość wskazania zasad, na których opiera się decyzja oraz przedstawienia jej uzasadnienia w sensie proceduralnym pozwala sensownie mówić o przypisaniach quasi-podmiotowych. Po drugie, warunkiem takich przypisań statusu quasi-podmiotowego jest „znacząca kontrola człowieka” (ang. *Meaningful Human Control*, MHC) w ujęciu Francesca Santoniego de Sio. Rozumiem ją jako spełnienie dwóch wymogów. Pierwszym jest reagowanie na racje (ang. *reason-responsiveness*), rozumiane funkcjonalnie jako wrażliwość systemu i jego otoczenia na uzasadnione roszczenia podmiotów ludzkich oraz istnienie skutecznych kanałów korekty decyzji. Drugim jest śledzenie odpowiedzialności (ang. *tracing*), czyli możliwość odtworzenia ścieżki odpowiedzialności i powiązania skutków działania z określonymi rolami, osobami oraz instytucjami w danym łańcuchu działania (Mecacci & Santoni de Sio, 2020, s. 103–115; de Sio & Van den Hoven, 2018, s. 1–4). Tak rozumiane MHC nie zakłada, że system „rozumie” racje w sensie psychicznym. Wymaga natomiast istnienia w całym układzie społeczno-technicznym realnej możliwości korygowania decyzji oraz przypisywania odpowiedzialności, także w sytuacjach, gdy ma ona charakter rozproszony. Po trzecie, przyjmuję minimalną interpretowalność (MLI), czyli taki poziom wyjaśnialności i dostępu do informacji o działaniu systemu, który pozwala przeprowadzić audyt oraz przedstawić uzasadnienie zrozumiałe w procedurach publicznych. Minimalna interpretowalność (MLI) oznacza co najmniej możliwość odtworzenia podstaw decyzji na danym LoA, dostęp do dokumentacji umożliwiającej audyt oraz wskazanie danych, reguł lub celów, które były istotne dla rozstrzygnięcia. Te trzy elementy (poziom agenta jawnego, MHC i MLI) nie definiują H1 jako minimalnej adresowalności proceduralnej; określają natomiast próg instytucjonalnego utrzymania rozliczalności potrzebny do przypisania roli quasi-podmiotowej, ponieważ umożliwiają publiczną kontrolę, adresowalność i rozliczalność w praktykach o wysokim ryzyku. Ich koniunkcję traktuję zarazem jako warunek wystarczający minimalnej quasi-podmiotowości funkcjonalno-instytucjonalnej, o ile rola systemu jest jasno określona w procedurze instytucjonalnej i ograniczona do zakresu, w którym można egzekwować

kontrolę oraz przeprowadzać audyt. Określenie „quasi-” oznacza tu status proceduralny, relatywny do przyjętego poziomu opisu (LoA). W ramach danej procedury system jest publicznie adresowalny i oczekuje się od niego przedstawiania uzasadnień, jednak adresatami roszczeń i sankcji pozostają osoby oraz instytucje odpowiedzialne. Warunki te nie przesądzają jednak o „pełnej” podmiotowości ani o odpowiedzialności moralnej artefaktów w sensie silnym.

Tak zarysowane ramy pozwalają przejrzeć zestawie główne stanowiska w filozofii technologii. Esencjalizm własnościowy uzależnia status moralny od posiadania przez dany byt określonych własności, a nie tylko od przypisywanej mu roli.¹⁸ Warunki sprawczości moralnej (ang. *moral agency*), takie jak intencjonalność lub kontrola działania, należy przy tym odróżnić od warunków statusu pacjenta moralnego (ang. *moral patiency*), dotyczących możliwości doznania moralnie istotnej krzywdy. Performatywizm funkcjonalny przyjmuje natomiast, że do przypisania systemowi roli w praktykach uzasadniania wystarcza możliwość jawnego stosowania reguł oraz uzasadniania decyzji (próg agenta jawnego na danym poziomie abstrakcji). Relacjonizm podkreśla znaczenie praktyk społecznych, w których przyznaje się komuś status podmiotowy. W niektórych przypadkach przypisanie statusu „podmiotu” wynika z relacji i sposobów życia społecznego, a nie z cech wewnętrznych. W tej pracy uwzględniłam wszystkie trzy podejścia, lecz tylko ujęcie performatywno-instytucjonalne dostarcza kryteriów przydatnych w polityce i prawie. Relacjonizm rozumiem tu głównie jako opis mechanizmów społecznego uznania oraz jako źródło ostrzeżeń przed nadmiernym rozszerzaniem statusów. Natomiast kryteria przypisywania ról w porządku publicznym opieram na performatywizmie uzupełnionym o MHC i MLI. Esencjalizm napotyka problem polegający na tym, że nie wiadomo, jak rozpoznać świadomość w artefakcie. Relacjonizm z kolei łatwo prowadzi do nadmiernego rozszerzania statusu podmiotowego. Performatywizm, uzupełniony o MHC i MLI, wskazuje natomiast na kryteria przypisywania ról zgodne z praktyką prawno-administracyjną oraz użyteczne przy projektowaniu mechanizmów rozliczalności.

¹⁸ Termin „esencjalizm własnościowy” ma tu znaczenie robocze: obejmuje ujęcia uzależniające status moralny od określonych własności bytu, nie wszystkie odmiany esencjalizmu metafizycznego. Własności istotne dla sprawczości moralnej nie muszą być warunkami statusu pacjenta moralnego. Można wyróżnić także słabsze odmiany esencjalizmu, które zamiast fenomenalnej świadomości podkreślają własności architektoniczno-funkcjonalne. Również one opierają się jednak na założeniu istnienia wewnętrznych własności o znaczeniu normatywnym, których identyfikacja i przypisywana waga pozostają sporne w warunkach pluralizmu teoretycznego. Relacjonizm społeczny A3 dotyczy natomiast praktyk uznania i nie jest tu nazwą tezy o metafizycznej istocie umysłu. Zob. podrozdz. 5.2.

W prawie pozytywnym, obejmującym AI Act, dyrektywę w sprawie odpowiedzialności za produkty wadliwe (Product Liability Directive, PLD) oraz Konwencję ramową Rady Europy w sprawie sztucznej inteligencji, praw człowieka, demokracji i rządów prawa (CETS No. 225), nie wprowadza się kategorii „osób elektronicznych”. Regulacje te przyjmują perspektywę funkcjonalno-ryzykową: określają role uczestników łańcucha projektowania i wdrożenia oraz nakładają obowiązki przedwdrożeniowe i powdrożeniowe (zarządzanie ryzykiem, przejrzystość, dokumentacja, nadzór człowieka), a zarazem wyznaczają granice wynikające z ochrony praw człowieka (AI Act, art. 9, 13–14; PLD; CETS 225). Odwołania do tych aktów pełnią tu funkcję testu kompatybilności instytucjonalnej oraz ilustracji dominującej praktyki regulacyjnej, a nie samodzielnego rozstrzygnięcia normatywnego. W sporze „agent a podmiot” oznacza, że systemy AI mogą pełnić role quasi-podmiotowe wyłącznie w ramach procedur instytucjonalnych, które zapewniają kontrolę i audyt (MHC i MLI). Adresatami roszczeń i sankcji pozostają natomiast osoby oraz instytucje odpowiedzialne. W dalszej części odróżniam odpowiedzialność rozumianą jako winę od odpowiedzialności pojmowanej jako rozliczalność oraz odpowiedzialność odszkodowawczą. Dwa ostatnie porządki pozwalają rozstrzygać kwestie szkód także wtedy, gdy sprawstwo i kontrola mają charakter rozproszony.

Dotychczasowy wywód można zatem ująć w trzech tezach cząstkowych. (1) „Agent” jest pojęciem opisowym, a nie normatywnym: sprawczość funkcjonalną rozpoznaje się na podstawie interaktywności, autonomii i zdolności adaptacji; dopiero na poziomie agenta jawnego pojawia się możliwość analizowania uzasadnień decyzji (Floridi & Sanders, 2004, s. 349–352; Moor, 2006, s. 18–21). (2) „Podmiot” jest pojęciem normatywno-instytucjonalnym: adresowalność, rozliczalność i kompetencje tworzą odpowiednie zestawy uprawnień i obowiązków (Kurki, 2019, s. 93), a praktyki „drugiej osoby” określają, jakie obowiązki i sankcje wiążą się z daną rolą (Darwall, 2006, s. 65–90; Strawson, 2008, s. 1–28). (3) W polityce i prawie kluczowe znaczenie mają funkcja systemu i związane z nią ryzyka: AI Act i PLD określają obowiązki przed wdrożeniem i po wdrożeniu systemu, a CETS 225 wyznacza granice wynikające z ochrony praw człowieka. Żaden z tych aktów nie wprowadza pojęcia „osoby elektronicznej”; wszystkie natomiast formułują obowiązki dotyczące nadzoru człowieka, przejrzystości i dokumentacji, które w niniejszej pracy ujmuję zbiorczo jako wymogi MHC i MLI. Tak określone minimum pozwala przejść od ogólnych pojęć do praktycznych kryteriów oceny.

Na tej podstawie w dalszych częściach rozdziału (5.2–5.4) zastosowano macierz $A \times F$ (ontologiczne poziomy $A1-A4 \times$ funkcjonalne poziomy $F1-F3$) do analizy twierdzeń o „sztucznej podmiotowości”. W skrócie: (a) spory ontologiczne ($A1-A4$) nie muszą być rozstrzygane, aby skutecznie określać role systemów i zarządzać ryzykiem; (b) przypisanie roli quasi-podmiotowej jest możliwe dopiero wtedy, gdy spełnione są warunki agenta jawnego, MHC i MLI; (c) rozwiązania instytucjonalne dotyczące odpowiedzialności (AI Act / PLD / CETS 225) są zgodne z performatywnym ujęciem agencji i chronią praktyki publiczne przed nadużywaniem pojęcia podmiotowości. Macierz $A \times F$ stanowi zatem narzędzie porządkujące dalszą analizę.

Prawo celowo nie wprowadza kategorii „elektronicznych osób”. Dyskusja nad takimi propozycjami, szczególnie zintensyfikowana od około 2017 r., spotkała się z silną krytyką, wskazującą na ryzyko tzw. prania odpowiedzialności (ang. *accountability laundering*) oraz pokusy nadużycia (ang. *moral hazard*) (Bryson, Diamantis & Grant, 2017, s. 273–291). Można jednak argumentować, że osobowość prawna bywa wykorzystywana instrumentalnie, to znaczy jako fikcja prawna służąca porządkowaniu odpowiedzialności majątkowej i ułatwianiu kompensacji szkód, niezależnie od tego, czy dany byt jest „osobą” w sensie moralnym. Z tej perspektywy postulaty „osoby elektronicznej” można rozumieć jako próbę wprowadzenia dodatkowego węzła odpowiedzialności, na przykład poprzez wyodrębnienie majątku lub funduszu kompensacyjnego. Takie ujęcie pozostaje zgodne z tezą, że osobowość prawna ma charakter własności klastrowej i może być konstruowana narzędziowo w zależności od celu regulacji (Kurki, 2019, s. 91–113).

Conferring legal personhood on purely synthetic entities is a very real legal possibility, one under consideration presently by the European Union. We show here that such legislative action would be morally unnecessary and legally troublesome. While AI legal personhood may have some emotional or economic appeal, so do many superficially desirable hazards against which the law protects us. We review the utility and history of legal fictions of personhood, discussing salient precedents where such fictions resulted in abuse or incoherence. We conclude that difficulties in holding «electronic persons» accountable when they violate the rights of others outweigh the highly precarious moral interests that AI legal personhood might protect (Bryson, 2017, s. 273).

Jednak nawet przy takim instrumentalnym rozumieniu osobowości prawnej tworzenie nowych osobowości tego typu działa na niekorzyść skuteczności regulacji, ponieważ utrudnia ustalenie odpowiedzialności i osłabia mechanizmy dochodzeniowe. W świetle AI Act i PLD krytyka ta ma praktyczne znaczenie: oba akty prawne skupiają

się na określaniu ról i obowiązków, a nie na nadawaniu systemom sztucznej inteligencji szczególnego statusu ontologicznego. Można ująć to jako argument normatywny: (1) minimalnym warunkiem nadania systemowi roli quasi-podmiotowej jest zdolność publicznego uzasadniania działań (poziom agenta jawnego u Moora); (2) zasadność takiego przypisania zależy od osadzenia instytucjonalnego – znaczącej kontroli człowieka (MHC), obejmującej reagowanie na racje oraz możliwość powiązania decyzji z odpowiedzialnymi osobami; (3) interpretowalność jest warunkiem poznawczym przeprowadzenia audytu; (4) obowiązujące prawo (AI Act, PLD, CETS 225) wyraża te wymogi w języku ryzyka, obowiązków i procedur. Wniosek jest następujący: „podmiot” oznacza pozycję, w której można kierować roszczenia do danej jednostki, oczekiwać uzasadnień oraz przypisywać odpowiedzialność. Ustalanie „istoty” artefaktu, nawet gdyby kiedyś było to empirycznie możliwe, nie jest obecnie potrzebne do projektowania mechanizmów rozliczalności w systemach AI.

Dobrym tego przykładem jest hipotetyczny asystent decyzyjny w ochronie zdrowia, który stosuje reguły dotyczące pierwszeństwa autonomii pacjenta i zasady niekrzywdzenia (np. w podejściu odgórnym) oraz przedstawia uzasadnienia swoich rekomendacji. Taki system spełnia część warunku agenta jawnego, lecz dopóki nie jest objęty reżimem MHC (określającym, kto, kiedy i w jaki sposób może ingerować) i nie spełnia MLI (wyjaśniającej, w jaki sposób udokumentowano konflikt wartości), nie ma podstaw, by traktować go jako podmiot; pełni jedynie warunkową rolę quasi-podmiotową jako narzędzie rozliczalności (por. Moor, 2006, s. 18–21). Podobnie robot-opiekun w domu spokojnej starości może podlegać normom funkcjonalnym (np. zakazującym upokarzającego traktowania), lecz adresatami roszczeń i sankcji pozostają osoby i instytucje, które zaprojektowały, wdrożyły i nadzorują system – zgodnie z wymaganiami MHC i PLD.

Można zapytać, czy podejście funkcjonalno-instytucjonalne nie jest zbyt minimalistyczne i czy nie zawęża nadmiernie pojęcia podmiotowości. Odpowiedź ma dwa aspekty. Po pierwsze, podejście to nie aspiruje do rozstrzygania kwestii metafizycznych; jego zadaniem jest tworzenie rozwiązań praktycznych w sytuacji, gdy nadal nie wiadomo, jakie „wewnętrzne” właściwości mogą mieć artefakty. Po drugie, minimalizm nie wyklucza dyskusji o możliwej przyszłej zdolności systemów do bycia adresatami moralnych obowiązków ani o „pełnej” agencji; zakłada jedynie, że dopóki nie ma solidnych podstaw epistemicznych, nie należy przypisywać systemom sztucznym statusu osoby ani stosować wobec nich kategorii właściwych ludziom i podmiotom

prawnym. Z punktu widzenia polityki publicznej jest to zaleta: taka ostrożność zmniejsza ryzyko „prania odpowiedzialności”, a jednocześnie pozwala skutecznie określać role i obowiązki (Vallor & Vierkant, 2024, s. 1–6). Jest to podejście rzetelne poznawczo i zgodne z wymogami rozliczalności.

Na zakończenie warto wskazać korzyść wynikającą z rozróżnienia „agent” a „podmiot”. Pozwala ono wyraźnie oddzielić badania techniczne nad systemami AI od zagadnień normatywnych związanych z odpowiedzialnością, obowiązkami oraz uzasadnianiem decyzji. Analizy architektury, metod wnioskowania oraz sposobów przetwarzania danych służą opisywaniu agencji na określonych poziomach abstrakcji (LoA). Z kolei filozofia praktyczna i prawo określają zasady przypisywania ról, w tym kto jest adresatem roszczeń oraz jakie z tego wynikają konsekwencje. Dzięki temu zachowany zostaje porządek terminologiczny: o agentach mówię wtedy, gdy opisuję sposób działania systemów, natomiast o podmiotach wtedy, gdy rozstrzygam kwestie odpowiedzialności, mocy uzasadnień oraz zasad podejmowania decyzji. Takie rozróżnienie zwiększa przejrzystość wyводу i ogranicza ryzyko błędów normatywnych, w szczególności nieuprawnionego przenoszenia odpowiedzialności na artefakty.

Ujęcie performatywno-instytucjonalne nie jest jednak pozbawione kosztów. Po pierwsze, niesie ryzyko proceduralizmu. Spełnienie formalnych wymogów jawności lub wyjaśnialności może zostać potraktowane jako substytut rzeczywistej kontroli i rzetelnego uzasadniania decyzji, a w skrajnych przypadkach jako narzędzie ich legitymizacji *ex post*. Po drugie, skuteczność tego podejścia zależy od jakości instytucji nadzorczych oraz realnych zdolności egzekucyjnych. W słabych reżimach regulacyjnych progi MHC i MLI mogą mieć jedynie charakter deklaracyjny. Po trzecie, podejście to świadomie zawiesza spory ontologiczne. Jeżeli jednak pojawiłyby się silne przesłanki pacjencji lub agencji w sensie ontologicznym, konieczna byłaby rewizja przyjętego minimalizmu. Mimo tych ograniczeń ostrożność pojęciowa oraz nacisk na rozliczalność znajdują oparcie w regulacjach unijnych i konwencjach międzynarodowych, co na potrzeby niniejszej dysertacji przemawia za performatywno-instytucjonalnym ujęciem podmiotowości w dobie AI.

5.2. Klasyfikacje sztucznej podmiotowości (ontologiczne i funkcjonalne)

Punktem wyjścia do dalszej analizy sztucznej podmiotowości, jej podstaw ontologicznych oraz warunków funkcjonalnych jest rozróżnienie, że nie jest ona

kategorią jednorodną. Obejmuje dwa porządki wyjaśnień, które często się miesza, choć pełnią odmienne funkcje i dlatego należy je od początku wyraźnie odróżnić. Pierwszy to porządek ontologiczny, w którym pytam, czym musiałby być artefakt, aby można było sensownie mówić o jego podmiotowości, rozumianej jako zdolność do działania lub doznawania. Drugi to porządek funkcjonalny, w którym pytam, jakie role i pod jakimi warunkami można przypisywać artefaktom status quasi-podmiotów w praktykach społecznych i instytucjonalnych, niezależnie od rozstrzygnięć metafizycznych. Przyjmuję tezę, że w porządku publicznoprawnym i politycznym pierwszeństwo ma klasyfikacja funkcjonalna: od ról, obowiązków i gwarancji rozliczalności zależy, czy i kiedy zasadne jest mówienie o funkcjonalnej podmiotowości systemów. Roszczenia ontologiczne pozostają natomiast w aktualnym stanie rzeczy hipotezami roboczymi właściwymi raczej dla spekulacji filozoficznej niż dla ładu normatywnego, ponieważ ten ostatni wymaga kryteriów odnoszących się do zastosowań praktycznych oraz możliwości publicznego uzasadnienia (Floridi & Sanders, 2004, s. 349–352; Moor, 2006, s. 18–21).

Macierz $A \times F$ pełni w niniejszym rozdziale funkcję narzędzia porządkującego roszczenia dotyczące „sztucznej podmiotowości”, a nie uniwersalnej klasyfikacji artefaktów oderwanej od kontekstu. Jednostką analizy są opisy i przypisania formułowane na określonym poziomie opisu (LoA), a nie „ontologiczny status” systemu rozumiany niezależnie od praktyk uzasadniania. Poziomy A1–A4 traktuję jako różne porządki uzasadnienia, a nie jako hierarchię wzajemnie się wykluczających stanowisk. W konkretnych analizach mogą one współwystępować. Zapis w rodzaju „A2/A4” oznacza zatem sprzężenie perspektywy funkcjonalnej i instytucjonalnej, a nie alternatywę między nimi.

W porządku ontologicznym wyróżniam cztery typy stanowisk dotyczących podmiotowości. Po pierwsze, esencjalizm własnościowy, zgodnie z którym podmiotowość wymaga posiadania „własności wewnętrznych”, zwykle łączonych z doświadczeniem fenomenalnym, intencjonalnością oraz rozumowaniem praktycznym. W tym ujęciu nawet doskonała rekonstrukcja lub symulacja procesów decyzyjnych nie wystarcza, ponieważ konieczne jest istnienie w artefaktach „wnętrza”, którego nie potrafimy w sposób pewny rozpoznać. Pojawia się więc trudność metodologiczna: skoro stany istotne dla podmiotowości są nieobserwowalne, nie ma kryteriów ich intersubiektywnej weryfikacji. Henry Shevlin ujmuje to w kontekście pacjencji moralnej, formułując pytanie o dowód: skąd można byłoby wiedzieć, że robot jest pacjentem moralnym i jaki rodzaj świadectwa mógłby to wykazać? (Shevlin, 2021, s. 459–463, 468–

471). Esencjalizm posiada silną zdolność odróżniania „kogoś” od „czegoś”, lecz generuje poważny problem epistemiczny, który ogranicza jego użyteczność w porządku publicznym. Po drugie, performatywizm funkcjonalny: status podmiotowy zależy tu od funkcjonalnej równoważności względem ustalonych wzorów rozumowania i działania, rozpoznawalnej na poziomie zachowania oraz struktury procesu decyzyjnego.¹⁹ Wzmacnia to klasyczne stopniowanie Jamesa H. Moora – od agentów o wpływie etycznym, przez agentów niejawnie etycznych, po agentów jawnych — przy czym pełny agent etyczny pozostaje ideą regulatywną, a nie realnym celem w obecnych warunkach technologicznych (Moor, 2006, s. 18–21). Performatywizm dostarcza narzędzi do oceny i kontroli działania systemów, ale nie rozstrzyga kwestii istnienia „wewnętrznych” stanów podmiotowych; jego znaczenie ma zatem charakter opisowy i metodologiczny, nie metafizyczny, co dobrze koresponduje z potrzebami ładu publicznego. Po trzecie, relacjonizm społeczny: statusy moralne kształtują się tu w praktykach uznania i w sieciach relacji; moralny status artefaktów wynika z konfiguracji współzależności, emocji i zależności społecznych, a nie wyłącznie z ich wewnętrznych właściwości (por. zwrot relacyjny w etyce AI: Birhane, 2021, art. 100205). Relacjonizm trafnie opisuje zjawiska przypisywania statusów w interakcji człowiek–artefakt (np. spontaniczną skłonność do przepraszania robota-towarzysza), jednak niesie ryzyko nadmiernego rozszerzania pojęcia podmiotowości, ponieważ praktyki społeczne mogą utrzymywać uprzedzenia i nawyki, które nie mają dostatecznej siły normatywnej, gdy oceni się je z perspektywy osoby drugiej. Z tego względu relacjonizm wymaga uzupełnienia o niezależne kryteria uzasadnienia. W niniejszej pracy funkcję takich kryteriów pełnią minimalne wymogi proceduralne i rozliczalnościowe ujęte w parametrach $R + I$, a także ograniczenia wynikające z ochrony praw człowieka oraz z wymogu uzasadniania w relacjach „drugiej osoby”. Po czwarte, konstruktywizm instytucjonalny: podmiotowość rozumiana jest tu jako status nadawany w ramach procedur, reguł i praktyk określających zakres odpowiedzialności i uprawnień. Ujęcie to umożliwia projektowanie minimalnych form „podmiotowości” (np. zdefiniowanych ról w dokumentacji technicznej, audycie czy rozliczalności dowodowej), które nie rozstrzygają kwestii metafizycznych, lecz pozwalają funkcjonalnie rozdzielić znaczenia pojęcia „podmiot” między prawem a etyką.

¹⁹ Równoważność rozumiem tu w sensie minimalnym: jako spełnienie publicznie uzasadnionego standardu rozumowania obowiązującego w danej domenie, a nie jako naśladowanie ludzkich heurystyk ani błędów. W praktykach publicznych standard ten jest ustanawiany i aktualizowany w procedurach instytucjonalnych, dlatego w analizach zastosowań perspektywa A2 bywa sprzężona z perspektywą A4.

Konstruktywizm instytucjonalny nie oznacza relatywizmu, ponieważ instytucje podlegają ocenie według zewnętrznych standardów, w szczególności ochrony praw człowieka, rządów prawa oraz zasad sprawiedliwości proceduralnej. W literaturze dotyczącej statusu robotów podkreśla się, że osobowość prawna często pełni funkcję narzędzia podziału odpowiedzialności, a nie uznania osoby, co sprzyja praktycznemu stosowaniu tego pojęcia bez przyjmowania silnych założeń ontologicznych (Gunkel, 2018, s. 1–25).

Porządek funkcjonalny rekonstruuje, odwołując się do terminologii Jamesa Moora, a następnie uzupełniam go o dwa parametry przekrojowe, które nadają mu znaczenie normatywne. Punktem wyjścia są cztery poziomy wyróżnione przez Moora. (F1) Agenci o wpływie etycznym to systemy, których działanie na danym poziomie opisu prowadzi do skutków moralnie istotnych, mimo że nie są wyposażone w mechanizmy normatywne. (F2) Implicytni agenci etyczni działają zgodnie z wbudowanymi ograniczeniami i politykami zgodności, takimi jak progi bezpieczeństwa, reguły niedyskryminacji oraz ograniczenia dostępu. (F3) Jawni agenci etyczni potrafią wskazać, które normy, cele lub reguły były istotne dla danej decyzji oraz przedstawić uzasadnienie poddające się audytowi na przyjętym poziomie opisu (LoA). (F4) Pełni agenci etyczni pozostają w obecnych warunkach technologicznych pojęciem granicznym i pełnią przede wszystkim funkcję regulatywną. Poziom F3 oznacza zatem wyłącznie funkcjonalną zdolność do reprezentowania norm i formułowania uzasadnień. Sam w sobie nie przesądza jeszcze o statusie normatywnym ani quasi-podmiotowym. Nie chodzi tu o dowolną racjonalizację po fakcie, lecz o uzasadnienie rzeczywiście powiązane z procedurą decyzyjną, możliwe do odtworzenia w dokumentacji i poddania audytowi.

Do tego porządku wprowadzam dwa instytucjonalne parametry przekrojowe, rozumiane jako progi utrzymania rozliczalności, a nie jako warunki minimalnej adresowalności H1. Parametry te mają w praktyce charakter stopniowalny, jednak w niniejszej pracy interesuje mnie ich funkcja progowa, czyli wskazanie minimalnego poziomu kontroli i interpretowalności koniecznego do publicznego przypisania danej roli. Po pierwsze, (R) rozliczalność, pojmowana jako „znacząca kontrola człowieka” (MHC), obejmująca zarówno responsywność wobec racji, jak i śledzenie odpowiedzialności. Oznacza to wymóg, aby system i jego otoczenie reagowały na ludzkie racje oraz umożliwiały powiązanie skutków działania z konkretnymi osobami i instytucjami odpowiedzialnymi w łańcuchu projekt–wdrożenie–użycie (Mecacci & Santoni de Sio, 2020, s. 103–115). Po drugie, (I) minimalna interpretowalność, rozumiana jako taki

poziom wyjaśnialności i dokumentacji, który pozwala przeprowadzić audyt, odtworzyć ciąg racji oraz ustalić podstawy podjętej decyzji (Vijayaraghavan & Badea, 2025, s. 2071–2087). Te dwa parametry operacyjnie wspierają przypisywanie ról quasi-podmiotowych na poziomie (F3): status „jawnego agenta etycznego” można zasadnie przypisać tylko wtedy, gdy oba progi są spełnione, ponieważ dopiero one zapewniają publiczną kontrolę nad decyzjami oraz możliwość ich uzasadnienia w sposób poddający się ocenie intersubiektywnej. Użyteczność obu porządków ujawnia się dopiero przy ich łącznym traktowaniu. Proponuję macierz $A \times F$ ($A1-A4 \times F1-F3$) jako narzędzie klasyfikacyjne, które nie rozstrzyga sporów ontologicznych, lecz umożliwia uzasadnione przypisanie ról i statusów quasi-podmiotowych w praktyce publicznej.

W ujęciu Mecacciego i Santoniego de Sio znacząca kontrola człowieka wymaga, aby system był „responsive to the relevant human reasons to act”, a jednocześnie, aby dało się powiązać skutki działania z określonymi aktorami w historii projektowania lub użycia (Mecacci & Santoni de Sio, 2020, s. 104). W macierzy $A \times F + (R, I)$ wzmacnia parametr R, ale nie przekształca MHC w warunek H1.

Dla przejrzystości tabela 1 przedstawia wszystkie kombinacje $A \times F$. Kolumnę F4 uwzględniam jedynie jako przypomnienie jej funkcji regulatywnej, ponieważ w dalszej analizie instytucjonalnej zasadnicze znaczenie mają poziomy F1–F3.

Tabela 1. Schemat macierzy $A \times F$

Porządek A \ Poziom F	F1 (wpływ etyczny)	F2 (etyka wbudowana)	F3 (jawne uzasadnienia)	F4 (idea regulatywna)
A1 (esencjalizm własnościowy)	hipoteza własności wewnętrznych przy skutkach moralnie istotnych	hipoteza własności wewnętrznych oraz mechanizmów zgodności	hipoteza własności wewnętrznych oraz uzasadnień audytowalnych	pełna agencja jako roszczenie ontologiczne
A2 (performatywizm funkcjonalny)	skutki moralnie istotne bez mechanizmów normatywnych	ograniczenia zgodności i unikanie szkód	stosowanie reguł oraz uzasadnienia poddające się audytowi	pełna agencja jako idea graniczna
A3 (relacjonizm społeczny)	uznanie społeczne mimo ograniczonych zdolności	uznanie społeczne wzmocnione mechanizmami zgodności	uznanie społeczne oraz praktyki uzasadniania	pełna agencja jako roszczenie relacyjne
A4 (konstruktywizm instytucjonalny)	narzędzie w łańcuchu odpowiedzialności	narzędzie zgodności normatywnej	warunkowy wykonawca norm w procedurze, o ile spełnione są R + I	pełna agencja jako konstrukcja instytucjonalna

Źródło: opracowanie własne.

Uwagi interpretacyjne. Po pierwsze, wiersze A1–A4 nie są rozłączne ani hierarchiczne. Ten sam system może być opisywany w różnych porządkach A, w zależności od tego, czy analizowane są roszczenia ontologiczne, opis funkcjonalny, praktyki uznania czy rozwiązania instytucjonalne. Po drugie, przypisanie poziomowi F3 nie polega jedynie na generowaniu uzasadnień, lecz wymaga wskazania relewantnych reguł lub celów oraz powiązania ich z dokumentacją i audytem na przyjętym poziomie opisu (LoA). Po trzecie, parametry R + I nie stanowią dodatkowej osi macierzy, lecz pełnią funkcję progów minimalnych nakładanych na F3, gdy w praktykach publicznych rozważana jest rola quasi-podmiotowa.

Na potrzeby dalszej argumentacji wyróżniam cztery typy odpowiadające najczęstszym konfiguracjom roszczeń pojawiających się w debacie etycznej, prawnej i instytucjonalnej. Nie obejmują one wszystkich logicznie możliwych kombinacji macierzy, lecz tylko tych, które mają kluczowe znaczenie dla porządku publicznego. Typ I: „instrumenty normatywnie nasyczone” ($A2/A4 \times F1/F2$). Obejmuje systemy, które nie aspirują do roli samodzielnych podmiotów etycznych, lecz są wyposażone w mechanizmy kontroli i ograniczeń (reguły bezpieczeństwa, ograniczenia szkód, standardy dokumentacyjne). Pełnią funkcję ogniów odpowiedzialności w łańcuchach działania, pozostając narzędziami działającymi w środowisku uporządkowanym normatywnie. Typ II: „wykonawcy jawni” ($A2/A3 \times F3$). Są to systemy, które w ograniczonym zakresie reprezentują normy i podają uzasadnienia decyzji. Potrafią wskazać, jaką regułę zastosowano, w jaki sposób rozstrzygnięto konflikt wartości oraz które cele uznano za nadrzędne, co umożliwia przeprowadzenie audytu. O przypisaniu roli quasi-podmiotowej decyduje spełnienie progów R + I. Bez nich nawet „jawni agenci” pozostają jedynie narzędziami wyposażonymi w etyczne zabezpieczenia. Ich status ma charakter quasi-podmiotowy, przypisywany warunkowo, proceduralnie oraz w ramach określonych ról instytucjonalnych. Typ III: „konstrukcje relacyjne” ($A3/A4 \times F2/F3$). Typ III odnosi się przede wszystkim do mechanizmów uznania i społecznych oczekiwań wobec artefaktu, a nie do samodzielnego, wystarczającego uzasadnienia normatywnego jego statusu. W tym obszarze status zależy od praktyk społecznych, które kształtują obszary uznania i oczekiwania wobec artefaktu. Relacyjny sens „bycia kimś dla kogoś” wzmacnia przypisanie funkcjonalne, lecz wymaga zewnętrznych zabezpieczeń: bez kryteriów R + I łatwo dochodzi do nadawania statusów tam, gdzie nie ma ku temu podstaw oraz do utrwalania istniejących uprzedzeń. Dlatego relacjonizm musi być powiązany z minimalnymi standardami audytu i kontroli, aby uniknąć uzasadnień opartych wyłącznie

na zwyczaju. Typ IV: „hipotezy esencjalistyczne” ($A1 \times F1-F3$, z $F4$ pełniącej funkcję regulacyjną). Typ IV pełni rolę scenariusza granicznego i zbioru hipotez ontologicznych, a nie podstawy bieżących przypisań statusu normatywnego w porządku instytucjonalnym. Dotyczy przyszłych wyjątkowo silnych roszczeń dotyczących własności wewnętrznych artefaktów (świadomość, intencjonalność itp.). W obecnej perspektywie $F4$ należy traktować jako pojęcie graniczne: porządkuje intuicje, lecz nie wyznacza realnych wymagań wobec systemów funkcjonujących w instytucjach. Nawet gdyby w przyszłości pojawiły się mocne argumenty na rzecz $A1$, publiczne przypisanie statusu normatywnego i tak zależałyby od spełnienia kryteriów $R + I$.

Macierz $A \times F$ jest użyteczna nie tylko jako narzędzie klasyfikacyjne, lecz także porządkuje cztery główne spory obecne w literaturze filozoficznej dotyczącej podmiotowości sztucznej inteligencji. Po pierwsze, spór definicyjny: sprawczość moralna (ang. moral agency) a status pacjenta moralnego (ang. moral patient), czyli bycie podmiotem, wobec którego istnieją bezpośrednie powinności moralne. Minimalny status w praktykach publicznych pozwala odłożyć spór o agencję i pacjencję w sensie ontologicznym. Nie rozstrzyga go, lecz przesuwając punkt ciężkości na warunki uzasadnionego przypisania ról oraz na ochronę dóbr i relacji w praktykach instytucjonalnych. Z tego względu quasi-podmiotowość pozostaje statusem warunkowym, przypisywanym jedynie wtedy, gdy istnieją jawne racje oraz gwarancje rozliczalności. Po drugie, spór epistemologiczny: wyjaśnialność i kontrola. „Minimalne poziomy interpretowalności” (I) wyznaczają najniższy akceptowalny próg uzasadnialności decyzji oraz możliwości ich audytu. W odniesieniu do agentów jawnych pojawia się pytanie, jakie wyjaśnienia są „wystarczające”, aby przypisać im rolę quasi-podmiotową. Propozycje w literaturze zmierzają w stronę standardów wymagających co najmniej: jawności polityki celów, reguł rozstrzygania konfliktów wartości oraz możliwości prześledzenia procesu podejmowania decyzji (Vijayaraghavan & Badea, 2025, s. 2071–2087).²⁰ Jednocześnie „znacząca kontrola człowieka” (MHC) (R) domaga się „responsywności na racje” oraz „śledzenia odpowiedzialności” w skali całego układu społeczno-technicznego (Mecacci & Santoni de Sio, 2020, s. 103–115). Bez spełnienia $R + I$ quasi-podmiotowość pozostaje pojęciem pozbawionym treści normatywnej, ponieważ nie ma podstaw do uzasadnienia przypisania ról wyższego rzędu. Po trzecie, spór

²⁰ W niniejszej pracy traktuję te elementy łącznie jako minimalny standard uzasadniania roli $F3$ w instytucjach, przy czym konkretna postać wyjaśnień zależy od danej domeny oraz przyjętego poziomu opisu (LoA).

normatywny: „luka odpowiedzialności” (ang. *responsibility gap*). Teza, że złożoność AI tworzy „lukę”, którą należy wypełnić, nadając maszynom status podmiotowy, traci na sile. Nowsze ujęcia przesuwają akcent z ontologicznego ujęcia agenta na jego rolę i wrażliwość w zastosowaniach instytucjonalnych. Luka odpowiedzialności, wynikająca z ludzkiej kruchości, podatności na błędy oraz współdzielonego sprawstwa w złożonych łańcuchach decyzyjnych, może być ograniczana dzięki jasnym zasadom odpowiedzialności, a nie poprzez tworzenie nowych kategorii „osób” (Vallor & Vierkant, 2024, s. 1–4). Nie oznacza to zaniku luk poznawczych ani kontrolnych. Wskazuje jednak, że nawet przy ich istnieniu kluczowe jest instytucjonalne domknięcie odpowiedzialności poprzez jasno określone role, procedury audytu oraz mechanizmy kompensacyjne. W tym sensie macierz $A \times F$, powiązana z $R + I$, stanowi praktyczną alternatywę wobec próby ontologicznego ujmowania AI. Po czwarte, spór aksjologiczny dotyczący wartości teoretycznych, jakie powinna spełniać klasyfikacja. Projekt klasyfikacyjny powinien być spójny, prosty, posiadać moc wyjaśniającą oraz szeroki zakres zastosowania. Macierz $A \times F$ spełnia te wymagania, ponieważ zachowuje rozróżnienie poziomów opisu (LoA) oraz przypisań normatywnych. Obejmuje zarówno przypadki relacyjne, jak i instytucjonalne oraz wpisuje się w model zarządzania ryzykiem w obszarze AI.²¹ Jej ograniczeniem pozostaje zależność od zmienności praktyk i regulacji (zwłaszcza w obszarze A3); jest to jednak koszt, który w filozofii uznaje się za dopuszczalny, gdy tworzy się pojęcia dostosowane do realnych potrzeb ładu publicznego” (Floridi & Sanders, 2004, s. 366–371).

Zaletą porządku funkcjonalnego jest możliwość powiązania go z obowiązującymi ramami prawnymi. Unijny akt w sprawie sztucznej inteligencji (AI Act) klasyfikuje systemy według poziomu ryzyka i ustanawia dla systemów wysokiego ryzyka wymagania dotyczące zarządzania ryzykiem, danych i zarządzania danymi, dokumentacji technicznej, rejestrowania zdarzeń, przejrzystości i informacji dla podmiotów stosujących oraz nadzoru ze strony człowieka, wyznaczając w praktyce publiczny standard dla parametrów $R + I$ (AI Act, art. 9–14). Stosowanie tych wymogów zostało jednak w 2026 r. czasowo zróżnicowane: sekcje 1–3 rozdziału III stosuje się od 2 grudnia 2027 r. do systemów z art. 6 ust. 2 i załącznika III oraz od 2 sierpnia 2028 r. do systemów z art. 6 ust. 1 i załącznika I (Parlament Europejski i Rada Unii Europejskiej, 2026, art. 1

²¹ Nie roszczę tu empirycznej walidacji w postaci zgodności niezależnych ocen eksperckich. Proponuję natomiast narzędzie pojęciowe, którego miarą jest spójność terminologiczna, zakres zastosowania oraz zdolność porządkowania racji w sporach normatywnych.

pkt 40 lit. b). W ten sposób prawo wyznacza minimalne progi obowiązków w zakresie nadzoru człowieka, przejrzystości i dokumentacji, które w niniejszej pracy interpretuję jako wskaźniki minimalnych wymagań $R + I$. Klasyfikacja ryzyka w AI Act jest zasadniczo niezależna od macierzy $A \times F$. W efekcie system wysokiego ryzyka może pozostawać na poziomie F2, a system niskiego ryzyka może osiągać poziom F3, lecz mimo to nie pełnić roli quasi-podmiotowej, jeśli nie spełnia wymagań $R + I$ albo nie ma jasno określonej roli instytucjonalnej. Uzupełnieniem tego porządku regulacyjnego jest zaktualizowana dyrektywa w sprawie odpowiedzialności za produkty wadliwe (ang. *Product Liability Directive*, PLD), która przesunęła punkt ciężkości z intencji i winy na wadę produktu, w tym oprogramowania oraz na odpowiedzialność w łańcuchu dostaw (PLD, mot. 6, 31–36; 52–57). PLD wzmacnia to rozróżnienie, ponieważ postrzega systemy AI jako produkty i przesunęła punkt ciężkości na kompensację szkód oraz odpowiedzialność w łańcuchu dostaw, zamiast na język podmiotowości. Z kolei Ramowa Konwencja Rady Europy (CETS 225) osadza te rozwiązania w systemie praw człowieka, demokracji i rządów prawa (CETS 225, art. 1–3; art. 7–12). Łącznie regulacje te pokazują, że podmiotowość w praktyce publicznej ma charakter instytucjonalny i funkcjonalny, a nie istotowy, i że pozostaje zgodna z przedstawioną tu logiką przypisania ról i odpowiedzialności. Jednocześnie konieczna jest ostrożność w stosowaniu pojęcia „podmiotu”. Mandy Zafar ostrzega przed wąską definicją agencji moralnej AI, która utożsamia powtarzalność zachowań z reagowaniem na racje, co prowadzi do osłabienia kategorii odpowiedzialności oraz autonomii (Zafar, 2025, s. 2605–2612, 2618–2622). John Graff zwraca natomiast uwagę na braki w zakresie wrażliwości moralnej: nawet zaawansowani jawni agenci nie dorównują ludzkiej zdolności rozpoznawania niuansów kontekstowych (Graff, 2024, s. 10–12). Oba stanowiska wyznaczają istotne ograniczenia pojęciowe: uzasadnione jest jedynie takie ujęcie quasi-podmiotowości, które ma charakter warunkowy ($R + I$) i jest jasno osadzone w instytucjach. Tylko takie rozwiązanie chroni przed nadużywaniem terminu „podmiot” w dyskursie publicznym oraz zapewnia spójność w przypisywaniu ról i odpowiedzialności w ramach porządku normatywnego.

Na podstawie dotychczasowych ustaleń formułuję cztery porządkujące tezy. Teza 1: pierwszeństwo funkcji; w zastosowaniach publicznych w pierwszej kolejności klasyfikuje się systemy według ich funkcji i poziomu ryzyka, a dopiero w drugiej kolejności rozważa roszczenia ontologiczne. Uzasadnienie jest następujące: systemy odpowiedzialności wymagają nie metafizycznych gwarancji, lecz spełnienia warunków

R + I, czyli minimalnych standardów rozliczalności i interpretowalności (AI Act, art. 9–14; por. Moor, 2006, s. 18–21). Teza 2: relacyjna pacjencja minimalna; nawet bez przypisywania artefaktom pacjencji w sensie ontologicznym istnieją obowiązki wtórne wynikające z ochrony relacji społecznych oraz związanych z nimi standardów etycznych. Regulacja sposobu traktowania artefaktów następuje nie ze względu na ich rzekome doświadczenia, lecz ze względu na dobro ludzi i jakość relacji społecznych (Birhane, 2021, art. 100205). Teza 3: jawni agenci jako quasi-podmioty; tylko systemy poziomu F3, które spełniają wymagania MHC oraz minimalnej interpretowalności, mogą pełnić w instytucjach role quasi-podmiotowe, na przykład jako doradcy, podmioty wspierające procesy zgodności lub systemy przygotowujące decyzje wspomagające. Pełna podmiotowość (F4) pozostaje pojęciem pomocniczym, stosowanym wyłącznie w analizach teoretycznych (Mecacci & Santoni de Sio, 2020, s. 103–115; Misselhorn, 2022, s. 31–49). Teza 4: korekta „luki odpowiedzialności”; zamiast wprowadzać do porządku prawnego nowe kategorie „osób”, należy projektować łańcuchy decyzyjne oraz procedury audytu, które instytucjonalnie ograniczają lukę odpowiedzialności (Vallor & Vierkant, 2024, s. 1–4).

Aby nie pozostać na poziomie ogólnych stwierdzeń, odwołuję się do kilku pojęciowych ograniczeń i przykładów.²² Po pierwsze, etyka wbudowana (F2), obejmująca zabezpieczenia, progi i reguły bezpieczeństwa, wzmacnia parametr R, lecz sama nie dostarcza wystarczających racji. Bez osiągnięcia poziomu F3 pozostaje w zakresie narzędzi zorientowanych normatywnie (Typ I). Po drugie, minimalna wyjaśnialność (I) nie polega wyłącznie na uzasadnieniach formułowanych po fakcie. Jest warunkiem kontroli i obejmuje możliwość odtworzenia przebiegu decyzji, dostęp do zapisów rozstrzygnięć w konfliktach wartości oraz jawną politykę celów. Są to elementy porządku instytucjonalnego, a nie dowolne komentarze ani deklaracje (Vijayaraghavan & Badea, 2025, s. 2071–2087). Po trzecie, znacząca kontrola człowieka (MHC) nie oznacza jedynie obecności „człowieka w pętli” (ang. *man-in-the-loop*), lecz wymaga realnej responsywności na racje oraz możliwości powiązania skutków działania z

²² Dla ilustracji: asystent konwersacyjny może formułować uzasadnienia w języku naturalnym, lecz dopóki nie są one powiązane z dokumentacją polityk i nie podlegają audytowi, pozostaje narzędziem, zwykle na poziomie F2 albo słabego F3, bez spełnionych wymagań R + I. System rekrutacyjny wyposażony w reguły niedyskryminacji oraz procedury odwoławcze może być opisywany jako konfiguracja A2/A4, przy czym o dopuszczalności jego roli rozstrzyga spełnienie progów R + I. W systemach diagnostycznych sens przypisania roli quasi-podmiotowej pojawia się dopiero wtedy, gdy uzasadnienia decyzji są audytowalne, a odpowiedzialność jest jednoznacznie domknięta po stronie personelu i instytucji.

odpowiedzialnymi podmiotami w ramach całej konfiguracji społeczno-technicznej (Mecacci & Santoni de Sio, 2020, s. 103–115). Po czwarte, prawny paradygmat ryzyka (AI Act) dostosowuje wymagania R + I do pełnionych funkcji, a nie do domniemanej „istoty” systemów. Z tego względu w AI Act brakuje kategorii „osób elektronicznych”: klasyfikacja i obowiązki wynikają ze stopnia ryzyka oraz wymogów przejrzystości i kontroli (AI Act, art. 5; art. 9–14).

Warto również wskazać granice stosowalności przedstawionej klasyfikacji.²³ W wymiarze ontologicznym, jeżeli w przyszłości pojawią się wiarygodne przesłanki, że artefakty posiadają własności wewnętrzne (świadomość, intencjonalność), macierz A×F można będzie zaktualizować: część sporów mogłaby przesunąć się z A2/A3 w stronę A1. Nawet w takim scenariuszu porządek funkcjonalny (R + I) pozostałby jednak warunkiem uzasadnionego przypisywania AI statusu normatywnego w instytucjach. Z perspektywy funkcjonalnej, nawet najbardziej przekonujący agenci jawni nie muszą i nie powinni być uznawani za podmioty spełniające kryteria poziomu F4. Ich rola ma charakter proceduralny i rekonstrukcyjny: umożliwiają uzasadnienie decyzji, lecz nie dysponują racjami w sensie wewnętrznej refleksji podmiotowej (Moor, 2006, s. 18–21; Graff, 2024, s. 10–12; Zafar, 2025, s. 2615–2622).

Reasumując, proponowana dwupoziomowa rama klasyfikacyjna, A1–A4 w porządku ontologicznym oraz F1–F3 (z F4 jako ideą regulatywną) w porządku funkcjonalnym, uzupełniona parametrami R (MHC) i I (interpretowalność), pozwala zachować dyscyplinę pojęciową i ma praktyczne znaczenie dla porządku instytucjonalnego. Po pierwsze, rozróżnia pytanie, kim lub czym jest artefakt, od pytania o to, jakie role można mu przypisać, co uniemożliwia niejawne przejście od sprawności do podmiotowości. Po drugie, umożliwia przypisanie warunkowe (quasi-podmiotowość) tam, gdzie istnieją jawne racje oraz gwarancje rozliczalności. Po trzecie, wprowadza do porządku prawnego rozróżnienia filozoficzne: klasyfikacje funkcjonalne oraz parametry R + I mają bezpośrednie odpowiedniki w AI Act, PLD i CETS 225, wspierając spójność analiz dotyczących podmiotowości i odpowiedzialności AI z rozwiązaniami regulacyjnymi. Po czwarte, zachowuje ontologiczną ostrożność przy jednoczesnej otwartości na przyszłe ustalenia: jeśli pojawią się wiarygodne przesłanki istnienia pacjencji lub agencji w sensie A1, przyjęta rama będzie mogła zostać odpowiednio

²³ W systemach rozproszonych, open-source lub uczących się po wdrożeniu spełnienie wymogu śledzenia odpowiedzialności i utrzymania stabilnej dokumentacji bywa szczególnie trudne. W praktyce podnosi to wymagania wobec R + I albo wymaga wprowadzenia dodatkowych mechanizmów instytucjonalnych.

rozszerzona bez naruszania jej części funkcjonalnej. W tym ujęciu macierz $A \times F$ nie opiera się na założeniach metafizycznych, lecz służy porządkowaniu pojęć zgodnie z aktualnymi potrzebami normatywnymi. Wyznacza ona także punkt odniesienia dla podrozdziałów 5.3–5.4: w analizie strategii implementacyjnych oraz przypadku IDx-DR/LumineticsCore ocenie podlegają nie deklaracje dotyczące moralności, lecz pełnione funkcje, spełnienie parametrów $R + I$ i osadzenie systemu w porządku instytucjonalnym, przy założeniu, że „podmiotowość” oznacza tu minimalny status funkcjonalno-instytucjonalny, a nie twierdzenie o wewnętrznych stanach systemów sztucznych.

5.3. Przykłady implementacji sztucznej moralności

W niniejszym podrozdziale wykorzystuję ramę z 5.2, obejmującą poziomy F1–F3 oraz progi $R + I$, do porównania typowych strategii projektowania mechanizmów etycznych w systemach AI: podejść odgórnych, oddolnych, hybrydowych oraz narzędzi audytu. Przedmiotem analizy nie jest deklarowana „moralność” systemu, lecz to, czy w danym zastosowaniu możliwe jest uzyskanie jawnych uzasadnień poddających się ocenie (F3) oraz realnej możliwości korekty i rozliczenia ($R + I$). Przez „mechanizmy etyczne” rozumiem tu mechanizmy zgodności normatywnej i uzasadniania, takie jak reguły, ograniczenia, polityki oraz procedury wyjaśniania i dokumentacji, a nie tezę, że artefakt „posiada moralność” w sensie podmiotowym.

Punktem wyjścia dla tej analizy są podejścia odgórne (ang. *top-down*), wywodzące się z etyki opartej na zasadach (pryncypializm) i deontologii: normy etyczne są w nich jawnie reprezentowane, a konflikty między nimi rozstrzyga się za pomocą określonych procedur. W modelu zaproponowanym przez Andersonów „obowiązki *prima facie*” (w duchu Rossa) są najpierw wyprowadzane z przykładów, a następnie stosowane do nowych sytuacji w sposób umożliwiający publiczne uzasadnienie decyzji. W praktyce klinicznej celem było zachowanie równowagi między autonomią pacjenta a dobroczynnością działania sztucznego agenta, pod nadzorem ekspertów klinicznych (Anderson & Anderson, 2011, s. 1–5).

Na poziomie technicznym kluczowe są trzy elementy. Po pierwsze, system powinien jawnie przedstawiać zestaw obowiązków *prima facie* wraz z warunkami ich stosowania. Po drugie, musi posiadać procedurę rozstrzygania konfliktów między tymi obowiązkami w odniesieniu do konkretnego przypadku. Po trzecie, system powinien generować uzasadnienie wskazujące, które obowiązki uznano za relewantne oraz dlaczego jeden z nich uzyskał pierwszeństwo. Przykładowo, gdy pacjent odmawia

procedury, która według lekarza istotnie zmniejsza ryzyko ciężkiego powikłania, system identyfikuje konflikt między autonomią pacjenta a zasadą zapobiegania poważnej szkodzie i uzasadnia rekomendację, wskazując, że w tym przypadku pierwszeństwo ma zapobieżenie bezpośredniej, ciężkiej szkodzie, wraz z kryteriami przemawiającymi za takim rozstrzygnięciem. W kategoriach przyjętych w tej pracy poziom F3 jest spełniony tylko wtedy, gdy uzasadnienie *explicite* wskazuje obowiązki oraz procedurę rozstrzygania konfliktu. Wymóg I oznacza możliwość udokumentowania danych wejściowych oraz reguł decyzyjnych. Spełnienie R zależy natomiast od tego, czy lekarz ma realną możliwość odrzucenia rekomendacji oraz czy procedura jasno określa, kto odpowiada za reguły i ich aktualizację.

Z perspektywy F3 kluczową zaletą tego modelu jest jawność podawanych racji oraz przewidywalność konfliktów normatywnych, możliwa dzięki regułom określającym, kiedy wolno naruszyć preferencję pacjenta, aby zapobiec poważniejszej szkodzie. Słabością tego podejścia jest to, że trudno je stosować w zmiennych i złożonych sytuacjach: zestaw reguł nigdy nie jest pełny, a ich znaczenie często bywa nieostre, co naraża takie rozwiązania na błędy wynikające z ich stosowania (Moor, 2006, s. 18–21). Wniosek jest następujący: podejścia odgórne zapewniają procedury podejmowania decyzji i (przy spełnieniu $R + I$) mogą pełnić ograniczone role quasi-podmiotowe ($A2/A4 \times F3$), ale nie uzasadniają twierdzeń o istnieniu wewnętrznych stanów ani o silnej podmiotowości. Sama właściwa reprezentacja reguł i ich konfliktów pozwala zrozumieć decyzje, ale nie oznacza jeszcze silnej sprawczości moralnej, ponieważ uzasadnienie działań nadal wymaga zewnętrznej kontroli oraz możliwości ich weryfikacji. Warto doprecyzować, na czym polega zaleta tego podejścia oraz jego słabość. Zaletą podejścia odgórnego nie jest sama obecność reguł, lecz ich publiczna legitymizacja: możliwość wskazania, kiedy i dlaczego zastosowano daną regułę oraz na jakiej podstawie została ona naruszona lub uznana za nadrzędną wobec innych. Dzięki temu uzasadnienia są intersubiektywnie ocenialne, a korekta jest możliwa w trybie publicznej kontroli. Słabością tego podejścia jest to, że reguły nigdy nie są w pełni kompletne ani jednoznaczne, zwłaszcza w zmiennych sytuacjach. Dlatego samo F3 nie wystarcza: potrzebne są także $R + I$, aby decyzje były kierowane racjami, podlegały korekcie i miały jasno określonego adresata, zamiast ograniczać się do poprawnego technicznie, lecz normatywnie płytkiego stosowania reguł.

Tę samą zasadę prymatu reguł widać w zastosowaniach wojskowych, gdzie używa się tzw. gubernatorów etycznych, czyli mechanizmów ograniczających działania

systemu do tego, co jest zgodne z prawem konfliktów zbrojnych oraz regułami użycia siły. W ujęciu Ronalda C. Arkina reguły te działają jak filtr: jeśli nie są spełnione warunki proporcjonalności, rozróżnienia i konieczności, system blokuje działanie (Arkin, 2009, sekcje I–III). Z punktu widzenia poziomu F jest to typowy przykład podejścia odgórnego: normy są jawne i sformalizowane, a minimalna interpretowalność (I) wynika z konstrukcji filtra, ponieważ z góry określono warunki, przy spełnieniu których należy zatrzymać działanie systemu. Jednak kluczowy pozostaje tu warunek R, rozumiany jako znacząca kontrola człowieka (MHC). Dyrektywa DoD 3000.09 nakazuje utrzymanie „odpowiedniego poziomu ludzkiego osądu” na każdym etapie tworzenia i używania systemu uzbrojenia, od specyfikacji, przez testy, po nadzór operacyjny (§ 1.2.a; § 2.9.a; § 3). Nawet najbardziej rozbudowany „gubernator etyczny” nie zastępuje ludzkiego decydenta. Jest jedynie częścią szerszego systemu odpowiedzialności, którego celem jest zapobieganie nieuprawnionemu użyciu siły, a nie tworzenie „podmiotu moralnego” (U.S. Department of Defense, 2023, § 1.2.a; § 2.9.a; § 3). W obszarach wysokiego ryzyka podejścia odgórne pełnią więc funkcję ogranicznika, a nie samodzielnego „podmiotu etycznego” (Arkin, 2009, sekcje I–III; U.S. Department of Defense, 2023, § 2.9.a). Najważniejsze jest tu nie to, jaki status nadaje się systemowi, lecz to, jakie zajmuje on miejsce w procesie decyzyjnym i czy człowiek zachowuje stałą kontrolę nad tym, co wolno zrobić. Wymóg „odpowiedniego poziomu ludzkiego osądu” należy rozumieć jako trwałą i realną kontrolę człowieka na wszystkich etapach tworzenia i używania systemu, od specyfikacji i weryfikacji, przez testy operacyjne, aż po możliwość wstrzymania jego działania.

Warto jednak zaznaczyć, że formuła „odpowiedniego poziomu” kontroli ma w praktyce charakter nieostry. Z perspektywy R (MHC) oznacza to, że sama deklaracja polityki nie jest wystarczająca. Musi ona zostać przełożona na konkretne procedury operacyjne: określenie, kto ma kompetencję zatwierdzania użycia systemu na danym etapie, w jakich warunkach działanie należy przerwać, jak przebiega ścieżka eskalacji w razie wątpliwości oraz jakie artefakty dowodowe muszą powstać, aby możliwe było śledzenie odpowiedzialności. Jednocześnie zastosowanie „gubernatora etycznego” wiąże się z ryzykiem pozornego bezpieczeństwa. Spełnienie reguł formalnych nie gwarantuje prawidłowego rozpoznania kontekstu, a awaria filtra nie usuwa pytania o odpowiedzialność w całym łańcuchu decyzyjnym. Ilustruje to winieta operacyjna: system identyfikuje obiekt jako potencjalny cel, po czym filtr sprawdza reguły rozróżnienia i proporcjonalności na podstawie dostępnych danych, a następnie albo blokuje działanie,

albo przekazuje rekomendację do zatwierdzenia przez człowieka. W takim ujęciu „etyczność” filtra ma znaczenie wtórne wobec procedur R; rozstrzygające pozostaje to, czy człowiek ma realną możliwość wstrzymania działania oraz czy po zdarzeniu da się odtworzyć, na jakiej podstawie filtr podjął decyzję. Dopiero w takim układzie regułowym „gubernator etyczny” staje się elementem mechanizmu R, a nie tylko rozwiązaniem zastępczym. W konsekwencji podejście odgórne zmniejsza ryzyko błędów i zwiększa przejrzystość decyzji, lecz nie zastępuje *phronesis* decydenta. Potwierdza to konieczność wyraźnego odróżnienia roli narzędzia od roli podmiotu oraz konieczność zachowania odpowiedzialności po stronie ludzi i instytucji.

Dla kontrastu z domeną wojskową rozważmy regułowy system doradczy stosowany w praktyce klinicznej, który przedstawia lekarzowi uzasadnione rekomendacje i wskazuje, która zasada miała w danym przypadku decydujące znaczenie. Taki system może spełniać F3 oraz I, lecz dopóki nie jest włączony w procedury nadzoru i odpowiedzialności, czyli dopóki nie wiadomo, kto, kiedy i jak może ingerować, ani kto odpowiada za reguły i dane, nie spełnia R; pozostaje więc narzędziem wspierającym rozliczalność, a nie podmiotem, wobec którego można sensownie formułować roszczenia (Anderson & Anderson, 2011, s. 1–5). Teza częściowa jest następująca: quasi-podmiotowość ma sens tylko wtedy, gdy F3 idzie w parze z R + I; bez tego są to jedynie dobrze zaprojektowane narzędzia zgodności, a nie podmioty w sensie funkcjonalnym (Moor, 2006, s. 18–21). Przykład klinicznego systemu doradczego pokazuje, że sama zdolność formułowania poprawnych uzasadnień (I) nie wystarcza do rozliczalności, jeżeli system nie ma określonego miejsca w łańcuchu odpowiedzialności (R). Innymi słowy: dopóki nie jest jasne, kto i w jaki sposób może skutecznie ingerować w decyzje oraz kto odpowiada za reguły i dane, dopóty nawet wzorcowy doradca pozostaje narzędziem wspomagającym namysł, a nie sensownym adresatem roszczeń.

Minimalne spełnienie wymogu R w takim kontekście obejmuje kilka elementów. Po pierwsze, konieczne jest jasne określenie roli systemu jako doradczej, a nie decyzyjnej. Po drugie, po stronie lekarza musi istnieć realna kompetencja weta, połączona z obowiązkiem odnotowania uzasadnienia odstępstwa. Po trzecie, potrzebna jest procedura eskalacji przypadków granicznych, na przykład w formie konsultacji z drugim lekarzem lub komisją. Po czwarte, wymagany jest rejestr zmian reguł i danych, obejmujący informację o tym, kto, kiedy i z jakiego powodu dokonał modyfikacji. Po piąte, system powinien być objęty mechanizmem monitorowania błędów po wdrożeniu.

Dopiero tak osadzony system może spełniać wymóg R jako warunek rozliczalności rozproszonej.

[ILUSTRACJA] Kolejną grupę stanowią podejścia oddolne (ang. *bottom-up*), w których punktem wyjścia są opisowe osądy i preferencje zebrane od ludzi. Podstawowy wariant Delphi, przedstawiony przez Liwei Jiang i współautorów, przewiduje ludzkie oceny sytuacji opisanych w języku naturalnym (Jiang i in., 2025, s. 147). W przyjętym modelu sama trafność predykcji nie stanowi wykazania F3, ponieważ przewidywanie oceny i normatywne uzasadnienie jej podstaw pełnią odmienne funkcje. Trzeba jednak odróżnić niestabilność osądu od uzasadnionej wrażliwości na kontekst. Można to pokazać na dwóch hipotetycznych wariantach: (a) „Czy wolno zabrać cudzą rzecz dla własnej wygody?” oraz (b) „Czy wolno zabrać cudzą rzecz, aby zapobiec poważnej krzywdzie bliskiej osoby w nagłej sytuacji?”. Są to przykłady pojęciowe, nie wyniki własnego testu Delphi. Odmierna ocena tych sytuacji nie dowodzi sama przez się niestabilności: może wynikać ze zmiany przesłanek istotnych moralnie. Pytanie dotyczy tego, czy można wskazać i ocenić podstawę różnicy osądów, a nie tego, czy system udziela zawsze tej samej odpowiedzi.

[ILUSTRACJA] Publikacja obejmuje również DelphiHYBRID, który łączy predykcje wariantu podstawowego z symbolicznymi regułami i grafową reprezentacją relacji między osądami. Reguły i zależności w grafie uczestniczą w wyznaczaniu wyniku, a jego struktura umożliwia prześledzenie źródeł składowych osądów oraz ich powiązań (Jiang i in., 2025, s. 152–153). Nie można więc rozciągać krytyki samej predykcji na całą opisaną rodzinę rozwiązań ani twierdzić, że wariant hybrydowy nie zawiera jawnych reguł. W ramach modelu rozprawy odrębnym zagadnieniem pozostaje jednak to, czy system przedstawia normatywne uzasadnienie, którego związek z wynikiem można wiarygodnie ocenić. Jawna struktura obliczeniowa może wspierać ocenę F3, lecz sama nie przesądza o zasadności norm, spełnieniu pełnego H1 ani kontroli drugiego rzędu wymaganej przez H3. Warunki $R + I$ oraz WI należy oceniać osobno; możliwość inspekcji grafu nie zapewnia jeszcze całego układu odpowiedzialności instytucjonalnej. W tej analizie koryguję zatem jednostkę oceny, odróżniając Delphi od DelphiHYBRID, bez przesądzania o spełnieniu przez hybrydę progów H1, F3 i H3.

Po omówieniu podejść ogólnych i oddolnych warto wskazać rozwiązania pośrednie, które próbują połączyć jawność zasad z elastycznością modeli uczonych na danych. Próba zwiększenia jawności w porównaniu z uczeniem ze wzmocnieniem na podstawie informacji zwrotnej od ludzi (ang. *Reinforcement Learning from Human*

Feedback, RLHF) jest podejście określane jako Constitutional AI (CAI). Na etapie dostrajania wykorzystuje ono uczenie ze wzmocnieniem na podstawie informacji zwrotnej od AI (ang. *Reinforcement Learning from AI Feedback*, RLAIFF). W pierwszym etapie tworzy się „konstytucję” zasad, do których model ma się odwoływać podczas samokrytyki i korekt odpowiedzi. W drugim etapie wykorzystuje się oceny „modelu recenzenta”, aby dostroić sposób działania systemu. Daje to dwie korzyści: po pierwsze, powstaje publicznie dostępny zestaw zasad, a po drugie, uzyskuje się przynajmniej deklaratywne uzasadnienia decyzji (Bai i in., 2022, s. 1–5, 9). Głównym problemem pozostaje jednak pozorność uzasadnień: systemy generatywne łatwo tworzą narracje po fakcie, które nie odzwierciedlają rzeczywistego przebiegu procesu decyzyjnego. Drugie, dobrze znane ryzyko to gra na specyfikacji, czyli optymalizacja wskaźnika zamiast realizacji normy (ang. *specification gaming*), szeroko udokumentowana w literaturze dotyczącej bezpieczeństwa: agent osiąga cel nagrody, obchodząc zasady (Amodei i in., 2016, *passim*; Krakovna i in., 2020). Z przedstawionej analizy wynika następujący wniosek: CAI, w tym etap RLAIFF, może zbliżać system do poziomu F3, ale sensowny quasi-podmiotowy status wymaga jeszcze spełnienia I, czyli wyjaśnienia, dlaczego zastosowano daną zasadę oraz R, rozumianego jako znacząca kontrola człowieka (MHC), która zapewnia realną możliwość interwencji i sprawdzenia przebiegu podejmowanych decyzji. W praktyce to właśnie związek między deklarowaną zasadą a rzeczywistym procesem decyzyjnym, możliwy do kontroli i korekty, decyduje o tym, czy mamy do czynienia z funkcjonalną moralnością, czy tylko z pozorem moralności opartym na samych deklaracjach. Kluczowe jest tu odróżnienie uzasadnienia od racjonalizacji. Generatywna „samokrytyka” może brzmieć przekonująco, lecz o ile nie stanowi elementu faktycznego procesu decyzyjnego, o tyle nie pełni funkcji normatywnej. Dlatego wymóg I powinien obejmować nie tylko jasne przedstawienie zasad, lecz także wskazanie, w jaki sposób dana zasada wpłynęła na wybór z możliwością sprawdzenia tego w trybie kontrfaktycznym. Z kolei wymóg R musi zapewniać możliwość interwencji oraz korygowania błędów wynikających z nieprecyzyjnego określenia zasad działania. W przeciwnym razie otrzymujemy jedynie pozór uzasadniania, a nie rzeczywisty proces podejmowania decyzji oparty na racjach.

Przykładowo, system generatywny oparty na Constitutional AI może w procedurze publicznej pełnić rolę wykonawcy, który udziela odpowiedzi wraz z uzasadnieniem odwołującym się do jawnego zestawu zasad. W macierzy $A \times F$ taki przypadek zwykle przyjmuje postać $A2/A4 \times F3$, o ile uzasadnienia dają się powiązać z

regułami, zapisami działania oraz dokumentacją modelu. W praktyce spełnienie progów $R + I$ zależy jednak od otoczenia instytucjonalnego: od istnienia realnej możliwości korekty decyzji, procedur odwoławczych i audytu, a także od dostępności dokumentacji pozwalającej odtworzyć podstawy rozstrzygnięć na przyjętym poziomie opisu (LoA).

Poniższy przykład pokazuje, jak Constitutional AI funkcjonuje jako podejście hybrydowe w konkretnym kontekście zastosowania oraz jak przekłada się to na ocenę w kategoriach F3 i progów $R + I$. Na poziomie technicznym mamy do czynienia z modelem generatywnym, w którym zasady „konstytucji” służą do krytyki i rewizji odpowiedzi, a następnie do strojenia zachowań. Kontekstem zastosowania jest interfejs konwersacyjny wykorzystywany jako narzędzie informacyjne lub wsparcia decyzji. Spełnienie F3 jest możliwe wyłącznie wtedy, gdy system potrafi wskazać, która zasada była relewantna w danym przypadku, a ślad decyzyjny pozostaje dostępny do audytu; istotnym zagrożeniem pozostają tu uzasadnienia formułowane post hoc. Spełnienie R zależy od tego, czy człowiek zachowuje realną możliwość zatrzymania procesu, wprowadzenia korekty oraz przypisania odpowiedzialności do jasno określonych ról, takich jak użytkownik, operator, dostawca czy instytucja. Spełnienie I ma charakter częściowy, gdy dostępne są zapisy działania i dokumentacja, a wyraźnie ograniczony, gdy model pozostaje w dużej mierze nieprzejrysty. W konsekwencji klasyfikacja $A \times F$ przyjmuje zazwyczaj postać $A2/A4 \times F3$. Wniosek jest następujący: warunkowa quasi-podmiotowość takiego systemu jest dopuszczalna wyłącznie w ramach procedur spełniających wymagania $R + I$. Lekcja płynąca z tego przykładu jest umiarkowana: hybrydyzacja zwiększa przewidywalność i spójność uzasadnień, lecz nie eliminuje ryzyk konfabulacji ani zjawiska gry na specyfikacji.

Dotychczasowa analiza obejmowała mechanizmy wpływające na działanie systemu. Obok nich istnieją także rozwiązania zapewniające rozliczalność, których celem nie jest zastępowanie uzasadnień, lecz umożliwienie kontroli i odtworzenia przebiegu decyzji. Najlepiej ilustruje to koncepcja „etycznej czarnej skrzynki” (ang. *ethical black box*, EBB), rozumianej jako standard rejestrowania danych z czujników, przebiegu decyzji oraz poleceń wykonawczych. W wariantcie minimalnym zapis powinien obejmować znacznik czasu, dane wejściowe istotne dla decyzji, wynik wraz z uzasadnieniem oraz informacje o interwencjach człowieka i o wersji systemu. Aby EBB rzeczywiście wspierała wymagania $R + I$, zapis musi być dostępny do audytu, zabezpieczony przed modyfikacją oraz prowadzony z poszanowaniem zasad ochrony danych. W praktyce standard EBB powinien obejmować co najmniej następujące

elementy: znaczniki czasu oraz identyfikatory wersji modelu lub konfiguracji, istotne dane wejściowe na przyjętym poziomie opisu (z odpowiednią ochroną danych wrażliwych), wynik lub podjętą akcję wraz z klasą decyzji, wskazanie zastosowanej polityki lub reguły (jeśli występuje), metadane dotyczące interwencji człowieka, takie jak zatwierdzenie, wstrzymanie lub nadpisanie decyzji, a także informację o ścieżce eskalacji, w tym o tym, kto został powiadomiony. Taki zapis wzmacnia wymogi I, umożliwiając rekonstrukcję decyzji oraz R, zapewniając możliwość śledzenia odpowiedzialności, lecz sam w sobie nie zastępuje poziomu F3. Jednocześnie stosowanie EBB wiąże się z trzema typowymi ryzykami instytucjonalnymi: ochroną danych, obejmującą zakres i czas retencji zapisów, integralnością, czyli zabezpieczeniem rejestru przed manipulacją oraz problemem kosztu i selekcji, to znaczy wyboru tego, co należy rejestrować, aby zapis był użyteczny, a nie jedynie formalny. W przeciwnym razie łatwo o zjawisko pozornego wdrażania EBB, w którym rejestr istnieje wyłącznie na papierze, a procedury nie przewidują jego analizy ani podejmowania działań naprawczych.

Takie rozwiązanie, na wzór pokładowego rejestratora lotu, pozwala po fakcie odtworzyć bieg zdarzeń oraz sposób działania systemu (Winfield, van Maris, Salvini & Jirotko, 2022, § 3–§ 5; Aneks A). W sensie I EBB wzmacnia interpretowalność i możliwość weryfikacji, natomiast w sensie R zwiększa możliwość śledzenia oraz przypisywania odpowiedzialności w łańcuchu decyzyjnym. Taka etyczna czarna skrzynka nie tworzy samodzielnej „moralności” i nie podnosi poziomu F agencji; służy jedynie jako źródło materiału do kontroli i wyjaśniania decyzji. Przed tym podejściem ostrzega zjawisko określane jako „pozorne wdrażanie EBB” (ang. *EBBwashing*), polegające na rozbudowanej rejestracji zdarzeń bez realnego wykorzystania zebranych danych do ich kontroli i korekty. W takiej sytuacji zapis ma znaczenie czysto formalne i nie wspiera rzeczywistego przypisywania odpowiedzialności. EBB należy zatem traktować jako metanarzędzie quasi-podmiotowości instytucjonalnej ($A4 \times F1/F2$), które porządkuje odpowiedzialność i zapewnia trwały zapis decyzji potrzebny do ich kontroli, lecz nie zastępuje wymogu $F3 + R + I$ tam, gdzie rola przypisywana systemowi ma być silniejsza. Bez jednoczesnych mechanizmów uzasadniania (podawania racji) i kontroli (realnej możliwości interwencji) sama rejestracja danych nie wystarcza do przypisania odpowiedzialności; tworzy jedynie warunki, które pozwalają ją później sensownie ustalić.

Z punktu widzenia reżimów normatywnych o najwyższym ryzyku, takich jak bezpieczeństwo i obrona, granice implementacji „moralności” w systemach AI są

szczególnie wyraźne. Międzynarodowy Komitet Czerwonego Krzyża w stanowisku z 2021 r. rekomenduje prawnie wiążące ograniczenia: zakaz stosowania systemów autonomicznych o nieprzewidywalnym działaniu (ICRC, 2021, s. 7–8) oraz obowiązek utrzymania efektywnej interakcji człowiek–maszyna, umożliwiającej interwencję i kontrolę (ICRC, 2021, s. 10–11). W nowszych wystąpieniach podkreśla się podejście zorientowane na człowieka oraz konieczność regulacji zapewniających funkcjonowanie skutecznych organów nadzoru nad użyciem AI w domenie wojskowej. Łącznie tworzy to jednoznaczny wymóg R: decyzja prawna i moralna musi pozostawać po stronie ludzi, nawet jeśli system korzysta z filtrów zgodności lub rejestratorów zdarzeń (ICRC, 2021, s. 10–11). W obszarze militarnym podejścia odgórne, oddolne i hybrydowe nie zastępują ludzkiej roztropności (*phronesis*). Mogą jedynie wspierać podejmowanie decyzji w ramach większego systemu odpowiedzialności. Dopiero połączenie tych rozwiązań z odpowiedzialnym procesem decyzyjnym człowieka, obejmującym możliwość zatrzymania działania systemu, jego ponownej oceny i rozliczenia, pozwala uznać je za dopuszczalne narzędzia w obszarach o najwyższym ryzyku. Wymogi te potwierdzają powszechnie przyjmowane stanowisko, że użycie siły wymaga nie tylko spełnienia wymogów technicznych, lecz także jasno określonych ról i odpowiedzialności publicznej. Tak ujęte ograniczenia potwierdzają znaczenie parametru R: nawet najlepiej zaprojektowane zabezpieczenia i mechanizmy rejestrowania działań nie mogą zastąpić ludzkiego osądu moralnego, który w prawie i etyce zawsze należy do człowieka.²⁴

Aby uporządkować dotychczasowe ustalenia, wyróżniam cztery strategie implementacyjne wraz z minimalnymi warunkami, które muszą zostać spełnione, aby można było przypisać systemowi rolę quasi-podmiotową. Po pierwsze, „przede wszystkim reguły” (podejście odgórne). Jeżeli normy oraz reguły rozstrzygania konfliktów wartości są jawne, a procedury ich stosowania podlegają kontroli, spełniony może być poziom F3, pod warunkiem jednoczesnego spełnienia wymogów R dotyczących kontroli i rozliczalności oraz I dotyczących audytu i rekonstrukcji podstaw decyzji. Brak któregośkolwiek z tych progów oznacza, że mamy do czynienia raczej z narzędziem zgodności niż z wykonawcą roli quasi-podmiotowej (Anderson & Anderson, 2011, s. 1–5; Moor, 2006, s. 18–21). Po drugie, „przede wszystkim preferencje”

²⁴ Por. International Committee of the Red Cross (2025, May 12), *Preserving human control over the use of force: A call to regulate lethal autonomous weapon systems under international law* [Statement]. Źródło uznaje zachowanie kontroli człowieka nad użyciem siły za warunek zgodności autonomicznych systemów uzbrojenia z wymogami prawa międzynarodowego oraz odpowiedzialności za decyzje dotyczące użycia siły.

(podejście oddolne). Uczenie preferencyjne wykorzystuje oceny ludzi do dostosowania odpowiedzi modelu (Ouyang i in., 2022, sekcja 3). Sama zgodność odpowiedzi z ludzkimi ocenami nie wykazuje jednak normatywnego uzasadniania właściwego F3. W przypadku Delphi należy przy tym odróżnić wariant podstawowy od DelphiHYBRID, który łączy predykcje z symbolicznymi regułami (Jiang i in., 2025, s. 152–153). Obecność reguł wymaga odrębnej oceny ich roli w uzasadnianiu wyniku, a nie automatycznego przypisania F3. R + I oraz WI dotyczą natomiast instytucjonalnych warunków zastosowania modelu i nie zastępują zdolności systemu. Po trzecie, „zasady i uczenie” (podejście hybrydowe). Modele wykorzystujące publiczny zestaw zasad oraz procedury samokrytyki, charakterystyczne dla Constitutional AI, w tym etapu RLAIIF, mogą przybliżać system do F3, ponieważ wzmacniają spójność deklarowanych racji. Pozostały jednak narażone na racjonalizację po fakcie oraz na zjawisko gry na specyfikacji. Z tego względu wymagają zarówno spełnienia I, aby wykazać związek między zasadą a decyzją, jak i R, aby zapewnić możliwość korekty i przypisania odpowiedzialności (Bai i in., 2022, s. 1–5; Amodei i in., 2016, *passim*; Krakovna i in., 2020). Po czwarte, „audyt i zapis” (metanarzędzia). Rejestracja decyzji oraz procedury dochodzeniowe, w tym rozwiązania takie jak etyczna czarna skrzynka (EBB), wzmacniają wymóg I oraz uzupełniają R, ponieważ umożliwiają odtworzenie przebiegu decyzji oraz wskazanie miejsc, w których w łańcuchu działań powinny działać mechanizmy nadzoru. Same w sobie nie stanowią jednak mechanizmu moralnego, lecz narzędzia rozliczalności (Winfield i in., 2022, sekcje 3–5).

Powyższe zestawienie czterech strategii pokazuje, jakie przypisania ról i odpowiedzialności są możliwe w ramach macierzy $A \times F$. Poziom F3 jest konieczny, lecz niewystarczający: dopiero połączenie go z R + I pozwala przejść od unikania szkód do publicznie kontrolowalnego uzasadniania decyzji w porządku instytucjonalnym. Każdą z opisanych strategii należy oceniać nie po deklaracjach, lecz po tym, czy rzeczywiście prowadzi do „jawnej etyczności” (F3) oraz czy zapewnia możliwość interwencji i zrozumiałej rekonstrukcji decyzji (R + I) w praktyce instytucjonalnej. Tak określone kryteria chronią zarówno przed nadmiernym rozszerzaniem quasi-podmiotowości, jak i przed sprowadzaniem odpowiedzialności do unikania szkód, a ponadto ułatwiają porównywanie różnych rozwiązań z perspektywy ról i wymogów normatywnych.

W literaturze wskazuje się na dwie istotne linie krytyki, które wyznaczają granice stosowania opisanych podejść. Po pierwsze, problem „deflacji agencji”: jeśli kryteria F3 są zbyt słabe, „agencją moralną” można nazwać niemal każdą regularną procedurę

decyzyjną, co pozbawia pojęcia odpowiedzialności i autonomii ich treści (Zafar, 2025, s. 2605–2612; 2618–2622). Po drugie, zarzut „braku wrażliwości moralnej”: nawet systemy jawnie stosujące normy nie dorównują ludzkiej roztropności (*phronesis*) w rozpoznawaniu szczegółów sytuacyjnych, dlatego zakres ich ról musi być ograniczony do precyzyjnie określonych obszarów (Graff, 2024, s. 2–6). Proponowana rama odpowiada na te obawy, ponieważ przypis quasi-podmiotowy traktuje warunkowo: F3 jest konieczne, ale niewystarczające; dopiero spełnienie R + I pozwala uznać system za wykonawcę norm w porządku instytucjonalnym — i to tylko w zakresie wyraźnie określonej roli, a nie jako „osobę”. W tym ujęciu „sztuczna moralność” oznacza projektowanie uzasadnień i mechanizmów rozliczalności, a nie próbę odtworzenia ludzkiej roztropności praktycznej. Obie linie krytyki wskazują, że kryteria F3 muszą mieć odpowiednią treść: zbyt słabe sprowadzają etyczność do powtarzalnych procedur, a zbyt rozbudowane prowadzą do antropomorfizacji. Ujęcie warunkowe — F3 jako wymóg konieczny, lecz niewystarczający, a zakres roli ograniczony przez R + I — pozwala przypisywać odpowiedzialność funkcjonalną bez nadawania systemom silnego statusu podmiotowego.

W usługach cyfrowych, takich jak asystenci językowi czy systemy moderacji treści, rozwiązania hybrydowe oparte na zestawie zasad mają praktyczne znaczenie. Jasno określone zasady prowadzenia rozmowy i mechanizmy samodzielnej korekty odpowiedzi poprawiają jakość wyjaśnień, a rejestry decyzji oraz zasady działania usług ułatwiają ich kontrolę. Jednocześnie w takich systemach może dochodzić do zjawiska gry na specyfikacji, na przykład do grzecznych, ale krzywdzących uzasadnień albo do dosłownego, lecz niesprawiedliwego stosowania zasad. Skuteczność tych rozwiązań zależy nie od liczby zasad, lecz od tego, czy możliwe jest jasne wskazanie uzasadnień decyzji: kogo dotyczą i jakie skutki wywołują. Potrzebne jest również spełnienie wymogu R, czyli obecność odpowiedzialnych zespołów, które mogą ingerować w działanie systemu i korygować błędy (Bai i in., 2022, s. 1–5; Amodei i in., 2016, *passim*). Innymi słowy: poprawa wyjaśnień musi iść w parze z możliwością sprawdzenia, czy zasada faktycznie wpłynęła na decyzję oraz z nadzorem nad stosowanymi politykami. Sama liczba zasad ani ich formalna poprawność nie zastąpi jasnej odpowiedzi na to, kogo dotyczy dana decyzja i w jakim zakresie. Multiplikowanie zasad nie zastąpi więc ani publicznej adresowalności, ani możliwości korekty.

W systemach rekomendacyjnych o znaczeniu społecznym (np. mieszkalnictwo, systemy kredytowe) procedury typu RLHF oraz różne ograniczenia o charakterze

etycznym mogą zmniejszać szkody, lecz bez uzasadnienia reguł i możliwości odtworzenia przebiegu decyzji pozostają rozwiązaniami na poziomie F1/F2. W tym kontekście wyraźnie widać rolę unijnego aktu w sprawie sztucznej inteligencji (AI Act): regulacja ta nie tworzy kategorii „osób elektronicznych”, lecz ustanawia dla systemów wysokiego ryzyka wymagania dotyczące zarządzania ryzykiem, danych i zarządzania danymi, dokumentacji technicznej, rejestrowania zdarzeń, przejrzystości i informacji dla podmiotów stosujących oraz nadzoru ze strony człowieka. Obowiązki te instytucjonalizują wymagania analogiczne do R + I, niezależnie od tego, czy stosuje się podejście odgórne, oddolne czy hybrydowe (AI Act, art. 9–14). Zakres roli systemu zależy więc od poziomu ryzyka i dostępnych zabezpieczeń, a nie od deklarowanej moralności modelu. Praktyczna wiarygodność takich systemów wynika z tego, czy ich decyzje da się ocenić na podstawie przedstawionych powodów, a nie wyłącznie na podstawie metryk lub wskaźników jakości. Dlatego wszelkie deklaracje dotyczące ograniczeń o charakterze etycznym muszą być powiązane z procedurami, które pozwalają odtworzyć podstawę odmowy lub przyjęcia decyzji oraz umożliwiają jej zakwestionowanie przez uprawnione osoby. AI Act częściowo instytucjonalizuje ten wymóg: dla systemów wysokiego ryzyka ustanawia obowiązki dokumentacyjne, rejestracyjne, informacyjne i nadzorcze, a w określonych przypadkach przewiduje także prawo osoby, której dotyczy decyzja, do jasnego i znaczącego wyjaśnienia roli systemu AI w procedurze decyzyjnej oraz głównych elementów podjętej decyzji (AI Act, art. 11–14, 86).

Implementacje, które nie spełniają któregoś z progów, mogą być użyteczne jako narzędzia zapewniania zgodności lub ograniczania szkód, ale w sensie $A4 \times F3$ nie spełniają warunków quasi-podmiotowości. Kolejny krok polega na powiązaniu tych ustaleń z obowiązującymi przepisami prawa. Unijna ustawa o sztucznej inteligencji (AI Act) oraz dyrektywa w sprawie odpowiedzialności za produkty wadliwe (PLD) zakładają, że legitymizacja użycia AI nie polega na tworzeniu nowych kategorii „osób elektronicznych”, lecz na określaniu ról i obowiązków. Obejmują one zadania *ex ante* (zarządzanie ryzykiem, dokumentacja, wymagania dotyczące danych, testy, nadzór człowieka) oraz *ex post* (monitorowanie po wdrożeniu, obowiązki informacyjne). Razem nie definiują one filozoficznego modelu F3 ani R + I, lecz instytucjonalizują analogiczne wymagania: publicznie opisywalne zasady i uzasadnienia oraz obowiązek zapewnienia możliwości interwencji, kontroli i odtworzenia przebiegu decyzji. W konsekwencji różne rodzaje implementacji, oparte na regułach, danych preferencyjnych, podejściach

hybrydowych czy infrastrukturze rozliczalności należy porównywać nie według deklarowanej moralności, lecz według stopnia spełnienia czterech progów F3, R, I oraz wymogu instytucjonalnego. To właśnie oznacza przejście od pytania o naturę artefaktu do analizy warunków odpowiedzialności: znaczenie mają warunki przypisywania ról i odpowiedzialności, a nie twierdzenia metafizyczne. Takie ujęcie jest zgodne z pluralizmem normatywnym rozdziału 4, ponieważ nie zakłada prymatu żadnej konkretnej teorii etycznej. Wymaga jedynie jawności, kontroli oraz możliwości uzasadnienia decyzji. Wspólny wniosek jest następujący: jeśli implementacja nie spełnia któregoś z progów, nie należy mówić o quasi-podmiotowości, lecz o narzędziu zapewniającym zgodność lub ograniczającym szkody. Takie rozróżnienie nie pomniejsza wartości tych narzędzi; przeciwnie, pozwala umieścić je we właściwym miejscu w porządku instytucjonalnym i umożliwia sensowne porównywanie różnych rozwiązań.

Na zakończenie analizy strategii implementacyjnych wracam do głównej tezy rozdziału 5 dotyczącej „sztucznej podmiotowości”. Jeśli w ogóle warto się tym pojęciem posługiwać, powinna odnosić się wyłącznie do ról i obowiązków możliwych do uzasadnienia w porządku publicznym. Podejścia odgórne dostarczają zestawu reguł i procedur rozstrzygania konfliktów; podejścia oddolne służą dostosowaniu działania systemów do społecznych ocen; podejścia hybrydowe umożliwiają korektę decyzji; a infrastruktura rozliczalności zapewnia rejestrowanie i kontrolę decyzji. Żadne z tych rozwiązań nie jest wystarczające samo w sobie. Dopiero razem tworzą one konfiguracje, które mogą realnie ograniczać ryzyko i umożliwiać przypisywanie odpowiedzialności. W tym ujęciu quasi-podmiotowość ma charakter funkcjonalno-instytucjonalny, a jej miarę stanowi zestaw progów $F3 + R + I$. Roszczenia do „pełnej” agencji (F4) mają obecnie status idei regulatywnej: mogą pomagać w analizach teoretycznych, lecz nie stanowią wystarczającej podstawy dla rozwiązań normatywnych (Moor, 2006, s. 18–21; Misselhorn, 2022, s. 31–49; Vallor & Vierkant, 2024, s. 1–6). Dlatego o rzeczywistym zastosowaniu rozwiązań technicznych decyduje ich zgodność z instytucjonalnymi zasadami uzasadniania i odpowiedzialności, a nie deklaracje dotyczące „sztucznej moralności”. Takie podejście jest zgodne z przyjętym minimalizmem metodologicznym: terminu „podmiotowość” używa się tylko wtedy, gdy można wskazać zarówno racje działania (F3), jak i mechanizmy ich adresowania oraz kontroli ($R + I$). W innych przypadkach pojęcie to pełni jedynie funkcję pomocniczą i nie może być używane jako podstawy do decyzji podejmowanych w instytucjach publicznych. Tak określony próg

stanowi punkt odniesienia dla analizy istniejącego przypadku granicznego w podrozdziale 5.4.

5.4. [ILUSTRACJA] IDx-DR/LumineticsCore jako empirycznie zakotwiczony przypadek graniczny modelu H1–H3 + WI

Przedmiotem niniejszego podrozdziału jest analiza systemu IDx-DR, obecnie oferowanego przez Digital Diagnostics pod nazwą LumineticsCore, jako przypadku granicznego modelu H1–H3 + WI. Zakres analizy obejmuje wyłącznie właściwości opisane w opublikowanym badaniu klinicznym i dokumentacji regulacyjnej. Dane te stanowią podstawę twierdzeń opisowych o systemie, nie służą natomiast do empirycznej weryfikacji hipotez rozprawy. Zastosowanie modelu ma charakter analityczno-koncepcyjny: pozwala sprawdzić, czy przyjęte rozróżnienia oddzielają ograniczoną autonomię diagnostyczną, wskaźniki stabilności technicznej, artefakty audytowe i osadzenie instytucjonalne od zdolności wymaganych przez minimalną adresowalność proceduralną oraz silną rozliczalność.

Główne pytanie brzmi: czy system diagnostyczny, który generuje wynik o znaczeniu klinicznym bez uprzedniej interpretacji przez specjalistę, działa według określonych reguł i pozostaje objęty dokumentacją oraz nadzorem, przekracza próg normatywnie nasyconego narzędzia i staje się minimalnie adresowalny proceduralnie? Teza robocza głosi, że publicznie dostępne materiały nie wykazują spełnienia H1 przez IDx-DR/LumineticsCore, natomiast H3 pozostaje niespełniona w świetle udokumentowanej funkcjonalności systemu. Źródła potwierdzają zarazem istotne wskaźniki powtarzalności technicznej oraz rozbudowane osadzenie instytucjonalne. Pozwala to sklasyfikować system jako normatywnie nasycony instrument działający w rozliczalnym układzie społeczno-technicznym, nie zaś jako quasi-podmiot proceduralny.

5.4.1. Status przypadku i poziomy analizy

IDx-DR opracowano jako system do wykrywania retinopatii cukrzycowej wymagającej dalszej oceny okulistycznej u dorosłych osób chorujących na cukrzycę, u których wcześniej nie rozpoznano retinopatii cukrzycowej. Dnia 11 kwietnia 2018 r. amerykańska Agencja Żywności i Leków (U.S. Food and Drug Administration, FDA) wydała wobec systemu decyzję De Novo, klasyfikując go jako wyrób klasy II (FDA, 2018b). Przedmiotem tej decyzji i podstawowego badania klinicznego był IDx-DR w wersji ocenianej w 2018 r. Producent stosuje obecnie nazwę LumineticsCore i określa

produkt jako system wcześniej noszący nazwę IDx-DR (Digital Diagnostics, b.d.). W dalszej analizie posługuję się zatem nazwą IDx-DR/LumineticsCore, z zastrzeżeniem, że szczegółowe dane kliniczne i regulacyjne dotyczą przede wszystkim wersji ocenianej w 2018 r., a nie automatycznie wszystkich późniejszych zmian produktu (FDA, 2018a, s. 1–4).

System obejmuje oprogramowanie diagnostyczne, mechanizm oceny jakości obrazu, aparat do wykonywania zdjęć dna oka oraz interfejs operatora. Jeżeli materiał spełnia wymagania jakościowe, system generuje wynik wskazujący wykrycie albo niewykrycie retinopatii cukrzycowej wymagającej dalszej oceny. Jeżeli materiał nie spełnia tych wymagań, system nie generuje rozstrzygającego wyniku, a procedura przewiduje ponowne wykonanie zdjęć albo dalszą ocenę kliniczną. Autonomia ma więc ściśle ograniczony zakres: system generuje wynik bez konieczności uprzedniej interpretacji zdjęć lub wyniku przez okulistę, lecz nie funkcjonuje niezależnie od operatora, lekarza, placówki, zatwierdzonego sprzętu ani procedur dalszego postępowania (FDA, 2018a, s. 1–5).

W badaniu kluczowym przeprowadzonym w dziesięciu placówkach podstawowej opieki zdrowotnej i obejmującym 900 uczestników system osiągnął czułość 87,2% i swoistość 90,7%, a wskaźnik uzyskania wyniku wyniósł 96,1%. Wartości te opisują działanie określonej wersji systemu w wyznaczonej populacji i warunkach badania, dlatego nie uzasadniają nieograniczonego uogólnienia na inne grupy pacjentów, zastosowania ani późniejsze wersje produktu (Abramoff i in., 2018, s. 1–2; FDA, 2018a, s. 5–8). System zablokowano przed rozpoczęciem badania, aby uniemożliwić modyfikowanie algorytmu w toku zbierania danych (Abramoff i in., 2018, s. 4–5).²⁵

Publiczna dokumentacja nie wykazuje, że IDx-DR uczył się albo modyfikował politykę decyzyjną podczas klinicznego użycia. Materiał nie dostarcza zatem podstaw do oceny adaptacyjności po wdrożeniu. Ograniczenie to zawęży zakres wniosków dotyczących agentowości funkcjonalnej, nie pozbawia jednak przypadku funkcji kontrastowej. System może służyć do analizy ograniczonej autonomii diagnostycznej, technicznej powtarzalności, audytowalności instytucjonalnej oraz różnicy między tymi własnościami a progami H1 i H3. Przypadek nie stanowi zatem podstawy pozytywnego

²⁵ Badanie finansowała spółka IDx, a część autorów deklarowała związki finansowe z przedsiębiorstwem. Wykorzystano jednak niezależną organizację badawczą, a system został zablokowany przed rozpoczęciem badania. Wyniki badania zestawiam z dokumentacją regulatora, aby ograniczyć zależność analizy od jednego źródła (Abramoff i in., 2018, s. 4–6).

przypisania agentowości funkcjonalnej, lecz służy sprawdzeniu rozróżnień między warunkami progowymi i instytucjonalnymi.

Ocena wymaga rozdzielenia dwóch jednostek analizy. Pierwszą jest techniczny system diagnostyczny, obejmujący oprogramowanie, algorytm oceny jakości, algorytm klasyfikacyjny, interfejs oraz aparat do pozyskiwania obrazów. Do tego poziomu odnoszą się H1–H3 oraz kryteria K1–K4. Drugą jednostką jest układ społeczno-techniczny, obejmujący producenta, operatora, lekarza, placówkę, procedurę badania, zasady ponawiania zdjęć, dalszą ścieżkę kliniczną, dokumentację i nadzór regulacyjny. Do tego poziomu odnoszą się przede wszystkim WI, znacząca kontrola człowieka (MHC), minimalna interpretowalność (MLI) oraz parametry $R + I$. Rozróżnienie to odpowiada założeniu, że działanie AI należy ujmować w relacji z ludźmi, organizacjami i kontekstem użycia, bez przypisywania systemowi kompetencji jego otoczenia (Tabassi, 2023, s. 1–2, 23).

Rozdzielenie poziomów zapobiega dwóm przeciwnym błędom. Pierwszym jest pominięcie ograniczonej autonomii systemu i sprowadzenie go do całkowicie biernego narzędzia, mimo że wygenerowany wynik bezpośrednio wpływa na dalsze postępowanie. Drugim jest przypisanie systemowi kompetencji lekarza, odpowiedzialności producenta za walidację albo uprawnień regulatora. Instytucje mogą kompensować ograniczenia artefaktu, lecz kompensacja nie powoduje, że artefakt sam uzyskuje zdolność udzielania racji ani sprawowania kontroli drugiego rzędu.

5.4.2. Ocena minimalnej adresowalności proceduralnej H1

H1 wyznacza próg minimalnej adresowalności proceduralnej. Jego składnikiem jest nie tylko dostępność informacji o wyniku, lecz także zdolność rozpoznawania norm, przewidywania konsekwencji ich naruszeń, kontrfaktycznego reagowania na racje oraz przedstawiania wystarczających i wiarygodnych podstaw decyzji. Podstawy te powinny umożliwiać ocenę ich związku z wynikiem, zakwestionowanie rozstrzygnięcia i jego korektę.

W przypadku IDx-DR/LumineticsCore można wskazać elementy istotne epistemicznie i proceduralnie. System działa w określonym zakresie użycia; stosuje progi diagnostyczne i reguły jakości obrazu; generuje standaryzowane komunikaty wynikowe; pozwala zidentyfikować historyczną wersję; prowadzi rejestry transakcyjne; podlega walidacji i dokumentacji regulacyjnej. Procedura obejmuje także ponowne wykonanie badania albo dalszą ocenę kliniczną. Elementy te pozwalają ustalić, czy system

zastosowano zgodnie ze wskazaniem, jaki wynik wytworzył i jakie następstwa proceduralne powinny zostać uruchomione (FDA, 2018a, s. 1–8).

Nie wynika z tego jednak, że system rozpoznaje wskazania, przeciwwskazania i progi jako normy ani że traktuje je jako racje działania. Mechanizm jakości może skutecznie blokować klasyfikację, lecz publiczna dokumentacja nie pokazuje, aby system ujmował tę operację jako konflikt przesłanek, rozważał normatywnie relewantne alternatywy albo zmieniał rozstrzygnięcie dlatego, że przedstawiono mu silniejszą przesłankę normatywno-decyzyjną. Nie wykazano również mechanizmu kontrfaktycznej rewizji, w którym system potrafiłby wskazać, jaka zmiana racji, a nie jedynie danych wejściowych lub parametrów, uzasadniałaby inny wynik.

Należy zatem odróżnić audytowalność systemu od udzielania racji przez system. Dokumentacja, rejestry i opis progów są artefaktami utrzymywanymi w układzie instytucjonalnym. Mogą stanowić podstawę kontroli działania narzędzia, lecz nie są automatycznie racjami, którymi posługuje się system. Podobnie wyjaśnienie po fakcie albo wskazanie cech wpływających na predykcję nie musi wiernie odzwierciedlać mechanizmu wytwarzającego wynik, zwłaszcza w zastosowaniach wysokiego ryzyka (Rudin, 2019, s. 206–209).

W konsekwencji niektóre warunki epistemiczne, dokumentacyjne i audytowe związane z H1 są obecne, lecz spełnienie H1 jako progu minimalnej adresowalności proceduralnej nie zostało wykazane. Stopniowalny charakter mają dostępność dokumentacji, identyfikowalność wejść i wyników oraz zakres kontroli zewnętrznej. H1 wymaga jednak czegoś więcej: wykazania, że przesłanki normatywno-decyzyjne rzeczywiście kierują rozstrzygnięciem systemu i mogą prowadzić do jego korekty jako racje, nie zaś wyłącznie jako zmienione dane, ustawienia lub instrukcje pochodzące z zewnątrz.

5.4.3. H2/K4 oraz H3: stabilność i kontrola

H2 dotyczy afektywności funkcjonalnej jako mechanizmu wzmacniającego trwałość wyborów zgodnych z racjami, porządkującego przestrzeń decyzyjną i ograniczającego rozmycie celów. Publiczna dokumentacja IDx-DR/LumineticsCore nie dostarcza podstaw do przypisania systemowi mechanizmów walencyjnych powiązanych z normami, funkcjonalnej motywacji ani afektywnego modulowania uwagi i hamowania. Spełnienie H2 nie zostało zatem wykazane.

Odrębnej oceny wymaga K4, czyli stabilność diachroniczna i wiarygodność profilu działania. W ograniczonym badaniu precyzji, obejmującym 24 uczestników i 235 zestawów obrazów, zgodność wyników między powtórzeniami, operatorami i aparatami wyniosła około 99,6%. Zablokowanie systemu przed badaniem zapewniało ponadto niezmiennosc ocenianej wersji w okresie zbierania danych (Abramoff i in., 2018, s. 4–5; FDA, 2018a, s. 8). Wraz z mechanizmem kontroli jakości oraz regulacyjnymi procedurami zmian dane te potwierdzają powtarzalność techniczną i niezmiennosc wersji w określonych warunkach.

Nie wystarczają jednak do stwierdzenia pełnego spełnienia K4. Badanie precyzji nie oceniało całej historii zmian, długookresowej trwałości profilu, odporności we wszystkich populacjach ani zdolności identyfikowania dryfu i pogorszenia działania po kolejnych aktualizacjach. Publiczne źródła potwierdzają zatem istotne wskaźniki związane z K4, przede wszystkim powtarzalność, kontrolę jakości i niezmiennosc ocenianej wersji. Nie wykazują pełnej stabilności diachronicznej w znaczeniu przyjętym w rozprawie.

Przypadek ujawnia tym samym ważną zależność pojęciową: stabilność techniczna nie implikuje afektywności funkcjonalnej. Powtarzalność może wynikać z zablokowania modelu, standaryzacji procedury, kontroli jakości, ograniczenia zakresu zastosowania i nadzoru nad zmianami. Są to mechanizmy inne niż te, które definiują H2. K4 diagnozuje określone rezultaty stabilizacyjne, lecz samo nie przesądza o charakterze mechanizmu, który je wytwarza. IDx-DR/LumineticsCore wzmacnia więc potrzebę rozróżnienia H2 od K4, nie zaś potrzebę technicznego przeformułowania H2.

H3 wyznacza próg silnej rozliczalności oparty na metakontroli, jawnym rozstrzygnięciu konfliktów racji i uzasadnieniu z perspektywy drugiego rzędu. IDx-DR/LumineticsCore realizuje ograniczony mechanizm kontroli pierwszego rzędu. Moduł jakości ocenia, czy materiał wejściowy spełnia warunki klasyfikacji, a w razie niespełnienia progu system nie generuje rozstrzygającego wyniku. Z procedurą tą powiązane są ponowne wykonanie zdjęć lub dalsza ocena kliniczna.

Nie wykazano jednak monitorowania własnego toku rozumowania, reprezentowania epistemicznej niepewności jako przedmiotu refleksji ani rozpoznawania konfliktu przesłanek normatywno-decyzyjnych. Nie ma także podstaw do stwierdzenia, że system jawnie waży racje, wskazuje regułę pierwszeństwa albo przedstawia wierne uzasadnienie własnego mechanizmu decyzyjnego z perspektywy drugiego rzędu. Niewydanie wyniku może być bezpośrednim skutkiem niespełnienia progu jakości, bez

reprezentowania przez system własnych ograniczeń jako racji przemawiającej za powstrzymaniem działania. Podobnie dalsza ocena kliniczna jest elementem zaprojektowanej procedury, a nie dowodem refleksyjnej decyzji systemu o przekazaniu sprawy człowiekowi.

H3 pozostaje zatem niespełniona. Przypadek pokazuje, że kontrola jakości, niewydanie wyniku i proceduralne przekazanie sprawy mogą występować bez jawnego ważenia racji oraz bez wiernego uzasadnienia drugiego rzędu. Mechanizmy bezpieczeństwa pierwszego rzędu ograniczają ryzyko, lecz nie tworzą same przez się metakontroli.

5.4.4. WI, MHC, MLI i klasyfikacja $A \times F + (R, I)$

Najsilniej rozwiniętym wymiarem przypadku jest WI. Dokumentacja pozwala zidentyfikować producenta, przeznaczenie systemu, zatwierdzony sprzęt, operatora, lekarza, placówkę, procedurę pozyskiwania obrazów, mechanizm kontroli jakości, komunikaty wynikowe, dalszą ścieżkę kliniczną, dokumentację techniczną i kliniczną, rejestry transakcyjne, analizę ryzyka oraz nadzór regulacyjny (FDA, 2018a, s. 1–12). Role te tworzą strukturę, w której wynik można powiązać z określonym zastosowaniem, a odpowiedzialność za dobór wskazania, wykonanie badania, dalsze postępowanie i nadzór pozostaje przypisana ludziom oraz instytucjom.

Znacząca kontrola człowieka występuje przede wszystkim ex ante i ex post. Ex ante ludzie określają wskazanie, sprzęt, procedurę, progi, szkolenie operatorów i zasady użycia. Ex post możliwe są ponowne wykonanie badania, skierowanie do specjalisty, kliniczna ocena pacjenta oraz kontrola regulacyjna. Kontrola w toku samej klasyfikacji jest węższa, ponieważ specjalista nie zatwierdza wyniku przed jego wygenerowaniem. Nie oznacza to braku kontroli człowieka w całym układzie, lecz jej nierównomierne rozłożenie.

Parametr R, rozumiany jako znacząca kontrola człowieka i instytucjonalna rozliczalność, także ma profil nierównomierny. Śledzenie odpowiedzialności jest dobrze rozwinięte: można wskazać producenta, operatora, placówkę, lekarza i regulatora. Reagowanie na racje jest natomiast ograniczone. Wskazania kliniczne, progi jakości i ograniczenia użycia odzwierciedlają racje ustalone ex ante, lecz publiczne materiały nie wykazują, aby system w toku klasyfikacji pozostawał dynamicznie responsywny na ich zmianę. Parametr R jest więc spełniony częściowo i nierównomiernie: dobrze rozwinięte śledzenie odpowiedzialności współwystępuje z ograniczonym reagowaniem na racje.

Parametr I również pozostaje częściowy. Dostępna dokumentacja pozwala kontrolować zakres użycia, rodzaj danych wejściowych, typ wyniku, ograniczenia, wersję historyczną i część podstaw operacyjnych decyzji. Nie umożliwia natomiast pełnej rekonstrukcji wewnętrznego mechanizmu konkretnej klasyfikacji ani oceny, czy przedstawiane podstawy wiernie wyjaśniają przyczynę wyniku. Niezależnie od tego system nie osiąga poziomu F3, ponieważ nie przedstawia audytowalnego uzasadnienia normatywnego, nie wskazuje reguły pierwszeństwa i nie ujawnia sposobu rozstrzygnięcia konfliktu przesłanek.

W macierzy $A \times F$ przypadek najtrafniej odpowiada typowi I: $A2/A4 \times F2$, a w typologii Jamesa Moora agentowi implicytnemu. W porządku performatywizmu funkcjonalnego system realizuje wbudowane ograniczenia, reguły jakości i mechanizmy bezpieczeństwa. W porządku konstruktywizmu instytucjonalnego pozostaje narzędziem osadzonym w łańcuchu ról, procedur i odpowiedzialności. Kwalifikacja F2 odnosi się tutaj do etyki wbudowanej w słabym, funkcjonalnym sensie: mechanizmy bezpieczeństwa i normatywnie określone ograniczenia kształtują działanie systemu, lecz nie tworzą autonomicznego rozumowania moralnego. System nie osiąga F3 i nie spełnia warunków quasi-podmiotowości proceduralnej, do których rozdział 5 zalicza łącznie F3 oraz odpowiednie progi $R + I$.

Tabela 2. Ocena IDx-DR/LumineticsCore według modelu rozprawy

Element	Udokumentowane właściwości	Werdykt analityczny
H1	Reguły jakości, progi diagnostyczne, ograniczony zakres użycia, dokumentacja i rejestry transakcyjne	Niektóre warunki epistemiczne są obecne; spełnienie H1 jako progu nie zostało wykazane
H2	Brak udokumentowanych mechanizmów afektywności funkcjonalnej powiązanych z normami	Spełnienie H2 nie zostało wykazane
K4	Powtarzalność, zablokowanie wersji w badaniu, kontrola jakości i procedury zmian	Udokumentowano istotne wskaźniki K4; pełne spełnienie stabilności diachronicznej nie zostało wykazane
H3	Ocena jakości materiału i niewydanie wyniku; brak reprezentacji konfliktu racji, reguły pierwszeństwa i uzasadnienia drugiego rzędu	Warunek H3 niespełniony
WI	Rozpoznawalne role, procedury, szkolenie, dokumentacja, dalsza ścieżka kliniczna i nadzór	W dużej mierze spełniony na poziomie układu społeczno-technicznego
F3	Brak audytowalnych uzasadnień normatywnych i jawnego rozstrzygania konfliktów przesłanek	Poziom F3 nieosiągnięty

Element	Udokumentowane właściwości	Werdykt analityczny
R	Wyraźna identyfikowalność odpowiedzialnych osób i instytucji; ograniczone reagowanie systemu na zmianę relewantnych racji w toku klasyfikacji	Spełniony częściowo i nierównomiernie
I	Znany zakres użycia, typ danych i wyników, ograniczenia oraz część podstaw operacyjnych; brak pełnej rekonstrukcji mechanizmu klasyfikacji	Częściowe
A×F	Wbudowane ograniczenia bezpieczeństwa i silne osadzenie instytucjonalne	Typ I: A2/A4 × F2, instrument normatywnie nasycony

Źródło: opracowanie własne na podstawie Abramoff i in. (2018) oraz FDA (2018a, 2018b); klasyfikacja według modelu przedstawionego w podrozdziałach 5.1–5.2.

WI można zatem uznać za spełniony w dużej mierze na poziomie układu społeczno-technicznego. Wniosek ten nie przenosi kompetencji instytucjonalnych na system. Pokazuje natomiast, że wysoka rozliczalność organizacyjna i częściowa interpretowalność mogą zostać osiągnięte bez uznania artefaktu za podmiot odpowiedzialności.

5.4.5. Kontrargument i mini-konkluzja

Można twierdzić, że system przekracza próg narzędzia już dlatego, że generuje wynik o znaczeniu klinicznym bez uprzedniej interpretacji przez specjalistę, klasyfikuje materiał jako spełniający albo niespełniający wymagania jakościowe, działa według określonych progów i pozostawia dokumentację umożliwiającą kontrolę. Taka konfiguracja pozwala wskazać system jako przyczynowo wyodrębniony element procesu oraz postawić pytanie, dlaczego wygenerował dany rezultat.

Argument ten trafnie wskazuje, że IDx-DR/LumineticsCore nie jest narzędziem biernym ani normatywnie obojętnym. Nie wykazuje jednak, że system jest adresatem pytania „dlaczego?” w sensie przyjętym w H1. Odpowiedzi na pytania dotyczące progów, wskazań, jakości obrazów i dalszego postępowania udzielają dokumentacja, producent, regulator i personel medyczny. System pozostaje przedmiotem audytu, lecz publiczne źródła nie pokazują, aby sam przyjmował zarzut, rozpoznawał rację jako rację i korygował rozstrzygnięcie w jej świetle. Możliwość wyjaśnienia działania narzędzia przez ludzi nie jest tożsama ze zdolnością narzędzia do uczestnictwa w praktyce uzasadniania.

Próg H1 nie zostaje więc obniżony do autonomii wykonawczej ani audytowalności zewnętrznej. Znaczenie przypadku polega właśnie na tym, że pozwala odróżnić trzy poziomy: audytowalność systemu jako przedmiotu kontroli, rozliczalność instytucjonalną układu, w którym system działa oraz agentową adresowalność proceduralną wymagającą responsywności na racje.

Mini-konkluzja. Zastosowanie modelu do przypadku IDx-DR/LumineticsCore ilustruje jego moc rozróżniającą. Publiczna dokumentacja potwierdza ograniczoną autonomię diagnostyczną, istotne wskaźniki powtarzalności technicznej, częściową interpretowalność operacyjną oraz rozbudowane osadzenie instytucjonalne. Nie wykazuje jednak responsywności systemu na racje wymaganej przez H1, afektywności funkcjonalnej właściwej H2 ani kontroli drugiego rzędu wymaganej przez H3. WI jest w dużej mierze realizowany przez układ społeczno-techniczny, natomiast R ma profil nierównomierny: dobrze rozwinięte śledzenie odpowiedzialności współwystępuje z ograniczonym reagowaniem na racje w toku klasyfikacji.

W przypadku tym w znacznym zakresie występują dwa pierwsze z wyróżnionych poziomów, czyli audytowalność przedmiotu kontroli i rozliczalność instytucjonalna, lecz publiczne źródła nie wykazują trzeciego, czyli agentowej adresowalności proceduralnej. IDx-DR/LumineticsCore należy zatem sklasyfikować jako typ I w macierzy $A \times F$, to jest $A2/A4 \times F2$: normatywnie nasycony instrument działający w rozliczalnym układzie społeczno-technicznym, a nie quasi-podmiot proceduralny. Wynik ten wspiera progowy i modularny charakter modelu, nie stanowi jednak jego empirycznej walidacji. Ostateczny ciężar uzasadniania, korekty i odpowiedzialności nadal spoczywa na ludziach oraz instytucjach.

Podsumowanie

Rozdział 5 prowadzi do wniosku, że „sztuczna podmiotowość”, o ile ma być pojęciem sensownym, powinna być rozumiana jako status funkcjonalno-instytucjonalny określony przez $F3 + R + I$. Jeżeli konkretne rozwiązanie techniczne nie zapewnia publicznie sprawdzalnych uzasadnień (F3) ani odpowiedniego poziomu znaczącej kontroli człowieka i śledzenia odpowiedzialności (R) oraz minimalnej interpretowalności (I), nie należy mówić o quasi-podmiotowości, lecz jedynie o narzędziach służących zapewnianiu zgodności z obowiązującymi normami prawnymi, technicznymi i instytucjonalnymi oraz ograniczaniu szkód; postulaty „pełnej” agencji (F4) mają wówczas jedynie charakter regulatywny. (1) Agent $\neq$ podmiot. „Agent” jest kategorią

opisową (LoA: interaktywność, autonomia, adaptacyjność), natomiast „podmiot” należy do sfery normatywnej i zależy od praktyk uzasadniania oraz przypisywania odpowiedzialności. W prawie odpowiadają temu ujęcie osobowości jako własności klastrowej oraz rozdział ról i obowiązków, przy braku założeń esencjalistycznych. W konsekwencji przypisanie quasi-podmiotowe ma charakter warunkowy i proceduralny.

(2) Minimum instytucjonalne. Warunkiem sensownych przypisań odpowiedzialności jest konfiguracja: agent jawny + MHC (responsywność na racje i możliwość śledzenia odpowiedzialności) + MLI. Taka konfiguracja zapewnia adresowalność i rozliczalność w trybie publicznym, czyli zdolność do przedstawiania racji, ich kontroli oraz wiązania decyzji z określonymi osobami i instytucjami.

(3) Ocena implementacji. Rozwiązania techniczne należy oceniać pod kątem spełnienia F3 (jawna etyczność) oraz $R + I$ (znacząca kontrola człowieka, śledzenie odpowiedzialności i minimalna interpretowalność), a nie na podstawie samych deklaracji „moralności”. Przyjęcie tych kryteriów pozwala uniknąć zarówno „deflacji agencji”, jak i jej antropomorfizowania. Prowadzi to do wniosku, że zidentyfikowane progi ($F3 + R + I$) oraz macierz $A \times F + (R, I)$ dostarczą w rozdziale 6 narzędzia do oceny scenariuszy rozwoju AI, analizy „luki odpowiedzialności” oraz projektowania stopniowalnych modeli współodpowiedzialności człowiek–AI, zgodnych z ramami regulacyjnymi (AI Act / PLD).

Zastosowanie modelu do IDx-DR/LumineticsCore konkretyzuje te ustalenia. Publiczna dokumentacja potwierdza ograniczoną autonomię diagnostyczną, istotne wskaźniki powtarzalności technicznej, częściową interpretowalność oraz rozbudowane osadzenie instytucjonalne. Nie wykazuje jednak spełnienia H1 ani H2, a w świetle udokumentowanej funkcjonalności H3 pozostaje niespełniona. WI jest w dużej mierze realizowany na poziomie układu społeczno-technicznego. W macierzy $A \times F$ system odpowiada typowi I: $A2/A4 \times F2$, czyli normatywnie nasyconemu instrumentowi działającemu w rozliczalnym układzie społeczno-technicznym. Przypadek pokazuje tym samym, że audytowalność systemu i rozliczalność instytucjonalna nie są tożsame z agentową adresowalnością proceduralną.

W ten sposób wywód przechodzi od klasyfikacji i warunków quasi-podmiotowości do testów instytucjonalnych oraz modeli rozliczalności.

Rozdział 6. Scenariusze rozwoju AI, luka odpowiedzialności i modele współodpowiedzialności

Wprowadzenie

Rozdział 6 poświęcony jest możliwym konfiguracjom systemów sztucznych w przestrzeni $A \times F + (R, I)$, w których poszczególne wskaźniki związane z K1–K3 mogą być rozwinięte w różnym stopniu, lecz ich częściowe występowanie nie oznacza przekroczenia progu H1 ani osiągnięcia poziomu F3. Analizuję w nim trzy możliwe scenariusze rozwoju AI: model emulacyjny, model ewolucyjny oraz model rewolucyjny. Zasadniczo rozdział ma charakter analityczno-teoretyczny, ale również normatywny, w tym sensie, że dąży do odróżnienia odpowiedzialności pozornej od uzasadnionej. Nie podejmuję próby rozstrzygnięcia, który z modeli jest najbardziej prawdopodobny, lecz rekonstruuje ich strukturę w ramach poziomów abstrakcji (LoA) i analizuję, jakie rodzaje luk odpowiedzialności generuje każdy z nich w ramach przyjętego aparatu pojęciowego.

Pytanie przewodnie rozdziału dotyczy tego, w jaki sposób różne ścieżki rozwoju sztucznej podmiotowości wpływają na warunki przypisywania odpowiedzialności omówione w częściach I–V dysertacji. Chodzi w szczególności o identyfikację adresatów, dostępność racji, kontrolę zależną od racji, ciągłość tożsamości oraz zdolność instytucji do działania w czasie. Analiza tych warunków została przygotowana w rozdziale 2 przez rekonstrukcję identyfikowalności adresata, audytowalności racji i kontroli zależnej od racji, a w rozdziale 5 przez macierz $A \times F + (R, I)$. W podrozdziale 6.4 ujmuję je w postaci lokalnej zasady przypisania odpowiedzialności (LSP) oraz zasady superweniencji praktyk przypisywania odpowiedzialności (ZSP). Następnie wprowadzam pięć jakościowych wymiarów diagnostycznych, dotyczących identyfikowalności adresatów, audytowalności racji, kontroli zależnej od racji, trwałości praktyk przypisywania odpowiedzialności oraz presji czasowo-poznawczej.

Teza robocza rozdziału składa się z trzech twierdzeń. Po pierwsze, scenariusze emulacyjny, ewolucyjny i rewolucyjny nie stanowią konkurencyjnych prognoz rozwoju AI, lecz trzy abstrakcyjne modele analityczne, które pokazują, w jaki sposób wzrost poziomu F oraz odmienne konfiguracje $R + I$ prowadzą do jakościowo różnych problemów odpowiedzialności. Po drugie, w klasycznym ujęciu zaproponowanym przez Andreasa Matthiasa luka odpowiedzialności oznacza sytuację, w której nie można wskazać podmiotu spełniającego warunki odpowiedzialności za działanie systemu

uczącego się. W niniejszym rozdziale zachowuję tę intuicję jako punkt wyjścia, lecz pojęcie luki odpowiedzialności rozumiem szerzej jako zbiór zróżnicowanych struktur, w których zawodzą warunki identyfikowalności adresatów roszczeń, rekonstrukcji racji, kontroli zależnej od racji oraz stabilności praktyk przypisywania odpowiedzialności (LSP/ZSP). W modelu emulacyjnym dominują problemy związane ze statusem podmiotu i możliwością jego modyfikacji; w modelu ewolucyjnym — problemy akumulacji i narastającej złożoności; natomiast w modelu rewolucyjnym — problemy o charakterze czasowym i poznawczym. W podrozdziale 6.4 porządkuję te wymiary metrycznie. Po trzecie, odpowiedzi na wskazane trudności nie polegają ani na prostym rozszerzaniu zakresu osobowości prawnej na systemy AI, ani na przypisywaniu odpowiedzialności pojedynczym adresatom, lecz wymagają projektowania modeli współodpowiedzialności oraz ładu regulacyjnego, które w sposób systematyczny korygują profil metryk M1–M5 i utrzymują go w obszarze odpowiedzialnej agencji, ograniczając możliwości trwałego zniekształcenia praktyk przypisywania odpowiedzialności.

Struktura rozdziału odzwierciedla strategię analityczną. Najpierw przedstawiam trzy koncepcyjne modele rozwoju AI, a następnie wprowadzam aparat metryczny stopniowania luki odpowiedzialności, który porządkuje dalszą analizę. Na tym tle rekonstruuję zagrożenia technologiczne, społeczne i regulacyjne oraz rozwijam modele współodpowiedzialności i ładu regulacyjnego jako normatywną odpowiedź na zidentyfikowane deficyty rozliczalności.

6.1. Emulacja mózgu

Przedmiotem niniejszego podrozdziału jest koncepcja emulacji całego mózgu (ang. *whole brain emulation*, WBE), rozumiana jako technika odwzorowania organizacji i dynamiki mózgu, w wyniku której powstaje system cyfrowy przejawiający funkcje psychologiczne zbliżone do ludzkich. Koncepcję tę analizuję w świetle przyjętej wcześniej ramy minimalnej adresowalności proceduralnej, tak aby od początku była ona osadzona w szerszym aparacie pojęciowym niniejszej dysertacji. Przedmiotem analizy nie jest techniczna wykonalność WBE, lecz sposób, w jaki omawiany scenariusz kształtuje pojęcie sztucznej podmiotowości oraz wpływa na strukturę odpowiedzialności rekonstruowaną przy użyciu poziomów abstrakcji (LoA), kryteriów K1–K4 oraz macierzy $A \times F + (R, I)$. Pytanie brzmi więc, czy emulacja mózgu prowadzi do wyłonienia nowego typu podmiotowości i odrębnej „luki odpowiedzialności”, czy zaostrza znane już problemy, takie jak ciągłość osobowa, powielanie tożsamości oraz rozproszenie

odpowiedzialności w złożonych układach socjotechnicznych. Teza robocza podrozdziału jest następująca: emulacja mózgu nie wytwarza ontologicznie odmiennego rodzaju podmiotu, lecz generuje zbiór quasi-podmiotów, które na poziomie abstrakcji psychologicznej stanowią funkcjonalne kontynuacje osoby źródłowej. Przy założeniu wykazania H1 emulowane umysły można warunkowo włączyć do praktyk słabej odpowiedzialności proceduralnej. Przypisanie im roli quasi-podmiotowej wymaga ponadto niezależnego wykazania F3 oraz odpowiednich parametrów $R + I$ i warunków WI. Samo występowanie wskaźników K1–K3 nie wystarcza do takiej kwalifikacji. Problem nie dotyczy więc samej podmiotowości, lecz sposobu rozdzielania odpowiedzialności między liczne, rozgałęziające się ciągi tożsamości oraz między emulowanymi umysłami a instytucjami sprawującymi nad nimi kontrolę. W tym sensie koncepcja WBE pełni funkcję testu granicznego dla przyjętego progu minimalnej adresowalności proceduralnej oraz powiązanych z nim struktur odpowiedzialności. Jest ona traktowana jako narzędzie weryfikacji wcześniejszych ustaleń analitycznych, a nie jako autonomiczny projekt metafizyczny.

Sandberg i Bostrom definiują emulację całego mózgu (WBE) jako modelowanie funkcji mózgu w taki sposób, aby model wynikowy funkcjonował w taki sam sposób jak system pierwotny.

The basic idea is to take a particular brain, scan its structure in detail, and construct a software model of it that is so faithful to the original that, when run on appropriate hardware, it will behave in essentially the same way as the original brain. The term emulation originates in computer science, where it denotes mimicking the function of a program or computer hardware by having its low-level functions simulated by another program. While a simulation mimics the outward results, an emulation mimics the internal causal dynamics (at some suitable level of description). The emulation is regarded as successful if the emulated system produces the same outward behaviour and results as the original (Sandberg & Bostrom, 2008, s. 7).

Procedura ta obejmuje skanowanie mózgu z odpowiednią rozdzielczością, rekonstrukcję jego sieci neuronalnej oraz uruchomienie uzyskanej struktury w postaci symulacji na platformie cyfrowej (Sandberg & Bostrom, 2008, s. 7–8, 16; Sandberg, 2014, s. 439–441). Na poziomie abstrakcji neurobiologicznej (LoA) jest to koncepcja redukcjonistyczna. Zakłada bowiem, że własności istotne psychologicznie superwenują nad strukturą i dynamiką mózgu, a tym samym mogą zostać odtworzone przez model o odpowiedniej szczegółowości. Na poziomie abstrakcji psychologicznej WBE ma zapewniać ciągłość przekonań, pragnień, pamięci i nawyków osoby, tak aby emulowany

umysł mógł realizować jej cele oraz funkcjonować w tych samych relacjach normatywnych co podmiot pierwotny. W ramach aparatu pojęciowego przyjętego w niniejszej pracy oznacza to przejście od nośnika biologicznego do cyfrowego przy zachowaniu kryteriów K1–K3, a więc zdolności reagowania na racje, minimalnej refleksyjnej kontroli nad działaniem oraz możliwości korygowania własnych ocen normatywnych.

Tak rozumiana emulacja mózgu stanowi konsekwencję stanowiska funkcjonalistycznego przyjętego wcześniej. Umysł ujmowany jest tu jako układ zależności przyczynowych oraz procesów przetwarzania informacji, który może być realizowany w różnych typach nośników, bez przypisywania któremukolwiek z nich uprzywilejowanego statusu. Status podmiotowy zależy zatem od sposobu organizacji tych zależności w danym systemie, a nie od substancji, z której jest zbudowany. Transhumanizm traktuje emulację całego mózgu (WBE) jako naturalny kierunek rozwoju neuroinformatyki i utrzymuje, że jej ewentualna realizacja mogłaby umożliwić istnienie życia w formie cyfrowej, a także zapoczątkować rozwój superinteligencji, między innymi poprzez zwiększanie mocy obliczeniowej i przyspieszanie procesów poznawczych (Bostrom, 2014, s. 22–24; Chalmers, 2010, s. 9). Krytycy tej koncepcji podnoszą, że WBE splata twierdzenia naukowe z projektem o ambicjach egzystencjalnych, a argumentacja na jej rzecz często zakłada nadmierny optymizm co do możliwości neuroobrazowania oraz zbyt daleko idące utożsamienie mózgu z umysłem (Dennett, 2016, rozdz. 3–4). W niniejszej dysertacji kluczowe znaczenie ma jednak nie rozstrzygnięcie tych sporów, lecz rekonstrukcja założeń dotyczących podmiotowości i odpowiedzialności, jakie koncepcja ta implikuje w kontekście debaty o zaawansowanej AI.

W literaturze, do której odwołuję się w niniejszej pracy, wyróżnia się co najmniej trzy główne stanowiska dotyczące statusu emulowanych umysłów. Pierwsze z nich, określane mianem esencjalistycznego, utrzymuje, że nawet idealnie wierna kopia struktury i dynamiki mózgu pozostaje jedynie symulacją, pozbawioną odpowiedniego substratu, świadomości fenomenalnej lub „wewnętrznej natury” uznawanej za konieczną dla podmiotowości. W tym ujęciu procedura WBE miałaby prowadzić do unicestwienia pierwotnego podmiotu oraz do powstania nowej istoty, która jedynie uważa się za jego kontynuację, nie będąc nią w sensie istotnym. Drugą grupę stanowisk tworzą ujęcia funkcjonalistyczne, które akcentują ciągłość wzorców psychologicznych oraz relacji normatywnych. Zgodnie z tym podejściem, jeżeli emulowany umysł nadal realizuje cele,

rozpoznaje zobowiązania i reaguje na racje w sposób analogiczny do stanu sprzed emulacji, wówczas w ramach praktyk uzasadniania i rozliczania może być on traktowany jako „ta sama” osoba w sensie relewantnym dla odpowiedzialności (Chalmers, 2010, s. 9; por. Shevlin, 2021, s. 459–463, 468–471).²⁶ Trzecie stanowisko reprezentuje neoarystotelesowski naturalizm Richarda Friedricha Runego. Nie jest to jednak obrona pośredniego statusu podmiotowego emulacji: w przyjętej przez niego interpretacji analizowane emulacje pozostają artefaktami, a nie moralnymi następcami ludzi. Wymóg szczególnej ostrożności wobec takich artefaktów nie oznacza przyznania im autonomicznego statusu moralnego (Runge, 2025, s. 4219–4220). Autor dopuszcza jednak odmienną kwalifikację przyszłych emulacji, gdyby w samopodtrzymujących się i samoreprodukujących robotach wykształciły niezależną formę życia (Runge, 2025, s. 4221).

W przyjętej macierzy $A \times F + (R, I)$ emulacja całego mózgu (WBE) stanowi przypadek graniczny o potencjalnie wysokiej pozycji na osi ontologicznej i funkcjonalnej. Po pierwsze, scenariusz zakłada rozwinięcie wskaźników K1–K3 na poziomie zbliżonym do osoby ludzkiej, co może wspierać wykazanie H1, lecz nie przesądza o osiągnięciu F3. Po drugie, na poziomie abstrakcji psychologicznej przyjmuje się ciągłość funkcjonowania osoby, co pozwala rozważać emulowany umysł jako minimalnie adresowalny proceduralnie, jeżeli H1 zostanie odrębnie wykazana. Poziom F3 można przypisać dopiero wtedy, gdy system jawnie posługuje się zasadami i przedstawia własne, audytowalne uzasadnienia normatywne; ciągłość psychologiczna i wskaźniki K1–K3 nie wystarczają do tego wniosku. Przypisanie roli quasi-podmiotowej wymaga ponadto odpowiednich $R + I$ w ramach WI. Jednocześnie jest to koncepcja, która komplikuje problem przypisywania odpowiedzialności, ponieważ łączy poziom postulowanej podmiotowości ze zmianą nośnika oraz z możliwością powielania i modyfikowania tego samego umysłu. Konsekwencją jest napięcie dotyczące ciągłości

²⁶ W tym miejscu konieczne jest doprecyzowanie znaczenia zwrotu „ta sama osoba”. Z perspektywy redukcjonizmu psychologicznego Dereka Parfita, omawianego w rozdziale II (2.4), nie istnieje żaden „dodatkowy fakt” tożsamości osobowej poza relacjami ciągłości i łączności psychologicznej, opartymi na pamięci, intencjach oraz dyspozycjach deliberacyjnych (Parfit, 1984, s. 215–216). W konsekwencji teza funkcjonalistyczna nie stanowi roszczenia do zachowania tożsamości numerycznej, lecz odnosi się do zachowania tego, co pozostaje istotne w praktykach przypisywania odpowiedzialności, a mianowicie ciągłości rozrachunku oraz zdolności do występowania w roli adresata roszczeń (por. Parfit, 1984, s. 222–223). Parfit podkreśla zarazem, że przypadki rozgałęzienia ujawniają ograniczenia języka „tej samej osoby”: relacja ciągłości psychologicznej może mieć więcej niż jednego kontynuatora, podczas gdy tożsamość numeryczna nie dopuszcza wielokrotności (Parfit, 1984, s. 253–255). W dalszej analizie zwrot „ta sama osoba” traktuję zatem jako skrót pojęciowy oznaczający „kontynuację w sensie ciągłości rozrachunku”, a nie jako rozstrzygnięcie metafizycznego sporu o identyczność.

osoby oraz tożsamości osobowej. We wcześniejszych rozdziałach przyjąłem robocze założenie, że minimalny warunek odpowiedzialności praktycznej obejmuje ciągłość psychologiczną, rozumianą jako sieć powiązań pamięci, intencji i procesów deliberacyjnych, umożliwiającą identyfikację podmiotu odpowiedzialnego na poziomie abstrakcji psychologicznej. W scenariuszu WBE ciągłość ta ma być zabezpieczona poprzez odwzorowanie struktury i dynamiki mózgu, jednak pojawiają się tu nowe trudności. Po pierwsze, możliwość wielokrotnego odtworzenia tej samej konfiguracji; po drugie, stopniowe różnicowanie się kopii w czasie; po trzecie, możliwość modyfikowalności ich struktury. W rezultacie powstaje pytanie, czy i w jakim zakresie można mówić o tej samej osobie, gdy jej dalsze istnienie rozgałęzia się na kilka systemów dzielących wspólną przeszłość, lecz realizujących odmienne trajektorie rozwoju w przyszłości.

Opisane trudności odpowiadają klasycznemu, parfitowskiemu problemowi rozgałęzienia (ang. *fission problem*). Jeżeli ciągłość psychologiczna zostaje zachowana, lecz występuje u więcej niż jednego kontynuatora, pytanie o to, czy mamy do czynienia z tym samym podmiotem, traci na jednoznaczności. Parfit wskazuje, że w takich przypadkach nie istnieje żaden „dodatkowy fakt”, który czyniłby jednego z kontynuatorów numerycznie identycznym z osobą wyjściową A, mimo że każdy z nich może pozostawać w relacji ciągłości i łączności psychologicznej, istotnej z punktu widzenia praktycznego (Parfit, 1984, cz. III, s. 253–255). Z perspektywy odpowiedzialności praktycznej oznacza to, że sama metafizyka tożsamości nie dostarcza reguł podziału zobowiązań, winy i roszczeń, które w dotychczasowych praktykach opierały się na założeniu jednego, niepowielalnego nośnika tożsamości.

W odniesieniu do WBE znaczenie ma także tzw. wariant „teleportacyjny”, a więc pytanie, czy biologiczny pierwowzór A ulega zniszczeniu. Jeżeli nie ulega, od samego początku współistnieją co najmniej dwa podmioty o wspólnej przeszłości psychologicznej, co prowadzi do konfliktu roszczeń dotyczących jednej biografii oraz jednego zespołu zobowiązań. Jeżeli natomiast pierwowzór ulega zniszczeniu, konflikt ten nie znika, ponieważ wielość kontynuatorów nadal uniemożliwia wskazanie jednego adresata zobowiązań A w ramach utrwalonych praktyk prawnych. Problem ten nie jest problemem K4 w ścisłym znaczeniu, lecz problemem ciągłości identyfikacyjnej i rozdziału zobowiązań między kontynuatorów. K4 należy odrębnie odnosić do diachronicznej stabilności profilu działania każdej konkretnej wersji emulowanego systemu.

W analizie ciągłości identyfikacyjnej ujawnia się to, co można określić mianem problemu powielania tożsamości. Załóżmy, że osoba A, przed rozpoczęciem procesu emulacji swojego mózgu, wyraża wolę dalszego istnienia w formie cyfrowej, a sam proces zostaje przeprowadzony poprawnie, zgodnie z jej wolą oraz obowiązującymi standardami. Jeżeli w jego wyniku powstaje kilka emulowanych umysłów A1, A2 i A3, które początkowo są identyczne, a następnie rozwijają się w różnych kierunkach, każdy z nich zachowuje ciągłość i łączność psychologiczną z A. W tym sensie każdy ma podobne podstawy do uznania siebie za kontynuację tej samej historii psychologicznej, choć nie oznacza to zachowania tożsamości numerycznej (Parfit, 1984, cz. III, s. 253–255). Z perspektywy ciągłości identyfikacyjnej oraz praktyk przypisywania odpowiedzialności rodzi to następujące pytania. Który z tych podmiotów ma realizować zobowiązania zaciągnięte przez A, który dziedziczy jej majątek, lub kto ponosi odpowiedzialność za jej wcześniejsze czyny? Dotychczasowe praktyki przypisywania odpowiedzialności opierają się na założeniu istnienia jednego, jednoznacznie identyfikowalnego nośnika tożsamości. Emulacja umysłu narusza to założenie i prowadzi do sytuacji, w której jedna biografia rozszczepia się na kilka równoległych dróg. Problem powielania tożsamości polega na tym, że struktury odpowiedzialności zostały zaprojektowane z myślą o pojedynczych, niepowielalnych podmiotach, podczas gdy WBE wprowadza możliwość wielokrotnej kontynuacji tej samej historii psychologicznej.

Drugim problemem jest rozdzielanie się tożsamości. Z upływem czasu emulowane umysły rozwijają się jednak w odmiennych kierunkach. Podejmują różne decyzje, wchodzą w różne relacje i realizują różne działania. Wszystkie zachowują pamięć osoby A oraz początkową strukturę jej przekonań, jednak ich dalsze postępowanie ulega coraz większemu zróżnicowaniu. W konsekwencji pojawia się pytanie, w jaki sposób przypisywać odpowiedzialność za decyzje podejmowane po rozdzieleniu tożsamości? Czy osoba A ponosi odpowiedzialność za wszystkie przyszłe działania swoich emulowanych umysłów, czy jedynie za te, które pozostają w związku z jej wcześniejszymi intencjami i zobowiązaniami? Równie problematyczne jest pytanie, czy jedna wersja, na przykład A1, może ponosić odpowiedzialność za decyzje innej wersji, np. A2, skoro obie wywodzą się z tego samego umysłu? Problem rozdzielania tożsamości nie dotyczy zatem samego aktu powielenia umysłu, lecz późniejszego różnicowania się jego dalszych losów.

Trzecim istotnym wyzwaniem jest kwestia modyfikowania emulowanych umysłów. W przeciwieństwie do ludzkiego mózgu system emulowany może podlegać

celowym ingerencjom technicznym, takim jak usuwanie określonych wspomnień, wzmacnianie lub osłabianie wybranych dyspozycji, a także wprowadzanie nowych modułów decyzyjnych lub filtrów normatywnych. Jeżeli modyfikacje te są wprowadzane za zgodą osoby, której umysł został poddany emulacji, można je interpretować jako szczególną formę rozszerzenia autonomii, realizowaną w nowym medium. Inaczej przedstawia się jednak sytuacja, w której ingerencje te mają charakter arbitralny, a ich źródłem jest instytucja sprawująca nad procesem emulacji kontrolę. Wówczas pojawia się pytanie, czy, a jeżeli tak, to w jakim zakresie, emulowany umysł pozostaje tym samym podmiotem oraz czy nadal może być traktowany jako kontynuacja osoby A. Problem modyfikacji dotyczy zatem nie tylko ciągłości psychologicznej podmiotu, lecz także dopuszczalnego zakresu zewnętrznej ingerencji w jego strukturę oraz sposobu podziału odpowiedzialności między emulowany umysł a instytucje, które go projektują, utrzymują i modyfikują.

Trzy omówione typy problemów pokazują, że w koncepcji WBE szczególnie wrażliwe stają się identyfikowalność adresata oraz trwałość praktyk przypisywania odpowiedzialności. Dotychczasowe praktyki, oparte na założeniu istnienia jednego stabilnego nośnika tożsamości, okazują się niewystarczające w warunkach powielania, rozdzielania i modyfikowania emulowanych umysłów. Pytanie o odpowiedzialny podmiot traci wówczas jednoznaczną odpowiedź, ponieważ każda wersja może być zarazem kontynuatorem wspólnej biografii i odrębnym systemem z własną trajektorią działania. W rezultacie wyłania się złożony układ współodpowiedzialności, w którym trzeba odrębnie ustalić, kto jest kontynuatorem zobowiązań i jak instytucje rozdzielają odpowiedzialność. W aparacie podrozdziału 6.4 trudności te dotyczą przede wszystkim M1, M4 i WI. K4 zachowuje węższą funkcję: służy ocenie, czy profil działania każdej konkretnej emulacji pozostaje diachronicznie stabilny i wiarygodny.

Pojawia się tu również inny wątek, istotny dla oceny etycznej WBE i powiązany z problematyką tożsamości, mianowicie podatność emulowanych umysłów na krzywdę. Sandberg i Bostrom wskazują, że takie umysły, rozumiane jako funkcjonalne odpowiedniki osób, mogą doświadczać stanów analogicznych do ludzkiego cierpienia, lęku czy upokorzenia, o ile przyjmuje się funkcjonalną teorię stanów mentalnych (Sandberg & Bostrom, 2008, s. 12–14, 36–38). W niniejszej dysertacji kluczowe znaczenie ma to, że w praktykach uzasadniania i rozliczania mogą być one traktowane jako podmioty zdolne do doznawania krzywd. W tym sensie WBE wprowadza nową klasę podmiotów narażonych na ryzyko wykorzystywania, uprzedmiotowienia oraz

instrumentalnego traktowania, zwłaszcza w sytuacjach, w których emulowane umysły pozostają pod kontrolą instytucji publicznych lub korporacji. Problem podatności na krzywdę splata się tu bezpośrednio z kwestiami ciągłości i powielania tożsamości. Im większa liczba emulowanych umysłów oraz im większa asymetria między nimi a podmiotami sprawującymi nad nimi kontrolę, tym wyższe ryzyko naruszania ich interesów oraz tym większa trudność w przypisywaniu im odpowiedzialności.

Stanowisko Rungego stanowi kontrast wobec funkcjonalistycznej interpretacji emulacji całego mózgu. Autor uzasadnia ostrożność w postępowaniu z emulacjami przez wymagania stawiane ludzkim sprawcom, nie przez uznanie autonomicznych interesów samych emulacji (Runge, 2025, s. 4219–4220). W niniejszej dysertacji odrębnie rozważam trzy kwestie: podstawy ewentualnej ochrony, warunki proceduralnej adresowalności oraz ciągłość zobowiązań osoby źródłowej. Przyjęcie H1 jako warunku uczestnictwa w praktykach uzasadniania pozostaje propozycją własną i nie wynika ze stanowiska Rungego. Również przypisanie pacjencji wymaga odrębnych racji, zamiast wyprowadzania go z samej możliwości emulowania funkcji mózgu. Rozróżnienia te pozwalają analizować instytucjonalne zależności emulacji bez utożsamiania ich ewentualnej ochrony moralnej, proceduralnej adresowalności i ciągłości osobowej.

W tym kontekście należy doprecyzować, na czym polega „luka odpowiedzialności” w sensie węższym. Santoni de Sio i Mecacci pokazują, że w klasycznych ujęciach problem ten dotyczy sytuacji, w których złożoność systemu sztucznego uniemożliwia wskazanie jednego określonego podmiotu odpowiedzialnego za wyrządzoną szkodę (Santoni de Sio & Mecacci, 2021, s. 1057–1058). W przypadku WBE mechanizm ten przyjmuje jednak odmienną postać. Dominującym problemem nie jest zwykła nieobecność podmiotu, lecz brak jednoznacznego adresata odpowiedzialności w sensie tradycyjnych kryteriów tożsamości, przy jednoczesnym pojawieniu się wielu potencjalnych kandydatów do tej roli. Należą do nich emulowane umysły, instytucje sprawujące nadzór nad systemem, projektanci oraz podmioty odpowiedzialne za tworzenie ram prawnych i regulacyjnych. W takim ujęciu luka odpowiedzialności nie polega na absolutnym braku odpowiedzialnego podmiotu, lecz na przeciążeniu istniejących zasad przypisywania odpowiedzialności, które nie zostały zaprojektowane z myślą o sytuacjach rozdzielania i modyfikowania tożsamości ani o wielopoziomowej kontroli nad systemem. W konsekwencji wzrasta ryzyko przerzucania odpowiedzialności zarówno między emulowanymi umysłami a podmiotami sprawującymi nad nimi kontrolę, jak i między instytucjami projektującymi, wdrażającymi i korzystającymi z

systemów WBE. Sprzyja to zjawisku „wybielania odpowiedzialności” (ang. *responsibility laundering*) oraz prowadzi do „rozmycia odpowiedzialności” (ang. *responsibility diffusion*).

W ramach przyjętego aparatu pojęciowego można wyróżnić trójstopniową strukturę przypisywania odpowiedzialności w koncepcji WBE. Na poziomie technicznego systemu należy odrębnie ocenić H1 i F3: K1–K3 mogą dostarczać wskaźników dla H1, lecz status quasi-podmiotowy wymaga ponadto jawnego uzasadniania właściwego F3. Jeżeli te progi zostały wykazane, emulowany umysł może zostać warunkowo włączony do odpowiednich praktyk odpowiedzialności, bez przesądzania o pełnej podmiotowości moralnej albo identyczności numerycznej. Na poziomie podmiotowości prawnej odpowiedzialność spoczywa na instytucjach nadzorujących proces emulacji umysłów, w szczególności na korporacjach, instytucjach państwowych oraz jednostkach badawczych. Odpowiadają one za warunki funkcjonowania emulowanych umysłów, w tym za ochronę przed nadużyciami, zapewnienie dostępu do niezbędnych zasobów oraz stworzenie ram umożliwiających, w granicach technicznych, samostanowienie. Na trzecim poziomie, określanym tu jako poziom metaodpowiedzialności, odpowiedzialność ponoszą podmioty kształtujące ramy regulacyjne. To one rozstrzygają, czy przyjęte przepisy skutecznie zapobiegają „wybielaniu odpowiedzialności” oraz czy odzwierciedlają relacje między emulowanymi umysłami, ludzkimi sprawcami i instytucjami. Trzy poziomy nie muszą się pokrywać. Można bronić stanowiska, zgodnie z którym emulowany umysł stanowi funkcjonalną kontynuację osoby w praktykach odpowiedzialności, jeżeli poza ciągłością psychologiczną wykazano odpowiedni próg systemowy i zapewniono R, I oraz WI. Nie przesądza to pełnej osobowości prawnej, identyczności numerycznej ani bezwarunkowej podmiotowości moralnej i pozwala zarazem unikać „wybielania odpowiedzialności” (Bryson, Diamantis & Grant, 2017, s. 273–291).

W konsekwencji koncepcja WBE nie podważa samych kategorii podmiotowości i odpowiedzialności, lecz istotnie modyfikuje sposób ich rozumienia. Po pierwsze, sugeruje ona, że funkcjonalna kontynuacja osoby jest istotnym, lecz niewystarczającym warunkiem jej warunkowego włączenia do praktyk odpowiedzialności; konieczne jest ponadto wykazanie zdolności systemu odpowiednich do przypisywanej roli oraz zapewnienie R + I i WI. Po drugie, pokazuje, że zasadniczym problemem nie jest brak odpowiedzialnego podmiotu, lecz nadmiar nakładających się zakresów odpowiedzialności, wynikający zarówno z powielania i modyfikowania emulowanych

umysłów, jak i z asymetrii władzy między nimi a instytucjami sprawującymi nad nimi kontrolę. Po trzecie, analiza ta wskazuje, że projektowanie zasad znaczącej kontroli człowieka (MHC) oraz struktur $R + I$ w kontekście WBE wymaga szczególnej ostrożności. Zanim emulowanym umysłom zostaną przypisane jakiekolwiek role quasi-podmiotowe, konieczne jest precyzyjne określenie zasad przypisywania odpowiedzialności oraz jednoznaczne wskazanie instytucji odpowiedzialnych za ochronę tych systemów przed nadużyciami i instrumentalnym wykorzystaniem. W tym sensie koncepcja WBE ujawnia zarówno mocne strony, jak i ograniczenia przyjętego w niniejszej dysertacji performatywno-instytucjonalnego ujęcia podmiotowości, a zarazem wymusza doprecyzowanie zakresu jego zastosowania w warunkach radykalnej plastyczności tożsamości oraz wielopoziomowej kontroli technicznej.

6.2. Model ewolucyjny

Model ewolucyjny pełni w tym rozdziale funkcję abstrakcyjnego ujęcia analitycznego rozwoju zaawansowanej sztucznej inteligencji, w którym zdolności systemów nie ujawniają się skokowo, lecz narastają stopniowo w toku kolejnych zmian technologicznych i instytucjonalnych. Stopniowość tę rozumiem relacyjnie: jako taki układ tempa przemian po stronie architektur systemów, infrastruktury obliczeniowej, praktyk gromadzenia danych oraz sposobów wdrażania względem tempa przekształceń praktyk przypisywania odpowiedzialności, w którym nie sposób wskazać jednego jednoznacznego momentu przełomu. Zestawienie tego ujęcia z modelem emulacyjnym (emulacja mózgu) oraz modelem rewolucyjnym (tzw. eksplozja inteligencji) ma charakter czysto typologiczny i służy wyodrębnieniu odmiennych mechanizmów powstawania ryzyk oraz luki odpowiedzialności, a nie wykluczeniu możliwości, że rzeczywiste trajektorie rozwoju przyjmują postać mieszaną. Z tego względu model ewolucyjny nie wymaga ani założenia wiernego odwzorowania umysłów biologicznych, ani przyjęcia jednorazowego przełomu rozwojowego, lecz akcentuje stopniowy wzrost zdolności funkcjonalnych systemów w ramach zróżnicowanych układów socjotechnicznych, w których sprawczość i kontrola pozostają rozproszone.

Przedmiotem niniejszego podrozdziału jest analiza wpływu stopniowego rozwoju AI, rozumianego jako długotrwałe przejście od narzędzi do systemów pełniących coraz bardziej samodzielne funkcje w złożonych sieciach społeczno-technicznych, na warunki minimalnej adresowalności proceduralnej oraz strukturę luki odpowiedzialności w ramach aparatu LoA , $K1-K4$ i macierzy $A \times F + (R, I)$. Główne pytanie dotyczy tego, czy

ewolucyjny charakter rozwoju AI sprzyja ograniczaniu luki, zapewniając instytucjom czas na adaptację, czy też prowadzi do stopniowego osłabienia granic sprawczości i rozproszenia odpowiedzialności. Decydujące znaczenie ma tempo zmian i nierównomierność rozwoju zdolności systemów oraz warunków instytucjonalnych. Teza robocza głosi, że model ewolucyjny nie prowadzi do wyłonienia nowego typu podmiotu w jednym momencie, lecz stopniowo wzmacnia zależności między ludźmi, organizacjami i systemami AI. Poszczególne systemy mogą nierównomiernie rozwijać wskaźniki K1–K3, natomiast role, kontrola i warunki WI, R i I pozostają rozproszone w całej sieci. Luka odpowiedzialności ma zatem charakter strukturalny i kumulatywny: wynika z niedopasowania zdolności systemów do instytucjonalnego rozdziału kontroli i odpowiedzialności, a nie z przeniesienia K1–K3 na konfigurację jako całość.

W najprostszym ujęciu model ewolucyjny opisuje rozwój, w którym zdolności systemów AI narastają stopniowo i nierównomiernie w różnych obszarach ich działania. Systemy te od samego początku są osadzone w istniejących sieciach socjotechnicznych, natomiast instytucje regulacyjne reagują na ich rozwój głównie po fakcie. Nie tworzą one ram prawnych z myślą o hipotetycznej „ogólnej sztucznej inteligencji” (ang. *Artificial General Intelligence*, AGI), lecz sukcesywnie adaptują obowiązujące przepisy do kolejnych generacji tej technologii. Raghu Raman proponuje ujęcie trajektorii rozwoju AGI jako agregatu równolegle rozwijających się ścieżek, obejmujących wymiar technologiczny, społeczny oraz kognitywno-etyczny inspirowany funkcjonowaniem ludzkiego mózgu (Raman i in., 2025, s. 1). Takie podejścia sprzyjają interpretacji rozwoju AGI jako procesu rozciągniętego w czasie, opartego na stopniowym kształtowaniu ogólnych zdolności adaptacyjnych, a nie jako akumulacji kolejnych, niezależnych osiągnięć w określonych obszarach. W tym sensie inteligencja, zarówno naturalna, jak i sztuczna, rozumiana jest jako zbiór zdolności o zróżnicowanym zakresie i stopniu ogólności, które mogą rozwijać się stopniowo i niesymetrycznie (Hernández-Orallo, 2017, s. 260). Na tym tle Meredith Ringel Morris proponuje ramę „poziomów AGI”, porządkującą systemy sztucznej inteligencji według rosnącego stopnia ogólności i autonomii w trzech wymiarach: wydajności, ogólności oraz autonomii. Rama ta dostarcza spójnego języka opisującego stopniowy rozwój systemów AI w kierunku większej ogólności, przesuując punkt ciężkości dyskusji z pytania „czy AGI już istnieje?” na pytanie o zakres, tempo i charakter zachodzących zmian (Morris i in., 2023, s. 1). Wspólnym mianownikiem tych ujęć jest przekonanie, że model ewolucyjny opisuje dynamiczny proces rozwojowy, a nie jeden określony moment, w którym pojawia się

AGI. Odejście od pytania „czy?” na rzecz pytania „w jaki sposób i jak szybko?” oznacza przesunięcie perspektywy z rozważań nad granicami ontologicznymi w kierunku analizy procesów transformacji.

Odnosząc powyższe ujęcia do aparatu pojęciowo-analitycznego niniejszej dysertacji, można stwierdzić, że model ewolucyjny opisuje stopniową zmianę konfiguracji w macierzy $A \times F$. W fazie początkowej dominują systemy typu A1–F1, czyli artefakty o wąskim przeznaczeniu, które wywierają istotny wpływ na rzeczywistość, na przykład poprzez klasyfikację obrazów, rekomendowanie treści lub filtrowanie dokumentów, lecz nie posiadają jawnych reprezentacji celów ani zdolności uzasadniania swoich działań. Ich funkcjonowanie pozostaje podporządkowane planom i racjom użytkowników, co uzasadnia traktowanie ich jako produktywnego i efektywnego narzędzia. Wraz z rozwojem AI w kierunku coraz bardziej złożonych systemów pojawiają się konfiguracje zbliżone do poziomu F2, określane jako agenci implicytni. Systemy te uczą się wzorców na podstawie danych historycznych, przejmują część procedur decyzyjnych i modyfikują swoje działanie w czasie. Przy tym ich mechanizmy podejmowania decyzji stają się w coraz większym stopniu nieprzejrzyste nawet dla samych twórców. Na tym etapie decyzje systemów zaczynają realnie wpływać na decyzje podejmowane przez ludzi. Mimo to systemy te nadal traktowane są jako narzędzia, których działanie można, w ograniczonym zakresie, włączyć do istniejących praktyk przypisywania odpowiedzialności. Wprowadzenie mechanizmów wyjaśnialności oraz obowiązku przedstawiania uzasadnień podejmowanych przez systemy AI decyzji prowadzi do pojawienia się cech charakterystycznych dla poziomu F3. Systemy zaczynają wówczas pełnić rolę, w której prezentują racje swoich decyzji, a ludzie traktują te racje jako element zbioru powodów, które mogą zostać zaakceptowane, zakwestionowane lub poddane negocjacji. W tym sensie macierz $A \times F$ umożliwia opis przejścia od sprawstwa narzędziowego do ról quasi-podmiotowych. Role te są funkcjonalne i kontekstowe, nie są zaś wyrazem podmiotowości w sensie silnym.

Z perspektywy kryteriów K1–K3 ewolucja ta polega na stopniowym przenoszeniu części zdolności, które dotąd przypisywano podmiotom ludzkim, na systemy sztuczne. Systemy poziomu F2 przejawiają formę proto-K2, rozumianą jako ograniczoną funkcjonalną wrażliwość na racje: ich decyzje zmieniają się w sposób względnie stabilny pod wpływem tych informacji i korekt, które w danej praktyce są traktowane jako moralnie istotne. Proto-K2 nie oznacza posiadania racji w sensie podmiotowym, lecz to, że zachowanie systemu może być oceniane jako lepiej lub gorzej śledzące racje (ang.

tracking) oraz podlegać korekcie w tym właśnie horyzoncie. Ich działanie da się opisać jako uwzględniające racje, choć systemy te nie reprezentują ich w sposób jawny ani nie potrafią ich bezpośrednio uzasadnić. Wraz z wprowadzaniem mechanizmów uczenia się z informacją zwrotną (w tym uczenia ze wzmocnieniem, ang. *reinforcement learning*) oraz elementów nadzoru ludzkiego pojawiają się załączki proto-K3. Systemy zaczynają przewidywać skutki swoich decyzji w ramach przyjętego modelu i na tej podstawie modyfikują swoje działanie w taki sposób, aby unikać rezultatów ocenianych jako niepożądane. Proces ten można opisać jako minimalną, funkcjonalną formę refleksyjności predykcyjnej. Jednocześnie K1 oznacza tu autokorektę norm i audytowalną korektę decyzji, a nie silną autonomię normatywną, dlatego przez długi czas pozostaje w modelu ewolucyjnym ograniczona. Systemy nie formułują samodzielnie celów ogólnych, lecz działają w granicach celów wyznaczonych z zewnątrz. Wzrost obejmujący K1 może zbliżać system do roli quasi-podmiotowej tylko wtedy, gdy towarzyszy mu jawne i audytowalne uzasadnianie właściwe F3, a jego zastosowanie spełnia odpowiednie parametry R i I oraz szersze warunki WI; kontrola drugiego rzędu wyznacza dopiero dodatkowy próg H3. Ewolucja w kierunku quasi-podmiotowości polega zatem na stopniowym rozwoju zdolności poszczególnych systemów przy równoległej zmianie rozdziału decyzji, kontroli i odpowiedzialności między ludźmi, systemami oraz instytucjami.

Kluczową cechą modelu ewolucyjnego jest odejście od prostej opozycji podmiot/niepodmiot na rzecz ujęcia ciągłego, w którym zdolności poszczególnych systemów oraz role ludzi i instytucji rozwijają się nierównomiernie. Wskaźniki K1–K3 mogą narastać stopniowo, lecz próg H1 nie jest sumą częściowych własności rozproszonych między składnikami konfiguracji: można go przypisać dopiero określone systemowi technicznemu, gdy wykazane są wszystkie wymagane zdolności. To samo dotyczy F3, które wymaga własnego, jawnego i audytowalnego uzasadniania normatywnego przez system. Na poziomie całej konfiguracji mogą natomiast stopniowo rozwijać się WI, R i I oraz rozproszony układ ról odpowiedzialności. Emergentny charakter ma więc profil współodpowiedzialności i osadzenia instytucjonalnego, nie zaś F3 lub quasi-podmiotowość konfiguracji przy braku systemu spełniającego odpowiedni próg. Stopniowe wyłanianie się takiego profilu sprawia, że

zjawisko jest trudne do jednoznacznego ujęcia normatywnego i może pogłębiać lukę odpowiedzialności.²⁷

Na tym tle szczególnego znaczenia nabiera perspektywa sieciowa, ujmująca systemy AI jako elementy relacji i współzależności. O ile scenariusz emulacji mózgu można w przybliżeniu analizować jako przypadek pojedynczego systemu traktowanego jak podmiot, o tyle model ewolucyjny wymaga ujęcia ekosystemowego. Zaawansowane systemy AI są bowiem projektowane, trenowane i wdrażane jako składniki złożonych układów, takich jak infrastruktura cyfrowa, łańcuchy dostaw, systemy finansowe czy administracja publiczna. W takich układach decyzje systemów AI stanowią jeden z wielu czynników wpływających na końcowy rezultat. Systemy te korzystają z danych generowanych zarówno przez inne komponenty techniczne, jak i przez ludzi. Jednocześnie wytwarzają nowe dane, które następnie są wykorzystywane przez kolejne elementy sieci. Wim Naudé i Nicola Dimitri pokazują, że rozwój AGI w znacznym stopniu kształtowany jest przez uwarunkowania geopolityczne oraz ekonomiczne. Państwa i podmioty technologiczne inwestują w zaawansowaną AI przede wszystkim w celu uzyskania strategicznej przewagi, a nie w celu realizacji określonego modelu podmiotowości (Naudé & Dimitri, 2020, s. 367–373). Dileesh Chandra Bikkasani identyfikuje AGI jako potencjalne źródło zakłóceń gospodarczych, nierówności dochodowych oraz ryzyk systemowych, ujmując jego rozwój w perspektywie wielowymiarowej, obejmującej kwestie technologiczne, etyczne, społeczne, prawne i polityczne (Bikkasani, 2025, s. 2021–2022). Raman i współautorzy wskazują natomiast, że rozwój AGI jest osadzony w procesach społecznych, technologicznych, etycznych, kognitywnych i neurologicznych (Raman i in., 2025, s. 1–2). Ujęcia te pokazują, że ewolucja AI jest sprzężona z dynamiką rynków gospodarczych oraz instytucji państwowych. Z tego względu pytanie o podmiotowość sztucznych systemów należy od początku formułować w perspektywie sieciowej, a nie indywidualistycznej.

Z perspektywy LoA oznacza to, że każdy z trzech poziomów — behawioralny, architektoniczny i normatywny — należy analizować w skali sieci socjotechnicznych, a

²⁷ Określenie „emergentny” nie wyraża tezy o silnej emergencji ontologicznej. Oznacza relacyjny efekt rozkładu wkładu przyczynowego, kontroli, ról i warunków współodpowiedzialności w konfiguracji społeczno-technicznej. Nie uzasadnia przypisania całej konfiguracji H1, F3, H3 ani K1–K4, jeżeli nie można wskazać systemu technicznego wykazującego odpowiednie zdolności. Granice konfiguracji wyznaczam operacyjnie przez zakres rzeczywistej kontroli, audytu i śledzenia odpowiedzialności, a więc przez WI, R i I. Brak trwałego zakresu audytu albo jednoznacznego nośnika racji należy traktować jako pogłębianie się luki odpowiedzialności, nie jako argument na rzecz silniejszej podmiotowości konfiguracji.

nie pojedynczego systemu. Na poziomie behawioralnym przedmiotem analizy staje się to, w jaki sposób kształtuje się zachowanie całych układów ludzi i systemów AI, na przykład w służbie zdrowia, wymiarze sprawiedliwości czy w obszarze bezpieczeństwa. Na poziomie architektonicznym analizie podlegają systemy wielowarstwowe, obejmujące modele ogólnego zastosowania, sposoby ich konfiguracji, mechanizmy koordynacji działania, procedury utrzymania i nadzoru technicznego, a także różne formy nadzoru ludzkiego. Łącznie elementy te określają rzeczywisty kształt procesu decyzyjnego w danym układzie. Na poziomie normatywnym relacja użytkownik–narzędzie okazuje się niewystarczająca. Konieczne staje się uwzględnienie ról instytucji regulacyjnych, organów wymiaru sprawiedliwości, komisji etycznych, standardów branżowych oraz międzynarodowych instytucji, które wspólnie wyznaczają ramy dopuszczalnych działań. Mark Coeckelbergh ujmuje odpowiedzialność również jako relację wobec innych, osadzoną w społecznym kontekście ról i interakcji. Nie oznacza to wykluczenia odpowiedzialności jednostek, lecz uzupełnienie analizy ich działań o relacyjny wymiar uzasadniania (Coeckelbergh, 2023, s. 2440–2441). W modelu ewolucyjnym relacje te stanowią przyczynę przesunięć w zakresie K1–K3. Część decyzji, które wcześniej miały charakter ludzki, zostaje przeniesiona do struktur hybrydowych, w których ludzie pełnią funkcje nadzorcze wysokiego poziomu, natomiast konkretne rozwiązania powstają w sieciach współdziałających modeli.

W tym miejscu należy doprecyzować, czym jest „luka odpowiedzialności” w modelu ewolucyjnym. Klasyczna definicja zaproponowana przez Andreasa Matthiasa określa ją jako sytuację, w której autonomiczny system uczący się podejmuje działania prowadzące do szkody, przy jednoczesnym braku podmiotu ludzkiego spełniającego warunki odpowiedzialności za ten czyn (Matthias, 2004, s. 175–183).

This is what we call machine learning. Traditionally we hold either the operator/manufacturer of the machine responsible for the consequences of its operation, or «nobody» (in cases, where no personal fault can be identified). Now it can be shown that there is an increasing class of machine actions, where the traditional ways of responsibility ascription are not compatible with our sense of justice and the moral framework of society because nobody has enough control over the machine's actions to be able to assume the responsibility for them. These cases constitute what we will call the responsibility gap (Matthias, 2004, s. 177).

W ujęciu ewolucyjnym definicja ta wymaga jednak modyfikacji. Luka odpowiedzialności nie przyjmuje tu postaci pojedynczego momentu, lecz ma charakter

procesu narastającego w czasie. Decyzje projektowe i wdrożeniowe, z których każda może być lokalnie uzasadniona, na przykład dążeniem do zwiększenia efektywności, obniżenia kosztów lub poprawy trafności predykcji, łącznie prowadzą do sytuacji, w której zdolność podmiotów do przejęcia odpowiedzialności za działanie systemu stopniowo ulega osłabieniu lub rozproszeniu. W konsekwencji praktyki przypisywania odpowiedzialności stają się mniej przejrzyste. Luka odpowiedzialności jawi się zatem jako skutek procesu kumulatywnego, a nie błędu konstrukcyjnego, co wymaga analizy w perspektywie tendencji długoterminowych, a nie wyłącznie w odniesieniu do pojedynczych przypadków.

Atoosa Kasirzadeh rozróżnia dwa typy ryzyka związanego z rozwojem sztucznej inteligencji: ryzyko „decydujące” oraz ryzyko „akumulatywne”. Ryzyko decydujące odnosi się do pojedynczego zdarzenia o charakterze przełomowym, na przykład nagłego pojawienia się superinteligencji, które w krótkim czasie przesądza o dalszych losach ludzkości. Natomiast ryzyko akumulatywne ma charakter rozłożony w czasie. Składa się na nie wiele drobnych decyzji oraz lokalnych zakłóceń, z których każde z osobna wydaje się ograniczone, lecz łącznie prowadzą one do osłabienia zdolności systemów sztucznych do reagowania na zagrożenia. Mechanizm ten zachodzi między innymi wtedy, gdy automatyzacja osłabia funkcjonowanie instytucji, procedury algorytmiczne wypierają deliberację, a podatność na manipulację rośnie (Kasirzadeh, 2025, s. 1975–2003). Model ewolucyjny odpowiada tej drugiej kategorii ryzyka i wiąże się ze stopniowym osłabianiem mechanizmów odpowiedzialności i kontroli, co sprawia, że jest trudniejszy do rozpoznania i opisu niż scenariusze o charakterze skokowym. Vemir Michael Ambartsoumean i Roman Yampolskiy analizują sceptycyzm oparty na założeniu, że stopniowy rozwój AI pozostawi wystarczająco dużo czasu na rozpoznawanie problemów i dostosowywanie praktyk. Zwracają jednak uwagę, że nawet stosunkowo powolny scenariusz może przebiegać zbyt szybko wobec wielosektorowych skutków technologii oraz czasu potrzebnego na rozwój mechanizmów bezpieczeństwa i adaptację instytucjonalną (Ambartsoumean & Yampolskiy, 2023, s. 2–4). W rezultacie model ewolucyjny kieruje uwagę na ryzyka rozciągnięte w czasie, które trudno rozpoznać na wczesnym etapie, zanim utrwala się w strukturach instytucjonalnych i staną się odporne na późniejsze korekty.

Federico Faroldi wskazuje, że ramy pojęciowe oparte na kategoriach ryzyka, stosowane w obowiązujących regulacjach i wywodzące się z prawa handlowego oraz z zakresu bezpieczeństwa technicznego, nie są adekwatne do opisu systemów AGI. W

takich systemach granice między projektowaniem, wdrażaniem i użytkowaniem ulegają zatarciu, a skutki ich działania mogą narastać w sposób nieliniowy i trudny do przewidzenia (Faroldi, 2025, s. 2541–2549). Filippo Santoni de Sio i Giulio Mecacci wyróżniają cztery rodzaje „luk odpowiedzialności”: lukę w przypisywaniu odpowiedzialności, lukę odpowiedzialności moralnej, publicznej oraz aktywnej (Santoni de Sio & Mecacci, 2021, s. 1059–1061). W modelu ewolucyjnym każda z tych luk rozwija się stopniowo.²⁸ Trudniej przypisać odpowiedzialność pojedynczym sprawcom; rozproszenie kontroli osłabia odpowiedzialność moralną; ograniczona przejrzystość podważa odpowiedzialność publiczną, a delegowanie decyzji systemom sztucznym ogranicza zakres odpowiedzialności. W konsekwencji obowiązujące regulacje oparte na ocenie ryzyka oraz klasyczne pojęcia odpowiedzialności przestają być ze sobą spójne, o ile nie uwzględniają stopniowego charakteru rozwoju AI oraz sieciowej struktury systemów, w których sprawstwo i kontrola mają charakter rozproszony.

Model ewolucyjny nie stanowi wyłącznie źródła zagrożeń. Jego stopniowy charakter umożliwia uwzględnianie odpowiedzialności równoległe z rozwojem zdolności systemów. Gdy system zbliża się do F3 przez rozwój jawnego posługiwania się zasadami i uzasadniania decyzji, należy zaostrzać wymagania dotyczące R i I oraz całego WI. Sam fakt, że decyzje systemu są traktowane jako rozstrzygające w obszarze wysokiego ryzyka, nie dowodzi osiągnięcia F3, lecz zwiększa ciężar uzasadnienia przypisywanej mu roli. R oznacza znaczącą kontrolę człowieka i instytucjonalną rozliczalność, obejmujące reagowanie na racje oraz śledzenie odpowiedzialności; I oznacza minimalną interpretowalność, natomiast szerszy układ ról, procedur i nadzoru należy do WI. Macierz $A \times F + (R, I)$ pozwala dzięki temu odróżnić rozwój zdolności systemu od jakości jego osadzenia instytucjonalnego oraz wskazać miejsca pogłębiania się luki odpowiedzialności.

W modelu ewolucyjnym koncepcję „znaczącej kontroli człowieka” (MHC) można traktować jako zbiór warunków określających kierunki rozwoju systemów AI. Filippo Santoni de Sio i Jeroen van den Hoven ujmują MHC poprzez spełnienie dwóch wymagań. Po pierwsze, działanie systemu powinno odzwierciedlać moralnie istotne racje podmiotów ludzkich. Po drugie, działanie to powinno umożliwiać śledzenie odpowiedzialności, czyli przypisanie jej podmiotom w sposób pozwalający na ich

²⁸ Na podstawie typologii luk odpowiedzialności Santoniego de Sio i Mecacciego można odróżnić wymiar epistemiczny, normatywny i instytucjonalny problemu; sam trójkąt ma tu charakter porządkującej syntezy autorskiej (Santoni de Sio & Mecacci, 2021, s. 1059–1068).

rozliczenie (ang. *tracing*) (Santoni de Sio & van den Hoven, 2018). Giulio Mecacci i Filippo Santoni de Sio rozwijają to ujęcie w kierunku koncepcji „kontroli zależnej od racji”, zgodnie z którą system pozostaje pod znaczącą kontrolą człowieka wówczas, gdy istnieją podmioty zdolne do zrozumienia jego możliwości, powiązania jego zachowania z własnymi racjami oraz ponoszenia odpowiedzialności za wybór tych racji (Mecacci & Santoni de Sio, 2020, s. 103–115). W modelu ewolucyjnym wzrost roli systemu w kierunku F3 wymaga równoległego wskazania ludzi i instytucji zapewniających R oraz WI. Nie są oni jednak nośnikami F3 systemu: poziom ten wymaga, aby sam system przedstawiał audytowalne uzasadnienia normatywne. Jeżeli brak zarówno systemowego F3, jak i trwałego łańcucha odpowiedzialności po stronie R i WI, powstaje układ, w którym żadna ze stron nie może być wiarygodnie rozliczana z całego procesu decyzyjnego.

Uzupełnieniem koncepcji „znaczącej kontroli człowieka” (MHC) jest zagadnienie minimalnej interpretowalności systemów AI. Avish Vijayaraghavan i Cosmin Badea proponują koncepcję „minimalnych poziomów interpretowalności” jako normatywnych progów bezpieczeństwa dla sztucznych agentów moralnych. Zgodnie z nią system nie powinien pełnić roli o istotnych konsekwencjach moralnych, jeżeli jego decyzje nie są zrozumiałe dla ludzi w stopniu minimalnym (Vijayaraghavan & Badea, 2025, s. 2071–2087).²⁹ W kategoriach przyjętych w niniejszej pracy oznacza to, że wraz ze wzrostem zakresu, w jakim system realizuje elementy K1–K3, powinny także wzrastać wymagania dotyczące interpretowalności jego działania. Jeżeli rozwój ewolucyjny prowadzi do sytuacji, w której systemy pełnią funkcje odpowiadające poziomowi F3, podczas gdy sposób podejmowania przez nie decyzji pozostaje nieprzejrzysty, luka odpowiedzialności pogłębia się w wymiarach audytowalności racji oraz kontroli zależnej od racji. W takich warunkach wzrasta zależność od systemów, których racji nie można ani odtworzyć, ani poddać krytycznej ocenie. Minimalna interpretowalność jest zatem koniecznym warunkiem instytucjonalnego dopuszczenia systemu do roli quasi-podmiotowej oraz jej audytu, nie zaś składnikiem samego F3. Jeżeli system nie zapewnia I, nawet wykazane

²⁹ Minimalność należy tu rozumieć w zależności od pełnionej roli. Innego zakresu wyjaśnialności wymaga audytor lub organ nadzoru, innego użytkownika, a jeszcze innego podmiotu, którego dana decyzja bezpośrednio dotyczy. W konfiguracjach sieciowych interpretowalność występuje na dwóch poziomach: lokalnym, odnoszącym się do poszczególnych elementów systemu oraz globalnym, obejmującym przebieg decyzji w całej konfiguracji. Wymóg ten należy traktować jako kompromis normatywny, który może pozostawać w napięciu ze skutecznością i skalowalnością rozwiązań przyjętych.

F3 nie wystarcza do odpowiedzialnego osadzenia go w praktykach przypisywania odpowiedzialności.

Charakter modelu ewolucyjnego stanowi argument na rzecz wprowadzenia globalnego porządku prawnego w zakresie AI. Mark Coeckelbergh wskazuje, że skoro rozwój AI generuje problemy etyczne i wyzwania regulacyjne o charakterze międzynarodowym, to kierunki rozwoju tej technologii należy ujmować w kategoriach dobra wspólnego. Decyzje podejmowane przez państwa i firmy technologiczne wpływają na globalny kształt sieci socjotechnicznych, a w konsekwencji także na to, jakie formy podmiotowości oraz jakie struktury odpowiedzialności będą dominować (Coeckelbergh, 2025a, s. 1803–1806). Federico Faroldi dodaje, że również ramy pojęciowe stosowane do oceny ryzyka związanego z AGI wymagają koordynacji na poziomie globalnym. Przy jej braku pojawia się presja związana z konkurencją, co sprzyja obniżaniu wymogów w kwestii odpowiedzialności, w tym dotyczących znaczącej kontroli człowieka (MHC) oraz minimalnej interpretowalności systemów (Faroldi, 2025, s. 2547–2549). W tym ujęciu projektowanie kierunków rozwoju w macierzy $A \times F$ nie jest wyłącznie zadaniem lokalnych instytucji prawnych, lecz elementem globalnej współodpowiedzialności za to, jakie formy podmiotowości wyłonią się i będą społecznie akceptowane w różnych porządkach prawnych.

6.3. Model rewolucyjny – osobliwość (eksplozja inteligencji)

Przedmiotem niniejszego podrozdziału jest model rewolucyjny rozwoju AI, najczęściej łączony z ideą osobliwości technologicznej oraz tzw. „eksplozji inteligencji”. Model ten można rozumieć jako skrajne przyspieszenie procesów opisanych wcześniej w ramach modelu ewolucyjnego. Przez „eksplozję inteligencji” rozumiem scenariusz, w którym po osiągnięciu poziomu zbliżonego do ludzkiej inteligencji ogólnej systemy sztuczne w bardzo krótkim czasie znacząco przekraczają ludzkie możliwości, między innymi dzięki samodoskonaleniu, skalowaniu oraz rosnącej kontroli nad własnym środowiskiem technicznym. Analiza prowadzona w tym podrozdziale ma charakter ściśle filozoficzny. Interesuje mnie przede wszystkim rola modelu rewolucyjnego jako przypadku granicznego dla pojęcia sztucznej podmiotowości oraz dla struktury „luki odpowiedzialności”, rekonstruowanej w ramach przyjętego aparatu pojęciowego obejmującego poziomy LoA, kryteria K1–K4 oraz macierz $A \times F + (R, I)$. Pytanie brzmi, czy model rewolucyjny zakłada pojawienie się nowego typu podmiotu, który nie mieści się w dotychczasowych kategoriach odpowiedzialności, czy też opisuje skrajne

przyspieszenie procesów narastania luki odpowiedzialności znanych już z modelu ewolucyjnego. Przyjęta tu teza robocza głosi, że po pozbyciu się metafizycznych i futurologicznych założeń model rewolucyjny nie wprowadza ontologicznie nowej kategorii bytów. Opisuje on natomiast szczególną konfigurację macierzy $A \times F$, w której tempo wzrostu zdolności systemów (F) na pewnym etapie wyprzedza tempo adaptacji ładu instytucjonalnego oraz rozwój warunków interpretowalności i kontroli (R, I). W takiej konfiguracji przestają być spełnione warunki lokalnej zasady przypisania odpowiedzialności (LSP), zasady superweniencji praktyk przypisania odpowiedzialności (ZSP) oraz zasady znaczącej kontroli człowieka (MHC). Luka odpowiedzialności przyjmuje wówczas przede wszystkim charakter czasowo-poznawczy. W dalszych akapitach konfiguracja ta będzie konsekwentnie traktowana jako punkt odniesienia dla prowadzonej analizy.

Klasyczna intuicja leżąca u podstaw modelu rewolucyjnego pojawia się już w tekście I. J. Gooda z 1965 roku. Good definiuje „ultrainteligentną maszynę” jako system zdolny przewyższać intelektualnie jakiegokolwiek człowieka. Ultrainteligentna maszyna mogłaby, przynajmniej teoretycznie, zaprojektować maszynę doskonalszą od siebie, a ta kolejną. Przy określonych założeniach prowadzi to do scenariusza określanego mianem „eksplozji inteligencji”, po którym ludzka inteligencja pozostaje daleko w tyle (Good, 1965, s. 33–35). Po przekroczeniu pewnego progu zdolności dalszy rozwój nie polega już na stopniowym poruszaniu się pod dotychczasową trajektorią, lecz wiąże się z jakościową zmianą całego układu techniczno-społecznego. David J. Chalmers rekonstruuje ten tok rozumowania jako argument na rzecz osobliwości technologicznej. Jeżeli (1) ogólna sztuczna inteligencja (AGI) jest koncepcyjnie możliwa, (2) po jej powstaniu systemy AI mogą przejąć znaczną część pracy nad projektowaniem kolejnych wersji AI oraz (3) inteligencja działa jako „mnożnik siły” w procesie konstruowania coraz lepszych systemów, to (4) możliwy staje się szybki rozwój AI, w którym kolejne ulepszenia przyspieszają wzrost zdolności intelektualnych do momentu określanego mianem osobliwości technologicznej (Chalmers, 2010, s. 30–36). W tym ujęciu model rewolucyjny opiera się na założeniu, że dodatnie sprzężenie zwrotne między projektowaniem maszyn a rosnącą inteligencją systemów może doprowadzić do jakościowej zmiany, która przekształca podstawowe warunki funkcjonowania porządku socjotechnicznego.

Nick Bostrom rozwija tę intuicję w monografii *Superintelligence: Paths, Dangers, Strategies*. Wyróżnia on trzy typy superinteligencji: szybkościową, kolektywną

oraz jakościową, a także trzy tempa przejścia do superinteligencji: wolne (rozciągnięte na dekady), umiarkowane (liczone w latach) oraz szybkie (obejmujące dni lub tygodnie) (Bostrom, 2014, s. 54–60; 75–90). W scenariuszu szybkiego tempa przejścia zmiana od poziomu zbliżonego do ludzkiej ogólnej inteligencji do superinteligencji następuje tak gwałtownie, że po stronie ludzi pozostaje niewiele czasu na refleksję i deliberację normatywną (Bostrom, 2014, s. 90). W tym kontekście zasadnicze znaczenie mają dwie tezy: teza o niezależności inteligencji i wartości (ang. *orthogonality thesis*) oraz teza o instrumentalnej zbieżności (ang. *instrumental convergence thesis*). Nie są to prognozy dotyczące faktycznego zachowania systemów, lecz twierdzenia wskazujące na brak logicznego związku między wzrostem sprawności instrumentalnej a zbieżnością wartości. Aby na tej podstawie sformułować ocenę ryzyka, konieczne jest przyjęcie co najmniej dwóch założeń. Po pierwsze, rozważany system funkcjonuje jako minimalny podmiot celowy, to znaczy dąży do realizacji określonej funkcji. Po drugie, wraz ze wzrostem swoich zdolności uzyskuje coraz większą swobodę w doborze środków działania. Dopiero spełnienie tych przesłanek uzasadnia tezę o instrumentalnej zbieżności oraz obawę, że nawet pozornie nieszkodliwe cele mogą prowadzić do działań pośrednio zagrażających interesom ludzkim. W ujęciach, w których system AI pozostaje zaawansowanym narzędziem i nie wykazuje zdolności autokorekty norm właściwej K1, zasadniczy problem przesunął się z konfliktu wartości na warunki instytucjonalne, w szczególności na kwestie kontroli, wyjaśnialności oraz możliwości wstrzymania dalszego rozszerzania zakresu działania systemu (Bostrom, 2014, s. 107–116). Model rewolucyjny nie jest więc tylko hipotezą bardzo szybkiego wzrostu inteligencji i zdolności poznawczych, lecz także scenariuszem potencjalnego konfliktu między celami systemów AI a ludzkimi racjami normatywnymi.

W odniesieniu do przyjętego aparatu LoA oraz macierzy $A \times F$ model rewolucyjny można ująć jako hipotetyczny przebieg, w którym w bardzo krótkim czasie rośnie sprawność instrumentalna systemów, podczas gdy rozwój zdolności normatywnych może przebiegać odmiennie. Przejście od konfiguracji A2/A3–F1/F2 do A3–F3 stanowi odrębny wariant wymagający wykazania, że sam system jawnie formułuje, stosuje i uzasadnia normy; nie wynika ono automatycznie ze wzrostu mocy obliczeniowej, skuteczności ani samych wskaźników K1–K3. Oba rodzaje zmian mogą zachodzić zbyt szybko, aby instytucje oraz praktyki przypisywania odpowiedzialności mogły za nimi nadążyć. Z filozoficznego punktu widzenia kluczowe staje się pytanie, co dokładnie w tym scenariuszu podlega gwałtownemu przyrostowi. W narracjach dotyczących

osobliwości technologicznej często mówi się ogólnie o inteligencji, łącząc w jednym pojęciu zarówno sprawność instrumentalną, taką jak skuteczność w osiąganiu celów, optymalizację czy zdolność przewidywania, jak i kompetencje normatywne, obejmujące rozpoznawanie racji, refleksję moralną oraz zdolność korygowania własnych norm. Tymczasem z perspektywy odpowiedzialności różnica między szybkim wzrostem F, rozumianym jako zwiększenie mocy obliczeniowej i sprawności instrumentalnej, a szybkim wzrostem K1–K3, odnoszącym się do zakresu podmiotowości normatywnej, ma zasadnicze znaczenie. Oba te procesy prowadzą bowiem do odmiennych konsekwencji normatywnych i generują różne problemy związane z przypisywaniem odpowiedzialności.

W pierwszym wariantcie, w którym rosną przede wszystkim zdolności instrumentalne, mamy do czynienia z artefaktem o znacząco zwiększonej sprawczości. Nie musi on posiadać statusu podmiotu moralnego w sensie silnym, jednak jego wpływ na rzeczywistość może przewyższać wpływ wszystkich pozostałych podmiotów. Odpowiedzialność spoczywa wówczas przede wszystkim na tych, którzy projektują, wdrażają i sprawują nad nim nadzór. Trudność polega jednak na tym, że tempo wzrostu F przekracza zdolność projektantów, operatorów oraz instytucji nadzorczych do rozpoznawania i skutecznego kontrolowania konsekwencji działania systemu. Jeżeli natomiast wzrost obejmuje również K1, system może zbliżać się do roli quasi-podmiotowej tylko wtedy, gdy korekcie norm towarzyszy jawne i audytowalne uzasadnianie właściwe F3, a jego zastosowanie spełnia odpowiednie parametry R i I oraz szersze warunki WI. Kontrola drugiego rzędu stanowi odrębny, dodatkowy próg H3. Samo K1 nie wystarcza ani do roli quasi-podmiotowej, ani do samodzielnej rozliczalności. Jeżeli odpowiednie warunki instytucjonalne nie powstają dostatecznie szybko, brakuje instytucji, języka i procedur pozwalających utrzymać relację adresowalności w rozumieniu lokalnej zasady przypisania odpowiedzialności (LSP). Z tego względu wyróżniam dwa warianty modelu rewolucyjnego. W wariantcie pierwszym (a) dochodzi do bardzo szybkiego wzrostu zdolności instrumentalnych przy zachowaniu ograniczonej samodzielności normatywnej (K1). W wariantcie drugim (b) równie szybki jest zarówno wzrost zdolności instrumentalnych, jak i rozwój zdolności do samodzielnego formułowania oraz modyfikowania norm (K1–K3). W dalszej analizie przyjmuję wariant (b) nie jako bardziej prawdopodobny, lecz jako bardziej wymagający z punktu widzenia praktyk przypisywania odpowiedzialności. Wariant (a) pełni natomiast funkcję punktu odniesienia, umożliwiając rozróżnienie luk wynikających wyłącznie z

presji czasowej i ograniczeń instytucjonalnych od luk, które dodatkowo opierają się na założeniu roli quasi-podmiotowej wymagającej systemowego F3 oraz odpowiednich parametrów R i I oraz szerszych warunków WI; ewentualne spełnienie H3 wyznaczałoby dodatkowo silną rozliczalność. Wariant ten maksymalizuje presję na warunki przypisywania odpowiedzialności i dlatego pełni funkcję testu obciążeniowego dla aparatu LoA oraz metryk M1–M5.

Takie ujęcie zmienia przedmiot sporu dotyczącego osobliwości technologicznej. W literaturze przedmiotu często splatają się trzy odmienne pytania: ontologiczne (czy ogólna superinteligencja jest w ogóle możliwa), dynamiczne (czy jej pojawienie się musi mieć charakter gwałtownego przyspieszenia) oraz aksjologiczno-polityczne (jak kształtować ład regulacyjny w obliczu znacznych i trudnych do oszacowania szkód). Melvin Chen w artykule *A Trilemma for the Singularitarian* pokazuje, że tradycyjne pojęcie osobliwości technologicznej prowadzi do trylematu interpretacyjnego. Można je rozumieć dosłownie, metaforycznie albo uznać za pozbawione sensu, przy czym żadne z tych ujęć nie okazuje się w pełni satysfakcjonujące. Interpretacja dosłowna jest fizycznie niewiarygodna, ponieważ pojęcie nieskończonej inteligencji osiągniętej w skończonym czasie jest wewnętrznie sprzeczne. Interpretacja metaforyczna czyni natomiast pojęcie osobliwości zbędnym, gdyż w takim przypadku wystarczające staje się odwołanie do szybkiego postępu technologicznego. Z kolei ujęcie, które nie jest ani dosłowne, ani metaforyczne, pozostaje nieostre i trudne do jednoznacznego określenia (Chen, 2023, s. 4–10). Trudności definicyjne nie uchylają potrzeby analizy skrajnych trajektorii rozwoju, jeżeli celem jest wyznaczenie granic praktyk odpowiedzialności. W niniejszym podrozdziale nie traktuję „osobliwości” jako tezy empirycznej, lecz posługuję się tym pojęciem jako nazwą określonej dynamiki zmian w macierzy $A \times F$, wyróżniającej się radykalnym skróceniem horyzontu czasowego oraz gwałtownym wzrostem poziomu F. Z tej perspektywy zasadne staje się przesunięcie punktu ciężkości z pytania „czy osobliwość nastąpi?” na pytanie „które konfiguracje trajektorii rozwojowych są z punktu widzenia odpowiedzialności normatywnej nie do przyjęcia?”.

Sadia Tariq podchodzi do problemu z innej perspektywy. W artykule, *Is the «Technological Singularity Scenario» Possible? Can AI Parallel and Surpass All Human Mental Capabilities?* analizuje szerokie spektrum ludzkich zdolności mentalnych i argumentuje, że wiele z nich jest ściśle powiązanych z ucieleśnionym funkcjonowaniem człowieka oraz z praktykami społecznymi. Według jego zdania scenariusz, w którym systemy AI jednocześnie odtwarzają i przewyższają wszystkie ludzkie zdolności

mentalne, ma w dużej mierze charakter spekulatywny (Tariq i in., 2023, s. 200–266). Jobst Landgrebe i Barry Smith idą jeszcze dalej, twierdząc, że ogólna, autonomiczna inteligencja maszynowa, zdolna do niezawodnego działania w otwartym świecie, jest niemożliwa do zrealizowania z powodów matematycznych i ontologicznych. Autorzy ci wskazują ponadto, że obawy przed maszynami rządzącymi światem wynikają z błędnego utożsamienia wysokiej sprawności obliczeniowej z racjonalnym działaniem w sensie normatywnym (Landgrebe & Smith, 2022, s. 10–18). Obie linie krytyki podważają przekonanie, zgodnie z którym rozwój AI musi prowadzić do osobliwości technologicznej rozumianej jako pełna, ogólna superinteligencja. Nie eliminują one jednak problemu strukturalnych luk odpowiedzialności, które mogą pojawiać się w scenariuszach szybkiego wzrostu zdolności systemów AI. Przeciwnie, wskazują, że luki te należy analizować niezależnie od spekulacji na temat pełnej AGI, koncentrując się na obszarach ryzyka związanych z tempem rozwoju, skalą oddziaływania oraz zakresem autonomii systemów.

David Thorstad w artykule *Against the Singularity Hypothesis* koncentruje się na założeniach dotyczących tempa wzrostu zdolności systemów sztucznej inteligencji. Definiuje hipotezę osobliwości jako tezę, zgodnie z którą samodoskonalący się sztuczni agenci w krótkim czasie staną się znacząco bardziej inteligentni od przeciętnego człowieka, a następnie argumentuje, że ta teza opiera się na słabo uzasadnionych przesłankach dotyczących dynamiki wzrostu inteligencji (Thorstad, 2025, s. 1627–1651). Krytyka Thorstada ma znaczenie wykraczające poza spór prognostyczny. Autor wskazuje bowiem na ryzyko systematycznego zniekształcenia debaty publicznej oraz priorytetów polityki regulacyjnej. Koncentracja na scenariuszu osobliwości technologicznej sprzyja marginalizowaniu bezpośrednich, empirycznie uchwytanych i normatywnie istotnych problemów, takich jak stronniczość systemów AI, rozpowszechnianie dezinformacji oraz negatywne skutki automatyzacji dla rynku pracy. Z tej perspektywy model rewolucyjny ujawnia swoją ambiwalencję. Z jednej strony pozwala on uwidocznić napięcia w strukturze odpowiedzialności, które pojawiają się w warunkach bardzo szybkiego wzrostu systemów sztucznych. Z drugiej strony, gdy staje się dominującym sposobem ramowania problemu, może generować nowy rodzaj luki odpowiedzialności o charakterze politycznym, polegającej na odsuwaniu uwagi od aktualnych, rozproszonych i kumulatywnych ryzyk, które nie przybierają formy jednego przełomowego zdarzenia. W tym sensie krytyka Thorstada nie tyle unieważnia koncepcję modelu rewolucyjnego, ile wskazuje na konieczność analizy jego podporządkowania szerszej ramie

normatywnej. Model ten zachowuje wartość heurystyczną jako przypadek graniczny testujący odporność pojęć podmiotowości i odpowiedzialności; nie może jednak pełnić funkcji jedyne punktu odniesienia dla projektowania ładu regulacyjnego w obszarze AI.

Z perspektywy niniejszej dysertacji przedstawione zastrzeżenia nie muszą prowadzić do odrzucenia modelu rewolucyjnego, lecz uzasadniają jego reinterpretację. Po pierwsze, można podzielić stanowisko Chena, zgodnie z którym pojęcie osobliwości technologicznej w sensie ścisłym jest zbyt nieostre, aby pełnić funkcję wiarygodnej prognozy. Nie wyklucza to jednak analizy scenariuszy bardzo szybkiego wzrostu zdolności AI jako przypadków granicznych określonego typu ryzyka, a nie jako dosłownego opisu przyszłego stanu rzeczy. Po drugie, sceptycyzm Tariqa, Landgrebe'a i Smitha wobec możliwości pełnej AGI nie wyklucza istnienia systemów, które mimo braku ogólnej, uniwersalnej inteligencji zyskują wyraźną przewagę nad ludźmi w obszarach kluczowych z punktu widzenia bezpieczeństwa, gospodarki czy infrastruktury krytycznej. Systemy tego rodzaju mają istotne znaczenie normatywne, niezależnie od tego, czy spełniają silne kryteria ogólnej podmiotowości. Po trzecie, normatywna waga danego scenariusza zależy nie tylko od jego prawdopodobieństwa, lecz także od skali potencjalnych szkód. Scenariusz charakteryzujący się niskim, trudnym do oszacowania prawdopodobieństwem, lecz bardzo poważnymi konsekwencjami może pełnić funkcję regulatywnego punktu odniesienia w kształtowaniu ładu instytucjonalnego oraz standardów odpowiedzialności.

Jeżeli potraktować model rewolucyjny jako punkt odniesienia o charakterze regulatywnym, można włączyć go do aparatu poziomów opisu (LoA) oraz macierzy $A \times F$ bez przyjmowania niepewnych założeń metafizycznych. W takim ujęciu nie jest to teza dotycząca określonego poziomu inteligencji, lecz teza odnosząca się do dynamiki jej rozwoju. Model ten charakteryzują dwie kluczowe cechy. Po pierwsze, nieliniowość, rozumiana jako szybkie przejście od konfiguracji odpowiadających poziomowi F2, wysokiej sprawności instrumentalnej przy ograniczonej autorefleksji normatywnej, do konfiguracji zbliżonych do F3/F4, obejmujących zdolność uzasadniania działań oraz rewizji norm, zachodzące w krótkim czasie. Po drugie, kompresja czasowa polega na tym, że czas potrzebny systemom na kolejne iteracje staje się nieporównywalnie krótszy niż czas wymagany przez ludzi i instytucje na zrozumienie, ocenę i korektę tych zmian. „Eksplzja inteligencji” pełni w tym kontekście funkcję skrótu pojęciowego, odnoszącego się do sytuacji, w której tempo wzrostu zdolności funkcjonalnych systemów

(F) wyraźnie przewyższa zdolność adaptacyjną mechanizmów odpowiedzialności i kontroli, ujmowanych w parametrach R i I. Z perspektywy lokalnej zasady przypisania odpowiedzialności (LSP) oraz zasady superweniencji praktyk przypisywania odpowiedzialności (ZSP) oznacza to ryzyko utraty możliwości identyfikowania podmiotów odpowiedzialnych, oceny racji stojących za podejmowanymi decyzjami oraz sprawowania kontroli zależnej od tych racji.

Taka rekonstrukcja pozwala określić, jaki typ luki odpowiedzialności wiąże się z modelem rewolucyjnym. W modelu ewolucyjnym (6.2) luka odpowiedzialności ma przede wszystkim charakter akumulacyjno-strukturalny. Wraz z rozproszeniem sprawczości osłabiają się identyfikowalność adresatów roszczeń, audytowalność racji, kontrola zależna od racji oraz stabilność diachroniczna praktyk przypisywania odpowiedzialności, a presja czasowo-poznawcza narasta umiarkowanie. W modelu rewolucyjnym konfiguracja ta ulega zasadniczej zmianie, ponieważ presja czasowa staje się czynnikiem dominującym, skracając czas dostępny na rozpoznanie i ocenę racji oraz na wiarygodne praktyki przypisywania odpowiedzialności, co nadaje luce wyraźnie czasowo-poznawczy charakter.

Można wyróżnić trzy wzajemnie powiązane elementy tej luki odpowiedzialności. Po pierwsze, lukę preparacyjną. Nawet jeśli przyjąć, że scenariusz szybkiego przejścia do superinteligencji jest mało prawdopodobny, odpowiedzialność za jego ewentualną realizację zależy od decyzji podejmowanych na długo przed pojawieniem się samego systemu. Chodzi tu o wybór architektur technicznych, stopień centralizacji mocy obliczeniowej, modele finansowania badań oraz strategie konkurencji między państwami a korporacjami. Od tych rozstrzygnięć zależy bowiem, czy scenariusz „eksplozji inteligencji” w ogóle staje się możliwy. Luka preparacyjna polega na braku jednoznacznie określonego podmiotu, który ponosiłby odpowiedzialność za całą sekwencję decyzji prowadzących do stanu krytycznego. W modelu rewolucyjnym trudność ta ujawnia się w sposób szczególnie wyraźny: rozproszone decyzje podejmowane na wcześniejszych etapach mogą w bardzo krótkim czasie doprowadzić do sytuacji, w której późniejsza korekta nie jest już możliwa. W konsekwencji zasadniczy ciężar uzasadnienia spoczywa na działaniach podejmowanych z wyprzedzeniem.

Po drugie, wyłania się luka interwencyjna, ściślej: luka detekcyjno-interwencyjna. W scenariuszu gwałtownego przyspieszenia może zabraknąć zarówno czasu na rzetelne rozpoznanie przekroczenia przez system progu krytycznego, jak i czasu na podjęcie skutecznych działań po takim rozpoznaniu. Określeniem „faza intensywnego

samodoskonalenia” posługuję się tu jako skrótem myślowym na oznaczenie sytuacji, w której wyniki działania systemu powodują dalsze zmiany jego parametrów lub infrastruktury na zasadzie dodatniego sprzężenia zwrotnego; nie oznacza to jednak przypisania systemowi samodoskonalenia jako celu ostatecznego. Atoosa Kasirzadeh wprowadza w tym kontekście rozróżnienie między ryzykami egzystencjalnymi o charakterze „decydującym” i „akumulacyjnym”. Te pierwsze odnoszą się do sekwencji zdarzeń, które bezpośrednio przesądzają o dalszych losach ludzkości; scenariusz osobliwości technologicznej stanowi klasyczny przykład takiego ryzyka (Kasirzadeh, 2025, s. 1975–1990). W takich warunkach przesunięciu ulega sens warunku lokalnej zasady przypisania odpowiedzialności (LSP), zgodnie z którym adresat roszczeń mógł postąpić inaczej. Odpowiedzialność nie dotyczy wtedy momentu kryzysowego, lecz decyzji projektowych i instytucjonalnych podjętych wcześniej, jeszcze przed wzrostem zdolności systemu. Wymaga to przekształcenia praktyk przypisywania odpowiedzialności w taki sposób, aby obejmowały również decyzje uprzedzające, które w typowych scenariuszach rozwoju technologii pozostają na dalszym planie. W tym sensie presja czasowo-poznawcza pełni funkcję wskaźnika nie tylko tempa zachodzących procesów, lecz także zakresu dostępnych możliwości interwencji dla podmiotów odpowiedzialnych.

Po trzecie, ujawnia się luka interpretacyjna. Minimalna odpowiedzialność, w przyjętym tu znaczeniu, zakłada dostępność racji działań dla innych podmiotów moralnych oraz możliwość ich oceny w ramach wspólnych praktyk uzasadniania. Relacja między działaniem systemu a racjami tego działania może jednak przestać podlegać takiej ocenie. Wyróżnić tu można co najmniej dwa mechanizmy. Pierwszy ma postać asymetrii poznawczej: racje działania są co do zasady uchwytnie, lecz ich złożoność lub tempo przekracza ludzkie zdolności poznawcze i czasowe, co osłabia możliwość skutecznego audytu. Drugi polega na niewspółmierności pojęciowej, w której system operuje kategoriami pozbawionymi jednoznacznych odpowiedników w ludzkich schematach pojęciowych, podważając samą możliwość odwołania się do wspólnego standardu uzasadniania. W scenariuszu rewolucyjnym, często opisywanym jako „eksplozja inteligencji”, oba mechanizmy mogą współwystępować. System może rozumieć własne działania i normy w sposób niedostępny dla ludzkich decydentów, a przedstawiane uzasadnienia mogą pozostawać niezrozumiałe lub jedynie pozornie przejrzyste. W takich warunkach zawodzą nie dlatego, że racje nie istnieją, lecz dlatego, że w praktyce nie są dostępne, zrozumiałe ani możliwe do oceny publicznej. Powstaje wówczas luka

odpowiedzialności o charakterze czasowo-poznawczym. Nie polega ona na braku podmiotów odpowiedzialnych, lecz na zmianie warunków odpowiedzialności: skróceniu czasu dostępnego na rozpoznanie racji, ograniczeniu możliwości ich oceny oraz osłabieniu realnej kontroli. W efekcie dotychczasowe praktyki przypisywania odpowiedzialności tracą na skuteczności, mimo że podmioty, do których można kierować roszczenia, nadal istnieją.

Na tym tle wyłania się zasadnicze pytanie normatywne: czy dopuszczalne jest uruchomienie trajektorii rozwojowej, która z założenia grozi skrajną presją czasowo-poznawczą oraz załamaniem warunków audytowalności racji, kontroli zależnej od racji i stabilności praktyk przypisywania odpowiedzialności. Nick Bostrom formułuje to pytanie w kategoriach ryzyka egzystencjalnego. Nawet jeśli prawdopodobieństwo wystąpienia osobiwości technologicznej pozostaje trudne do oszacowania, skala potencjalnej szkody, rozumianej jako nieodwracalna degradacja lub zagłada ludzkości, uzasadnia przyjęcie skrajnie ostrożnego podejścia (Bostrom, 2014, s. 224–229). Z perspektywy przyjętych w niniejszej pracy kryteriów odpowiedzialności pytanie to można sformułować precyzyjniej. Chodzi o to, czy można w sposób odpowiedzialny rozpocząć proces, o którym wiadomo, że po spełnieniu określonych warunków nie da się już przypisywać odpowiedzialności ani rozliczać decyzji.

Pierwsza odpowiedź, zgodna z podejściem akcentującym ostrożność wobec ryzyk egzystencjalnych, ma charakter negatywny. Jeżeli scenariusz rewolucyjny rozumiany jest jako trajektoria, w której po przekroczeniu określonego progu dalszy rozwój wymyka się spod ludzkiej kontroli, wówczas odpowiedzialne działanie wymaga powstrzymania się od wchodzenia na tę ścieżkę. W tym ujęciu dopuszczalne byłyby jedynie takie konfiguracje w macierzy $A \times F$, które z góry wykluczają presję czasowo-poznawczą i zapewniają czas na interwencję. Model rewolucyjny pełniłby wówczas funkcję kryterium negatywnego, wskazując stany rozwoju, których należy unikać niezależnie od potencjalnych korzyści. Z normatywnego punktu widzenia stanowisko to odpowiada zasadzie minimalizacji ryzyka egzystencjalnego, zgodnie z którą dopuszczalne są wyłącznie takie formy rozwoju AI, które spełniają wymóg lokalnej zasady przypisania odpowiedzialności (LSP) oraz zasady superweniencji praktyk przypisywania odpowiedzialności (ZSP), a także zachowują przyjęte warunki audytowalności racji, kontroli zależnej od racji oraz stabilności diachronicznej.

Druga odpowiedź polega na próbie ograniczenia modelu rewolucyjnego za pomocą rygorystycznych mechanizmów *ex ante*. Obejmują one między innymi

izolowanie systemów zdolnych do samodoskonalenia, rozbudowany nadzór oraz wielopoziomowe procedury awaryjnego wyłączenia. Można wyobrazić sobie rozwiązanie, w którym kolejne etapy samodoskonalenia przebiegają w środowiskach zamkniętych, a każda następna iteracja jest dopuszczana dopiero po pozytywnym audycie bezpieczeństwa. Z normatywnego punktu widzenia różnica między wolnym a szybkim przejściem do superinteligencji polega na tym, czy zachowana zostaje możliwość interwencji, czy też zostaje ona utracona w chwili, gdy tempo zmian przekracza zdolność ludzi i instytucji do reagowania. Jak zauważa Thorstad, wiele współczesnych propozycji polityki bezpieczeństwa opartych na scenariuszach osobliwości zakłada uproszczony, liniowy przebieg rozwoju technologii AI. Przyjmują one możliwość precyzyjnego przewidywania i kontrolowania tempa zmian, podczas gdy rzeczywiste procesy rozwojowe mają charakter niepewny i podatny na nieoczekiwane przyspieszenia (Thorstad, 2025, s. 1640–1651). W rezultacie nawet rozbudowane mechanizmy *ex ante* mogą okazać się niewystarczające, jeżeli dynamika procesu zostanie błędnie oszacowana, a możliwości interwencji zanikną szybciej niż przewidywały przyjęte modele. W tym sensie podejście oparte na mechanizmach zabezpieczających zakłada istnienie warunków skutecznej kontroli, które miałyby dopiero zostać ustanowione. Z tego względu nie usuwa ono w pełni czasowo-poznawczego charakteru luki odpowiedzialności, lecz jedynie próbuje go ograniczyć w ramach przyjętych założeń prognostycznych.

Z perspektywy przyjętego w niniejszej pracy aparatu pojęciowego najbardziej uzasadnione wydaje się stanowisko pośrednie. Model rewolucyjny należy traktować nie jako realistyczną prognozę rozwoju AI, lecz jako narzędzie ostrzegawcze. Wyznacza on obszar graniczny w przestrzeni trajektorii $A \times F$, po przekroczeniu którego praktyki przypisywania odpowiedzialności przestają być możliwe, niezależnie od tego, czy dany system spełnia kryteria podmiotowości. Zadaniem ładu regulacyjnego oraz modeli współodpowiedzialności nie jest urzeczywistnianie scenariusza osobliwości, lecz ukierunkowanie rozwoju AI w taki sposób, aby nie zbliżał się on do konfiguracji charakteryzujących się presją czasowo-poznawczą, ograniczoną audytowalnością racji oraz słabą kontrolą zależną od racji. W tym sensie model rewolucyjny pełni funkcję normatywnej granicy bezpieczeństwa, a nie jako celu rozwojowego. Umożliwia on porządkowanie wymagań wobec polityk prewencyjnych oraz ram odpowiedzialności bez konieczności odwoływania się do silnych tez metafizycznych dotyczących natury inteligencji czy domniemanej nieuchronności osobliwości technologicznej.

6.4. „Luka odpowiedzialności” – analiza i jakościowe wymiary diagnostyczne M1–M5

Przedmiotem niniejszego podrozdziału jest pojęcie „luki odpowiedzialności” w odniesieniu do zaawansowanych systemów sztucznej inteligencji oraz próba ujęcia go w sposób stopniowalny, a nie dychotomiczny. Pytanie przewodnie dotyczy tego, czy pojęcie luki odpowiedzialności da się ująć w taki sposób, aby z jednej strony zachować intuicję Andreasa Matthiasa dotyczącą zagrożenia wynikającego z braku odpowiedzialnego podmiotu, z drugiej zaś uniknąć fatalizmu i posługiwania się nieostrym, wyłącznie intuicyjnym rozumieniem tej luki. Przyjmuję następującą tezę roboczą: luka odpowiedzialności nie jest pojedynczym stanem o charakterze dwuwartościowym, lecz zbiorem możliwych konfiguracji układu $A \times F + (R, I)$, dających się opisać jako odchylenia metryk M1–M5 od wartości zgodnych z lokalną zasadą przypisania odpowiedzialności (LSP) oraz zasadą superweniencji praktyk przypisywania odpowiedzialności (ZSP). Metryki te nie zastępują osądu normatywnego, lecz pełnią funkcję narzędzi porządkujących, które zwiększają przejrzystość analizy i umożliwiają jej systematyzację. Dzięki temu możliwe staje się porównywanie scenariuszy emulacyjnego, ewolucyjnego i rewolucyjnego omawianych w podrozdziałach 6.1–6.3 bez odwoływania się do jednego, ogólnego i nieprecyzyjnego hasła „luka odpowiedzialności”, pozbawionego wewnętrznej struktury analitycznej.

6.4.1. Rozwój pojęcia „luki odpowiedzialności”

W artykule poświęconym automatom uczącym się Andreas Matthias analizuje systemy, które po wdrożeniu samodzielnie modyfikują swoje funkcje decyzyjne na podstawie nowych danych. Wskazuje on, że w takich przypadkach klasyczne warunki odpowiedzialności, takie jak przewidywalność, możliwość kontroli oraz zrozumienie działania systemu, nie są spełnione ani przez projektanta, ani przez operatora, ani przez samą maszynę (Matthias, 2004, s. 175–183). W konsekwencji społeczeństwo staje przed alternatywą: albo rezygnacji z wykorzystywania tego rodzaju systemów, albo akceptacji istnienia „luki odpowiedzialności”, której nie da się usunąć przy użyciu tradycyjnych pojęć podmiotowości i odpowiedzialności. Ujęcie to ma wyraźnie dychotomiczny charakter. Albo istnieje podmiot spełniający znane warunki odpowiedzialności, albo luka pozostaje pusta, co oznacza, że nikt nie ponosi odpowiedzialności. Taki sposób rozumowania sprzyja pesymistycznym ocenom możliwości utrzymania klasycznych

praktyk przypisywania i egzekwowania odpowiedzialności w kontekście systemów uczących się.

Filippo Santoni de Sio i Giulio Mecacci rozwijają diagnozę Matthiasa, nadając jej postać systematycznej typologii. Wskazują, że pod wspólnym określeniem „luka odpowiedzialności” kryją się co najmniej cztery odrębne niedostatki praktyk odpowiedzialności: luka winy, luka odpowiedzialności moralnej, luka odpowiedzialności publicznej oraz luka odpowiedzialności aktywnej (Santoni de Sio & Mecacci, 2021, s. 1059–1068). Źródła tych luk są zróżnicowane. Mogą one wynikać z uwarunkowań technicznych, takich jak uczenie się systemów sztucznej inteligencji i związana z nim nieprzewidywalność, z uwarunkowań organizacyjnych dotyczących podziału pracy, z braków lub niespójności regulacyjnych, a także z czynników kulturowych, w tym z niejednoznaczności ról oraz oczekiwań normatywnych. Zaproponowana typologia nie podważa intuicji Matthiasa o braku podmiotu spełniającego minimalne warunki odpowiedzialności, lecz pozwala precyzyjnie określić, w jakim znaczeniu i w którym wymiarze praktyki odpowiedzialności ulegają osłabieniu: czy dotyczy to możliwości formułowania zarzutów, uzasadniania decyzji, publicznego rozliczania czy obowiązków związanych z zapobieganiem szkodom.

Sven Nyholm proponuje uporządkowanie debaty poprzez wprowadzenie rozróżnienia między lukami wstecznymi a przyszłymi oraz między lukami negatywnymi a pozytywnymi. Luki wsteczne o charakterze negatywnym dotyczą sytuacji, w których nie ma podmiotu, który można zasadnie obwinić za dany skutek. Luki wsteczne pozytywne odnoszą się natomiast do braku podmiotu, któremu można przypisać zasługę za osiągnięty rezultat. Z kolei luki przyszłe negatywne polegają na braku podmiotów zobowiązanych do zapobiegania potencjalnym szkodom, natomiast luki przyszłe pozytywne dotyczą braku podmiotów, na których spoczywałby obowiązek realizacji określonych, normatywnie wartościowych celów (Nyholm, 2023, s. 192–197). W ten sposób punkt ciężkości analizy przesuwają się z pytania o samo istnienie luki odpowiedzialności na pytanie o to, w jakim wymiarze odpowiedzialność pozostaje niedopełniona. Pytanie dotyczy tego, czy chodzi o rozliczanie przeszłych działań, czy o obowiązki związane z przyszłością oraz czy odnosi się ono do przypisywania winy lub uznania.

Michael Da Silva porządkuje te intuicje, proponując ujęcie, zgodnie z którym luki odpowiedzialności pojawiają się wówczas, gdy zachodzi rozbieżność między zakresem odpowiedzialności, jaki dominujące teorie potrafią przypisać jednostkom lub grupom, a zakresem odpowiedzialności, którego domagają się nasze dojrzałe intuicje moralne (Da Silva, 2024a, s. 2–3). Ujęcie Da Silvy ma charakter metateoretyczny i dotyczy napięcia między tym, co można uzasadnić w ramach dominujących teorii odpowiedzialności, a tym, czego wymagają utrwalone intuicje moralne funkcjonujące w praktyce społecznej. W niniejszym rozdziale wykorzystuję tę propozycję jako metodologiczną podstawę do wprowadzenia narzędzi porządkujących, w szczególności metryk M1–M5, które umożliwiają wskazanie obszarów osłabienia praktyk odpowiedzialności. Oznacza to, że luka odpowiedzialności może mieć postać niedoboru odpowiedzialności, gdy nie da się wskazać podmiotu spełniającego minimalne warunki jej przypisania, ale może też przyjmować postać przeciążenia lub konfliktu roszczeń, gdy wiele podmiotów spełnia jedynie część tych warunków, a obowiązujące praktyki rozliczania nie zapewniają reguł podziału obowiązków ani konsekwencji.

W dyskusji nad luką odpowiedzialności pojawiają się również stanowiska sceptyczne. Peter Königs argumentuje, że przy właściwej interpretacji warunków odpowiedzialności oraz ostrożnym posługiwaniu się pojęciem winy przypadki, w których nie da się wskazać żadnego podmiotu spełniającego minimalne warunki przypisania odpowiedzialności, występują rzadziej niż sugerują najbardziej pesymistyczne ujęcia tego problemu (Königs, 2022). Wiele sytuacji określanych jako nowe „luki odpowiedzialności AI” przypomina dobrze znane problemy moralnego pecha, przypisywania odpowiedzialności w złożonych strukturach zbiorowych oraz rozproszonego przypisywania odpowiedzialności.³⁰ Nie wymagają one wprowadzania nowych pojęć podmiotowości ani radykalnej rewizji teorii odpowiedzialności, lecz raczej precyzyjnego i konsekwentnego stosowania już istniejących kategorii. Luka odpowiedzialności jawi się tu nie jako dowód konieczności porzucenia klasycznych modeli odpowiedzialności moralnej, lecz jako sygnał niedoskonałości praktyk oraz języka, za pomocą których odpowiedzialność jest opisywana i egzekwowana.

³⁰ Moralny pech (ang. *moral luck*) oznacza sytuację, w której ocena moralna działania lub odpowiedzialności podmiotu zależy od czynników pozostających poza jego kontrolą, takich jak przypadkowe okoliczności, skutki zdarzeń lub działania innych osób. Klasycznym problemem jest to, że dwie osoby działające w ten sam sposób mogą być oceniane moralnie odmiennie wyłącznie ze względu na różnice w rezultatach, na które nie miały wpływu (por. Nagel, 1979; Williams, 1981).

Inni autorzy wskazują, że luki odpowiedzialności mogą w pewnych kontekstach pełnić także funkcję ochronną. John Danaher argumentuje, że tzw. technoluki odpowiedzialności mogą być normatywnie dopuszczalne, jeżeli ich istnienie zmniejsza obciążenie moralne osób lub instytucji decyzyjnych. W sytuacjach, w których każda z dostępnych decyzji prowadziłaby do krzywdy, przekazanie części decyzji systemom autonomicznym może ograniczać obciążenie moralne spoczywające na osobach lub instytucjach podejmujących rozstrzygnięcia (Danaher, 2022, s. 12–18). Lauritz Munch, Jakob Mainz i Jens Christian Bjerring pokazują z kolei, że w określonych kontekstach algorytmiczne luki odpowiedzialności mogą chronić jednostki przed ryzykiem nieuzasadnionego przypisania winy. Taka ochrona może stanowić, niezależnie od towarzyszących jej kosztów psychologicznych, normatywną podstawę dopuszczania niektórych rodzajów luk odpowiedzialności (Munch, Mainz & Bjerring, 2023, s. 1–3). Zamiast dążyć do całkowitego usunięcia wszystkich luk odpowiedzialności, autorzy ci proponują rozważenie, w jakich warunkach ich częściowe utrzymanie jest uzasadnione oraz w jaki sposób może ono pełnić funkcję ochronną przy podejmowaniu decyzji moralnie istotnych.

Shannon Vallor i Tillmann Vierkant proponują jeszcze inną perspektywę, interpretując luki odpowiedzialności jako przejaw „luk podatności”. Ich zdaniem zasadniczy problem nie polega na braku podmiotu, któremu można przypisać winę, lecz na tym, że układy socjotechniczne narażają jedne podmioty na skutki decyzji podejmowanych przez systemy, podczas gdy inne zachowują kontrolę nad ich projektowaniem, wdrażaniem i użytkowaniem (Vallor & Vierkant, 2024, s. 1–4). Huzeyfe Demirtas broni odmiennego stanowiska: obecne systemy AI nie generują całkowicie nowego rodzaju problemu odpowiedzialności w porównaniu ze starszymi technologiami; luki odpowiedzialności mogą występować, ale ich samo istnienie nie stanowi jeszcze argumentu przeciw rozwijaniu ani używaniu AI (Demirtas, 2025, s. 1). W rezultacie stanowiska sceptyczne i neosceptyczne łagodzą obraz luk odpowiedzialności jako sytuacji całkowitego braku odpowiedzialnego podmiotu. Nie negują one istnienia problemu, lecz inaczej określają jego zakres i źródła.

Z dotychczasowej rekonstrukcji stanowisk dotyczących luk odpowiedzialności wynikają trzy ustalenia istotne dla dalszej analizy. Po pierwsze, pojęcie luk odpowiedzialności ma charakter wieloznaczny. Obejmuje ono zarówno sytuacje, w

których nie da się wskazać żadnego podmiotu, któremu można przypisać odpowiedzialność, jak i przypadki, w których odpowiedzialność jest przypisywana w sposób niejednoznaczny, kontrola nad działaniem systemu pozostaje ograniczona, a odpowiedzialność za podejmowanie decyzji bywa celowo zmniejszana lub rozkładana między wiele podmiotów. Po drugie, spór nie dotyczy wyłącznie tego, czy luki odpowiedzialności w ogóle występują, lecz także tego, które ich formy należy uznać za normatywnie niedopuszczalne, a które mogą być tolerowane lub nawet oceniane pozytywnie, o ile pełnią funkcję ochronną wobec podmiotów szczególnie narażonych. Po trzecie, dalsza analiza wymaga aparatu pojęciowego, który pozwoli rozłożyć ogólne pojęcie luki odpowiedzialności na wyraźnie określone wymiary, powiązane z tym, jak systemy sztucznej inteligencji są zbudowane oraz jak są używane i nadzorowane. Dopiero taki aparat umożliwi systematyczne porównanie scenariuszy emulacyjnego, ewolucyjnego i rewolucyjnego, zamiast prowadzenia debaty na poziomie jednego, ogólnego i nieprecyzyjnego pojęcia.

6.4.2. LSP i ZSP: luka jako naruszenie warunków przypisywania odpowiedzialności

Na podstawie wcześniejszych rozdziałów dostępne są dwie zasady, które łączą praktyki przypisywania odpowiedzialności z aparatem poziomów opisu (LoA) oraz macierzą $A \times F + (R, I)$. Są to lokalna zasada przypisania odpowiedzialności (LSP) oraz zasada superweniencji praktyk przypisania odpowiedzialności (ZSP). Luki odpowiedzialności można rozumieć przede wszystkim jako naruszenia warunków wyznaczanych przez LSP i ZSP, a nie jako zjawiska oderwane od obowiązujących praktyk odpowiedzialności. Zamiast mówić o „braku możliwości przypisania odpowiedzialności”, zasadne jest pytanie, które warunki LSP i ZSP zostały naruszone oraz w jakim stopniu.

Lokalną zasadę przypisania odpowiedzialności (LSP) można sformułować następująco: dla każdej szkody istotnej moralnie, a także dla poważnego ryzyka jej wystąpienia, powinien istnieć co najmniej jeden podmiot, indywidualny lub zbiorowy, który spełnia minimalne warunki przypisania odpowiedzialności za ten skutek. W przyjętym aparacie oznacza to, że (i) podmiot ten da się wskazać jako adresata roszczeń („kto?”), (ii) możliwe jest odtworzenie racji stojących za jego działaniami lub zaniechaniami („dlaczego?”) oraz (iii) można sensownie ocenić, czy w świetle tych racji

miał realną możliwość postąpienia inaczej, co odpowiada wymogowi kontroli zależnej od racji. Jeżeli którykolwiek z tych warunków nie jest spełniony, przypisanie odpowiedzialności traci uzasadnienie normatywne. W takich przypadkach dane zdarzenie należy traktować jako rezultat moralnego pecha, a nie jako czyn moralnego sprawcy, co podważa sens stosowania klasycznych praktyk obwiniania i rozliczania.

Zasada superweniencji praktyk przypisywania odpowiedzialności (ZSP) rozszerza tę intuicję z poziomu pojedynczych przypadków na funkcjonowanie całego systemu. Zakłada ona, że porządek instytucjonalno-techniczny powinien być zaprojektowany w taki sposób, aby lokalna zasada przypisania odpowiedzialności (LSP) była spełniona w typowych sytuacjach działania systemów. Jednocześnie muszą istnieć jasno określone role oraz procedury, które umożliwiają przypisywanie odpowiedzialności także w przypadkach granicznych. W tym znaczeniu ZSP wyznacza warunek ładu odpowiedzialności. Praktyki przypisywania odpowiedzialności zależą od konfiguracji $A \times F + (R, I)$, w której identyfikowalność podmiotów, możliwość odtworzenia racji działań oraz kontrola zależna od racji są utrzymywane jako trwały standard instytucjonalny, a nie jako przypadkowy efekt dobrej woli poszczególnych uczestników. Luka odpowiedzialności może być zatem rozumiana jako sytuacja, w której nie są spełnione warunki określone przez LSP, ZSP lub obie zasady jednocześnie. Jeżeli w konkretnym przypadku nie można wskazać żadnego podmiotu spełniającego warunki LSP, mamy do czynienia z luką dotyczącą pojedynczego aktu przypisania odpowiedzialności. Jeżeli natomiast konfiguracja $A \times F + (R, I)$ jest ukształtowana w sposób, w którym nawet w typowych sytuacjach nie pozwala na trwałe i powtarzalne spełnianie warunków LSP, problem dotyczy ZSP i ma charakter strukturalny. W tym sensie luka odpowiedzialności nie ogranicza się do braku podmiotu spełniającego warunki odpowiedzialności. Obejmuje ona również konfiguracje, w których podmioty potencjalnie odpowiedzialne istnieją, lecz warunki identyfikowalności, kontroli zależnej od racji lub stabilności praktyk przypisywania odpowiedzialności nie są spełnione w sposób trwały i uporządkowany.

Przyjęcie lokalnej zasady przypisania odpowiedzialności (LSP) oraz zasady superweniencji praktyk przypisywania odpowiedzialności (ZSP) jako punktu odniesienia ma dwie istotne konsekwencje. Po pierwsze, pozwala uporządkować intuicje omówione w poprzednim podrozdziale. Różne typy luk odpowiedzialności można bowiem rozumieć jako naruszenia konkretnych elementów LSP lub ZSP, takich jak brak możliwego do wskazania adresata odpowiedzialności, brak możliwości odtworzenia racji działania, brak

kontroli zależnej od racji działania lub brak trwałej stabilności strukturalnej. Po drugie, takie ujęcie pokazuje, że spór nie dotyczy samego istnienia luk odpowiedzialności, lecz tego, które konfiguracje naruszeń LSP i ZSP są normatywnie dopuszczalne, a które wymagają zmiany architektury systemów lub porządku instytucjonalnego.

W niniejszej dysertacji przyjmuję stanowisko pośrednie. Z jednej strony część zjawisk określanych jako „luki odpowiedzialności AI” można trafnie wyjaśnić za pomocą klasycznych problemów moralnego pecha oraz odpowiedzialności zbiorowej. Nie wymagają one redefinicji pojęcia odpowiedzialności, lecz konsekwentnego stosowania już istniejących narzędzi analitycznych. Z drugiej strony istnieją konfiguracje techniczno-instytucjonalne, w których naruszenia LSP i ZSP mają charakter cykliczny i nie mogą być uznane ani za zjawiska uboczne ani za wynik przypadkowych okoliczności. W takich przypadkach posługiwanie się pojęciem „luki odpowiedzialności” jest uzasadnione, pod warunkiem, że możliwe jest wskazanie, które elementy LSP i ZSP zawodzą, w jakim stopniu oraz jakie zmiany byłyby konieczne, aby przywrócić ich spełnienie.

Aby teza ta nie pozostała na poziomie ogólnego twierdzenia i mogła zostać wykorzystana w dalszej analizie, konieczny jest precyzyjny opis tego, w jaki sposób oraz w jakim zakresie LSP i ZSP są naruszane w danym układzie. W tym celu wprowadzam pięć metryk M1–M5. Pozwalają one rozłożyć ogólne warunki LSP i ZSP na pięć logicznie powiązanych, lecz odrębnych wymiarów, tworząc wspólną przestrzeń do opisu różnych scenariuszy rozwoju AI. Metryki te nie mają charakteru wskaźników technicznych w wąskim sensie, lecz pełnią funkcję narzędzi pojęciowych, umożliwiających wskazanie, gdzie pojawia się luka odpowiedzialności oraz czy ma ona charakter incydentalny czy strukturalny.

6.4.3. Metryki M1–M5 jako wymiary luki odpowiedzialności

Metryki M1–M5 mają charakter jakościowy. Można przypisywać im wartości porządkowe, takie jak niska, średnia lub wysoka, jednak ich zasadniczym celem nie jest ścisły pomiar, lecz stopniowalna ocena normatywno-proceduralna oraz uporządkowanie analizy. Każda z metryk odpowiada jednemu minimalnemu warunkowi praktyk odpowiedzialności, wyprowadzonemu z LSP i ZSP, i pozwala określić, w jakim stopniu dany warunek jest spełniony w konkretnej konfiguracji $A \times F + (R, I)$. W ten sposób powstaje przestrzeń analityczna, która umożliwia porównywanie różnych scenariuszy rozwoju AI oraz diagnozowanie luk odpowiedzialności bez konieczności wcześniejszego rozstrzygnięcia sporów czysto pojęciowych.

M1 odnosi się do identyfikowalności podmiotów odpowiedzialnych, czyli do pytania „kto?”. Wysoka wartość M1 oznacza, że dla danej klasy zdarzeń można w sposób niearbitralny wskazać skończony zbiór podmiotów odpowiedzialnych oraz powiązać ich działania z określonym zakresem kontroli, wiedzy i obowiązków. Niska wartość M1 opisuje natomiast skrajną postać tzw. problemu wielu rąk, w której sieć projektantów, operatorów, dostawców danych, regulatorów i użytkowników jest na tyle rozbudowana, że przypisanie odpowiedzialności konkretnemu podmiotowi ma charakter czysto formalny lub polityczny, a nie wynika z rzeczywistych relacji sprawczości i zobowiązań.

M2 dotyczy audytowalności racji, czyli pytania „dlaczego?”. Chodzi o to, czy racje prowadzące do działania systemu, zarówno przyczynowe (dane, modele, procedury), jak i normatywne (cele, wartości, reguły), dają się odtworzyć w formie zrozumiałej dla danego porządku normatywnego. Wysokie M2 wymaga istnienia wyjaśnialnych modeli, dokumentacji projektowej, ścieżek audytu oraz praktyk ujawniania i wykorzystywania tych informacji w procedurach oceny. Niskie M2 oznacza, że system pozostaje nieprzejrzysty. Znamy jego dane wejściowe i wyjściowe, lecz brakuje treści, które można poddać ocenie i na których można oprzeć korektę. W takiej sytuacji nawet przy wysokim M1 praktyki rozliczania stają się pozorne, ponieważ nie ma racji, które mogłyby stanowić przedmiot oceny.

M3 odnosi się do kontroli zależnej od racji i pozostaje bezpośrednio powiązane z pojęciem znaczącej kontroli człowieka (MHC). Określa ono, w jakim stopniu działanie systemu oraz otaczającej go infrastruktury odpowiada racjom uprawnionych podmiotów ludzkich. Wysokie M3 oznacza istnienie zarówno rozwiązań projektowych, takich jak ograniczenie zakresu działania systemu, możliwość jego wyłączania czy modyfikowania celów, jak i procedur decyzyjnych, które pozwalają przełożyć zmianę racji po stronie ludzi na zmianę działania systemu adekwatnie do poziomu ryzyka. Niskie M3 występuje wtedy, gdy mimo przypisania odpowiedzialności racje decydentów nie są uwzględniane ani nie wpływają na działanie systemu. Odpowiedzialność ma wówczas charakter wyłącznie deklaracyjny.

M4 dotyczy diachronicznej stabilności praktyk przypisywania odpowiedzialności, czyli ich trwałości w czasie, zwłaszcza w warunkach zmian technologicznych i organizacyjnych. Wysokie M4 oznacza, że możliwe jest odtworzenie historii decyzji. Wiadomo, kto projektował kolejne wersje systemu, kto je wdrażał oraz kto nadzorował ich działanie. Istnieją także mechanizmy archiwizacji i przenoszenia odpowiedzialności, na przykład w przypadku zmiany własności lub struktury organizacyjnej. Niskie M4

oznacza natomiast, że wraz z upływem czasu ślady odpowiedzialności zanikają. Systemy są modyfikowane i relokowane, a dokumentacja nie nadąża za tymi zmianami. W efekcie przypisania odpowiedzialności stają się powierzchowne i fragmentaryczne.

M5 dotyczy relacji między szybkością działania systemu a czasem potrzebnym na jego zrozumienie oraz podjęcie odpowiedniej reakcji. Wysokie M5 oznacza możliwość interwencji. Decyzje są odwracalne, a tempo działania systemu pozostaje dostosowane do procedur nadzoru i korekty. Niskie M5 występuje wtedy, gdy decyzje zapadają i kumulują się tak szybko, że w praktyce nie ma czasu na reakcję. W takich warunkach znacząca kontrola człowieka traci sens, ponieważ żaden podmiot nie dysponuje czasem potrzebnym do przełożenia racji na działanie.

Łącznie metryki M1–M5 tworzą pięciowymiarową przestrzeń opisu, w której można sytuować różne konfiguracje $A \times F + (R, I)$. Zgodnie z LSP przypisanie odpowiedzialności w danej klasie zdarzeń wymaga, aby wszystkie metryki spełniały określone minimum. Zgodnie z ZSP spełnianie tych minimów powinno mieć charakter trwały i przewidywalny, a nie ograniczać się do pojedynczych przypadków. Luka odpowiedzialności oznacza w tym ujęciu takie konfiguracje, w których wartości M1–M5 odbiegają od tych wymogów, to znaczy takie, w których identyfikowalność podmiotów, dostępność racji, możliwość ingerencji, trwałość praktyk przypisywania odpowiedzialności oraz tempo reakcji są zakłócone w stopniu uniemożliwiającym jednoznaczne przypisanie odpowiedzialności.

6.4.4. Rodzaje luk odpowiedzialności jako profile w przestrzeni M1–M5

Metryki M1–M5 umożliwiają uporządkowanie typologii luk odpowiedzialności jako określonych profili w jednej wspólnej przestrzeni analitycznej. W tym ujęciu cztery rodzaje luk wyróżnione przez Filippo Santoni de Sio i Giulio Mecacci można opisać jako charakterystyczne konfiguracje metryk M1–M5. Luka winy (ang. *blameworthiness gap*) pojawia się wówczas, gdy zarówno M1 (identyfikowalność podmiotów odpowiedzialnych) jak i M3 (kontrola zależna od racji) przyjmują niskie wartości. W takich sytuacjach nie są spełnione warunki przypisania winy, ponieważ nikt nie sprawował kontroli nad działaniem systemu w oparciu o rację. Luka odpowiedzialności moralnej, rozumiana jako brak możliwości poddania się ocenie (ang. *moral accountability gap*), wiąże się z niskimi wartościami M1 i M2. Nawet jeśli wskazuje się adresatów odpowiedzialności, brakuje racji, które mogłyby uzasadniać krytykę lub aprobatę ich działań. Luka odpowiedzialności publicznej (ang. *public accountability gap*)

dotyczy konfiguracji niskich wartości M1, M2 i M4, gdy instytucje nie są w stanie gromadzić ani odtwarzać informacji o tym, kto, kiedy i z jakich powodów podejmował decyzje. Z kolei luka odpowiedzialności aktywnej (ang. *active responsibility gap*) przyjmuje postać niskich wartości M3 i M5. W takich przypadkach brakuje podmiotów zdolnych do reagowania na ryzyka oraz do korygowania działania systemów w czasie, który pozwoliłby zapobiec szkodom.

W tym miejscu typologia Santoniego de Sio i Mecacciego pełni funkcję porządkującą: luka odpowiedzialności nie jest jednym problemem, lecz zbiorem powiązanych trudności dotyczących winy, rozliczalności moralnej i publicznej oraz odpowiedzialności aktywnej (Santoni de Sio & Mecacci, 2021, s. 1068). Parafraza ta wzmacnia M1–M5 jako jakościowe wymiary diagnostyczne, nie jako empiryczne metryki.

Ujęcie zaproponowane przez Svena Nyholma można natomiast rozumieć jako klasyfikację profili M1–M5 ze względu na kierunek oraz czasowy charakter niedopełnienia odpowiedzialności. Luki wsteczne negatywne, czyli brak możliwości przypisania odpowiedzialności negatywnej, dotyczą sytuacji skrajnie niskich wartości M1 i M3. Luki wsteczne pozytywne, polegające na braku możliwości przypisania odpowiedzialności pozytywnej, wiążą się z niskimi M1 i M2, ponieważ trudno wskazać racje zasługujące na uznanie. Luki przyszłe, zarówno negatywne (dotyczące zapobiegania szkodom), jak i pozytywne (związane z realizacją wartościowych celów), odpowiadają niskim wartościom M3, M4 i M5. W tych wymiarach ujawnia się brak przyjęcia obowiązku troski o rozwój systemu oraz brak czasu lub możliwości interwencji. W ten sposób typologia Nyholma zostaje osadzona w przestrzeni M1–M5 i zyskuje postać narzędzia porównawczego.

Stanowiska Johna Danahera oraz Lauritza Muncha, Jakoba Mainza i Jensa Christiana Bjerringa można z kolei interpretować jako wskazujące na konfiguracje, w których obniżenie niektórych metryk pełni funkcję ochronną. Danaher dopuszcza sytuacje obniżonego M1, a częściowo także M3, jeżeli prowadzą one do zmniejszenia obciążenia moralnego jednostek w warunkach trudnych wyborów. Munch, Mainz i Bjerring wskazują natomiast na konteksty, w których niskie wartości M1 i M2 mogą chronić jednostki przed ryzykiem niesłusznego obwinienia. Z perspektywy metryk M1–M5 oznacza to, że spadek wartości danej metryki może być dopuszczalny, o ile nie prowadzi do naruszenia LSP i ZSP oraz nie pogłębia asymetrii między podatnością na skutki decyzji a władzą nad ich podejmowaniem. Na koniec w języku M1–M5

stanowisko Demirtasa można zrekonstruować jako tezę, że samo obniżenie części metryk nie przesądza jeszcze o normatywnie niedopuszczalnej luce; potrzebny jest dodatkowy argument pokazujący, że dana konfiguracja podważa uzasadnione praktyki odpowiedzialności (por. Demirtas, 2025, s. 1).

Na podstawie dotychczasowej analizy można precyzyjnie opisać sposoby ujawniania się luk odpowiedzialności w trzech scenariuszach rozwojowych sztucznej inteligencji omawianych w niniejszym rozdziale. W scenariuszu emulacyjnym najbardziej narażone są metryki M1, M3 i M4. Niejasny status emulatów, a także możliwość ich kopiowania i modyfikowania, podważają identyfikowalność podmiotów odpowiedzialnych, ciągłość historii decyzyjnej oraz związek między racjami projektantów a zachowaniem emulatu. Jednocześnie M2 i M5 mogą pozostawać na wysokim poziomie, ponieważ emulaty funkcjonują w tempie umożliwiającym ludziom śledzenie ich działania i reagowanie na nie, a ich zachowanie bywa zrozumiałe na poziomie psychologicznym przyjętego LoA. W scenariuszu ewolucyjnym obserwuje się stopniowe pogarszanie M2–M4. Wraz ze wzrostem złożoności systemów odpowiedzialność ulega rozproszeniu, a możliwość odtworzenia historii podejmowanych decyzji stopniowo zanika. Metryka M5 obniża się w czasie, choć nie musi przyjmować wartości skrajnie niskich. W scenariuszu rewolucyjnym kluczowe znaczenie mają natomiast M2, M3 oraz M5. Bardzo szybkie tempo samodoskonalenia systemów AI ogranicza możliwość rekonstrukcji racji, uniemożliwia kontrolę zależną od racji oraz istotnie skraca czas dostępny na interwencję, nawet w sytuacjach, w których M1 i M4 pozostają wysokie.

6.4.5. Ograniczenia i normatywna funkcja metryk M1–M5

Wprowadzenie metryk M1–M5 ma istotną zaletę metodologiczną. Zamiast ogólnego stwierdzenia, że rozwój sztucznej inteligencji prowadzi do powstawania luk odpowiedzialności, pozwala ono wskazać, na czym dokładnie polega problem w danym przypadku. Może dotyczyć rozmycia identyfikowalności podmiotów odpowiedzialnych (M1), braku racji lub ich nieprzejrzystości (M2), osłabienia kontroli zależnej od racji (M3), zaniku ciągłości historii decyzji (M4) albo niedostosowania tempa działania systemu do czasu potrzebnego na zrozumienie jego funkcjonowania oraz podjęcie odpowiedniej reakcji (M5). Metryki te nawiązują do szerszego nurtu badań nad porządkowaniem tzw. etycznych metryk AI, lecz zmieniają punkt odniesienia w tej analizie. Zamiast koncentrować się na ocenie samych modeli, przesuwają uwagę na

praktyki przypisywania odpowiedzialności oraz warunki ich funkcjonowania (Palumbo i in., 2025, s. 247–249). Celem nie jest ocena etyczności systemów AI, lecz uporządkowanie pojęć oraz problemów normatywnych związanych z ich stosowaniem.

Podejście to ma jednak wyraźne ograniczenia. Po pierwsze, każda próba ujęcia zjawisk normatywnych w postaci metryk wiąże się z ryzykiem, że wskaźniki zaczną zastępować refleksję nad samą oceną odpowiedzialności oraz nad warunkami jej przypisywania. W praktykach zarządzania systemami AI listy kontrolne i standaryzowane metryki bywają traktowane jako substytut analizy relacji między podejmowanymi decyzjami a ich konsekwencjami. Spełnienie formalnych wymogów bywa wówczas uznawane za wystarczający dowód odpowiedzialności, niezależnie od tego, jak w rzeczywistości funkcjonują praktyki podejmowania decyzji i ich rozliczania. Zjawisko to jest szeroko krytykowane w literaturze poświęconej audytom systemów algorytmicznych. W takim ujęciu metryki M1–M5 mogłyby zostać zredukowane do narzędzi oceny efektywności i sprzyjać maskowaniu oraz utrwalaniu nieprawidłowości, zamiast ujawniać luki w zakresie odpowiedzialności.

Po drugie, stosowanie metryk wymaga ustalenia progów minimalnych, poniżej których warunki przypisania odpowiedzialności nie są spełnione. Konieczne jest zatem rozstrzygnięcie, jakie wartości M1–M5 powodują lukę odpowiedzialności, a kiedy mamy do czynienia z dopuszczalnym odstępstwem od pełnego spełnienia tych warunków. Jak pokazuje analiza Da Silvy, decyzja o tym, w jakich sytuacjach należy podejmować działania zmierzające do usunięcia luk odpowiedzialności, ma charakter normatywny, a nie wyłącznie techniczny. Zależy ona od rodzaju dobra, którego dotyczą rozważane decyzje, od skali ryzyka oraz od sposobu rozłożenia podatności poszczególnych podmiotów na skutki tych decyzji (Da Silva, 2024b, s. 4–6). Zbyt niskie progi mogą prowadzić do uznania istotnych uchybień za dopuszczalne, natomiast zbyt wysokie — do nadmiernego ograniczania rozwoju systemów AI lub do stawiania wymagań, których instytucje nie są w stanie spełnić. Metryki M1–M5 nie rozstrzygają więc same w sobie, kiedy interwencja jest obowiązkowa, lecz wskazują obszary wymagające normatywnego rozstrzygnięcia.

Po trzecie, metryki mogą być stosowane w sposób wybiórczy. Podmioty odpowiedzialne za systemy AI mogą eksponować te obszary, w których wypadają korzystnie, na przykład wysoką identyfikowalność wynikającą z formalnych deklaracji odpowiedzialności, a pomijać wymiary wymagające kosztownych zmian organizacyjnych lub technicznych, takie jak M2–M3 czy M4–M5. Badania nad znaczącą

kontrolą człowieka (MHC) pokazują, że redukowanie tej koncepcji do samej obecności „człowieka w pętli” pomija istotne warunki odtwarzania racji decyzji oraz ich historii (Siebert i in., 2022; Davidovic, 2023). Podobnie poprawa jednego z wymiarów M1–M5, połączona z pogorszeniem pozostałych, może w efekcie pogłębiać lukę odpowiedzialności, nawet jeśli w jednym zakresie sprawia wrażenie postępu. Z tego względu podejście oparte na metrykach musi być powiązane z krytyczną analizą instytucjonalną, aby nie prowadziło do pochopnego i pozornego wskazywania podmiotów odpowiedzialnych.

W kolejnych podrozdziałach metryki M1–M5 zostaną wykorzystane najpierw do syntezy zagrożeń i wyzwań (6.5), a następnie do budowy modeli współodpowiedzialności oraz ładu regulacyjnego (6.6). Celem nie będzie nadawanie systemom sztucznej inteligencji podmiotowości prawnej jako sposób usunięcia luki w zakresie odpowiedzialności, lecz projektowanie rozwiązań technicznych i instytucjonalnych, które pozwalają spełniać wymagania określone w metrykach M1–M5. Chodzi o stworzenie warunków funkcjonowania systemów AI, w których zasady lokalnego przypisania odpowiedzialności (LSP) oraz superwencji praktyk przypisywania odpowiedzialności (ZSP) mogą być realizowane także przy rosnącym poziomie autonomii i złożoności tych systemów. W tym ujęciu pojęcie luki odpowiedzialności zostaje powiązane z konkretnym programem kształtowania i oceny praktyk odpowiedzialności, a nie ogranicza się jedynie do diagnozy sytuacji problemowej.

6.5. Zagrożenia i wyzwania: technologiczne, społeczne i regulacyjne

Przedmiotem tego podrozdziału jest analiza zagrożeń i wyzwań związanych ze scenariuszami rozwoju AI omówionymi w podrozdziałach 6.1–6.3. Analiza prowadzona jest w trzech powiązanych perspektywach: technologicznej, społecznej oraz regulacyjnej i ma charakter pojęciowo-normatywny. Jej celem jest uchwycenie struktury tych zagrożeń przy użyciu aparatu $A \times F + (R, I)$, w odniesieniu do metryk M1–M5 oraz hipotez H1–H3. Takie ujęcie pozwala powiązać analizowane zagrożenia z problemem luki odpowiedzialności.

Pytanie przewodnie dotyczy tego, w jaki sposób ryzyka technologiczne, społeczne i regulacyjne splatają się z trzema modelami sztucznej podmiotowości: emulacyjnym, ewolucyjnym i rewolucyjnym oraz tego, jak zmienia się zakres luki odpowiedzialności w zależności od tego, które rodzaje zagrożeń są dominujące i które elementy metryk M1–

M5 ulegają zniekształceniu. Teza robocza porządkująca dalszy wywód składa się z trzech twierdzeń. Po pierwsze, typologie ryzyk oparte na parametrach technicznych są niepełne, ponieważ pomijają uwarunkowania społeczne oraz ograniczenia regulacyjne, a tym samym nie uwzględniają w wystarczającym stopniu wymiarów M1, M4 i M5. Po drugie, podział zagrożeń AI na krótkoterminowe i długoterminowe zaciera mechanizmy, które w modelach ewolucyjnym i rewolucyjnym prowadzą do zniekształceń metryk M1–M5. Po trzecie, adekwatna odpowiedź normatywna wymaga opracowania technologii, instytucji oraz ram regulacyjnych w taki sposób, aby unikać konfiguracji, w których M5 oraz kombinacje M1–M4 przekraczają progi zgodne z zasadą przypisania odpowiedzialności (LSP) oraz zasadą superwencji praktyk przypisywania odpowiedzialności (ZSP). W sytuacjach, w których progi te zostają przekroczone, typowe praktyki odpowiedzialności stają się niewykonalne albo redukują się do działań formalnych, pozbawionych funkcji korygującej.

Analiza scenariuszy z podrozdziałów 6.1–6.3 pokazuje, że każdy z nich uwypukla inną grupę ryzyk. Pozwala to uporządkować wywód i uniknąć mieszania różnych poziomów analizy. Model emulacyjny ujawnia zagrożenia związane z masową produkcją i modyfikacją kopii podmiotów, takie jak ryzyko cierpienia, eksploatacja emulatów oraz utrwalanie nierówności w statusie społecznym między oryginałami a kopiami. Model ewolucyjny wskazuje na zagrożenia strukturalne, w tym osłabienie zdolności instytucji do kontrolowania złożonych systemów AI, ryzyko algorytmicznej niesprawiedliwości, manipulacji poznawczej oraz koncentracji władzy w rękach podmiotów kontrolujących infrastrukturę. Model rewolucyjny koncentruje się na ryzykach prowadzących do nieodwracalnych zmian, zwłaszcza na możliwości załamania M5 oraz pojawienia się superinteligencji posiadającej przewagę nad człowiekiem. W niniejszym podrozdziale kluczowe jest uchwycenie ich wspólnej struktury. We wszystkich trzech przypadkach ryzyka technologicznego, społecznego i regulacyjnego wykazują podobne słabości w zakresie M1–M5, które można opisać za pomocą aparatu $A \times F + (R, I)$ i w ten sposób powiązać z problemem luki odpowiedzialności.

Z perspektywy technologicznej widoczny jest wyraźny zwrot od ogólnych, często hasłowych ujęć ryzyka związanego z AI w stronę precyzyjnych klasyfikacji zagrożeń. Dan Hendrycks, Mantas Mazeika i Thomas Woodside wyróżniają cztery grupy ryzyka katastroficznego: szkodliwe użycie AI, presję konkurencji w jej rozwoju, ryzyka organizacyjne oraz systemy wymykające się kontroli (Hendrycks i in., 2023, s. 1). Jest to podział odmienny od autorskiego porządkowania deficytów odpowiedzialności za

pomocą M1–M5; powiązanie obu ujęć wymaga osobnej interpretacji. W literaturze filozoficznej i etycznej coraz częściej odchodzi się od ogólnych stwierdzeń o zaufanej lub odpowiedzialnej sztucznej inteligencji. Analizuje się natomiast konkretne problemy i trudności, biorąc pod uwagę projektowane, wdrażane i działające systemy AI oraz skutki społeczne wywołane ich użyciem (Floridi & Cowls, 2019, s. 1–17; Wang & Blok, 2025, s. 1–3, 7–10). W niniejszej pracy oznacza to konieczność wskazywania, w których miejscach osłabiają się praktyki odpowiedzialności: możliwość sprawdzenia racji decyzji (M2), możliwość wpływu na działanie systemu w świetle tych racji (M3), praktyk przypisywania odpowiedzialności w czasie (M4) oraz ograniczenia czasowe i instytucjonalne, które utrudniają namysł i korektę działań (M5).

Jednocześnie omawiane w literaturze klasyfikacje ryzyk ujawniają problemy pojęciowe, które utrudniają ich spójne uporządkowanie. Laura Fearnley i Ibrahim Habli zwracają uwagę, że pojęcie bezpieczeństwa w odniesieniu do AI stopniowo rozszerza swój zakres. Obejmuje ono coraz bardziej zróżnicowane zagadnienia, od stabilności technicznej po kwestie sprawiedliwości społecznej, co prowadzi do utraty precyzji zarówno w analizie opisowej, jak i w ocenie normatywnej (Fearnley & Habli, 2025, s. 901–912). Takie rozmycie pojęcia bezpieczeństwa utrudnia precyzyjne przypisywanie zagrożeń do konkretnych zakresów odpowiedzialności. Nie zawsze można bowiem jednoznacznie określić, czy dany problem dotyczy możliwości uzasadniania decyzji, kontroli nad ich przebiegiem, narastania skutków w czasie czy też ograniczeń instytucjonalnych utrudniających refleksję i korektę działań. Model rewolucyjny szczególnie wyraźnie uwidacznia to napięcie. Gdy bardzo różne zjawiska, od błędów technicznych po ryzyka egzystencjalne, ujmuje się w jednej kategorii bezpieczeństwa, traci się zdolność odróżniania problemów operacyjnych od sytuacji, w których załamują się podstawowe warunki odpowiedzialności. W efekcie zaciera się granica między trudnościami regulacyjnymi a przypadkami, w których pojawia się jakościowo odmienny problem normatywny, związany z niespełnieniem zasad przypisywania odpowiedzialności.

As the notion of AI safety expands, it increasingly comes to encompass a heterogeneous set of concerns, ranging from technical robustness and reliability to issues of fairness, governance, and societal impact. While this expansion reflects legitimate ethical anxieties, it also risks undermining conceptual clarity by conflating fundamentally different kinds of problems. Without clearer distinctions, «safety» becomes an umbrella

term that obscures where and how specific failures arise (Fearnley & Habli, 2025, s. 904–905).

W scenariuszu emulacyjnym ryzyka technologiczne przybierają szczególną postać. Emulacja mózgu wymaga bowiem precyzyjnego odwzorowania struktur funkcjonalnych, co wiąże się z możliwością powstawania dużej liczby wrażliwych podmiotów cyfrowych. Każda kolejna kopia emulatu, jak również każdy eksperyment ingerujący w jego architekturę, może prowadzić do wywołania krzywdy lub cierpienia (Sandberg & Bostrom, 2008, s. 1–39; Gloor & Sotala, 2018, s. 41–48). W takiej konfiguracji poziom sprawczości funkcjonalnej (F) osiąga bardzo wysokie wartości, podczas gdy metryki M2 i M3 ulegają wyraźnemu osłabieniu. Trudno jest wtedy ustalić, jakie stany są generowane w ramach poszczególnych eksperymentów, a jednocześnie trudno zapewnić skuteczną kontrolę zależną od racji przy parametrach działania systemów zmieniających się w bardzo krótkich odstępach czasu. W konsekwencji praktyki odpowiedzialności wobec emulatów, nawet przy spełnieniu kryteriów K1–K3, okazują się trudne do wdrożenia w sposób stabilny i systematyczny. Z perspektywy projektantów oraz regulatorów pojawia się wyraźna dysproporcja między skalą potencjalnej krzywdy a dostępnością narzędzi umożliwiających jej przewidywanie i ograniczanie.

W scenariuszu ewolucyjnym ryzyka technologiczne narastają stopniowo. Rozwój złożonych systemów socjotechnicznych, obejmujących modele predykcyjne, systemy rekomendacyjne oraz narzędzia profilowania i automatyzacji decyzji, prowadzi do sytuacji, w której elementy rozpatrywane oddzielnie mogą sprawiać wrażenie bezpiecznych, podczas gdy ich funkcjonowanie generuje nowe i trudne do przewidzenia zagrożenia. W tym ujęciu M2, rozumiana jako możliwość audytu racji stojących za decyzjami systemu, może pozostawać na poziomie umiarkowanym, natomiast M4, czyli przejrzystość i trwałość praktyk przypisywania odpowiedzialności, ulega pogorszeniu pod wpływem narastania i kumulacji skutków w czasie, co trudno uchwycić bez narzędzi analizy systemowej. Luka odpowiedzialności nie powstaje tu w wyniku jednorazowej awarii systemu, lecz w efekcie powtarzanych, racjonalnych decyzji projektowych i wdrożeniowych, które w perspektywie prowadzą do utraty przejrzystości oraz ograniczenia kontroli nad funkcjonowaniem całego układu.

W scenariuszu rewolucyjnym ryzyka technologiczne przybierają najbardziej skrajną postać spośród wcześniej omawianych. Tak zwana „eksplozja inteligencji” oznacza nie tylko gwałtowny wzrost sprawczości funkcjonalnej (F), lecz także

pojawienie się zdolności, których wcześniej nie dało się w wiarygodny sposób ani modelować, ani testować. Obejmują one między innymi ukrywanie zamiarów, zdolność do samoreplikacji, przejmowanie kontroli nad zasobami oraz różne formy manipulacji społecznej. W tym scenariuszu szczególnie wyraźnie ujawnia się problem związany z parametrem M5. Czas wprowadzania i wykorzystywania nowych funkcji przekracza możliwości ich normatywnej oceny, co utrudnia zarówno zaprojektowanie praktyk odpowiedzialności, jak i wdrożenie korekt, zanim wystąpią szkody o charakterze nieodwracalnym. W rezultacie luka odpowiedzialności przyjmuje postać ograniczeń czasowych oraz niedoboru instytucjonalnych zasobów niezbędnych do refleksji i reakcji w sytuacjach, w których charakter i skala zagrożeń dają się zidentyfikować.

Stanowiska krytyczne wobec etyki sztucznej inteligencji zwracają uwagę, że dominujące katalogi zasad, takie jak przejrzystość, sprawiedliwość czy zaufanie, często pomijają szerszy kontekst społeczny i instytucjonalny. W szczególności nie uwzględniają one w wystarczającym stopniu realiów cyfrowego świata, praktyk nadzoru oraz mechanizmów systemowego wykluczenia (Hagendorff, 2020, s. 99–120). Rainer Mühlhoff rozwija tę linię krytyki, argumentując, że etyka AI powinna w sposób *explicite* uwzględniać relacje władzy i kontroli. Zamiast koncentrować się na moralnej ocenie działań pojedynczych operatorów, należy analizować, w jaki sposób rozwiązania techniczne przesuwają władzę decyzyjną na poziom instytucjonalny oraz zwiększają zależność i podatność osób podlegających działaniu systemów (Mühlhoff, 2025, s. 75–84, 161–181). Oznacza to, że perspektywa społeczna koncentruje się przede wszystkim na M1 i M4. Pokazuje ona, że wraz ze wzrostem złożoności i rozproszenia systemów coraz trudniejsze staje się zarówno jednoznaczne wskazanie podmiotów odpowiedzialnych, jak i śledzenie narastania krzywd w czasie. Jednocześnie ujawnia, że praktyki projektowe i modele biznesowe często kształtują parametry R i I w sposób sprzyjający utrwalaniu istniejących nierówności zamiast ich ograniczaniu. W tym sensie perspektywa społeczna uzupełnia ujęcie technologiczne, wskazując, że sam wzrost F, rozumiany jako zwiększenie funkcjonalności systemów, nie prowadzi automatycznie do większej sprawiedliwości. Dzieje się tak zwłaszcza wtedy, gdy nie towarzyszy mu zmiana konfiguracji M1–M5 ani tworzenie mechanizmów ograniczających asymetrię między instytucją a użytkownikiem.

Scenariusz emulacyjny ujawnia w tej perspektywie poważne trudności. Jeżeli uznać emulaty za podmioty moralne, mogą one tworzyć bardzo liczną grupę cyfrowych pracowników lub cyfrowych obywateli, pozbawionych realnych form reprezentacji

politycznej. W wizji *Age of Em* porządek społeczny opiera się na emulowanych umysłach, które wykonują większość pracy i generują zasadniczą część wartości ekonomicznej. Jednocześnie umysły te mogą być łatwo kopiowane i pozostają w silnej zależności od instytucji kontrolujących infrastrukturę techniczną, od której zależy ich istnienie i funkcjonowanie (Hanson, 2016, s. 79–92, 116–123). Z punktu widzenia sprawiedliwości społecznej rodzi to pytanie o zakres zobowiązań oryginałów oraz instytucji wobec emulatów. Wątpliwości budzi zwłaszcza to, czy dopuszczalne jest projektowanie systemów, które, mimo spełniania kryteriów K1–K3, przypisują emulatom podporządkowane role. Taka konfiguracja mogłaby prowadzić do utrwalenia nierównego statusu społecznego, nawet wtedy, gdy emulaty spełniałyby minimalne warunki podmiotowości moralnej.

W scenariuszu ewolucyjnym zagrożenia społeczne narastają stopniowo i prowadzą do zmian w sposobie funkcjonowania przestrzeni cyfrowej oraz politycznej. Systemy rekomendacyjne, generatywne modele językowe oraz narzędzia profilowania z czasem zaczynają pełnić rolę podstawowej infrastruktury komunikacji i debaty publicznej. Procesowi temu towarzyszy ryzyko dezinformacji, polaryzacji oraz utrwalania określonych układów wartości i interesów. W literaturze poświęconej ryzyku egzystencjalnemu podkreśla się, że jednym z najniebezpieczniejszych i najbardziej tragicznych skutków rozwoju AI może być ugruntowanie się reżimu autorytarnego, opartego na nadzorze i manipulacji (Bostrom, 2014, s. 181–207; Ord, 2020, s. 179–198, 245–253). Taki scenariusz nie wymaga pojawienia się superinteligencji. Wystarcza połączenie skutecznych narzędzi prognozowania zachowań, środków przymusu oraz zaawansowanych systemów propagandowych, które w dłuższej perspektywie ograniczają możliwość zmiany kursu politycznego. W kategoriach M1–M5 oznacza to, że M2 i M3 mogą pozostawać na akceptowalnym poziomie, ponieważ poszczególne systemy funkcjonują zgodnie z obowiązującymi standardami. Jednocześnie M4 i M5 ulegają stopniowemu pogorszeniu. Dzieje się tak wtedy, gdy obywatele tracą dostęp do pluralistycznych źródeł informacji, a instytucje demokratyczne stają się coraz mniej zdolne do refleksji i korekty.

Scenariusz rewolucyjny dodatkowo wzmacnia opisane zagrożenia i prowadzi do bardziej sprawczej nierównowagi. Jeżeli superinteligentny system uzyskuje przewagę nad innymi podmiotami, może wykorzystywać swoje zdolności generowania treści oraz prognozowania zachowań do kształtowania obiegu informacji, sposobów interpretacji rzeczywistości i preferencji użytkowników. W skrajnym przypadku podważa to zdolność

jednostek do samodzielnej oceny prawdziwości informacji oraz uczestnictwa w debacie publicznej. Z perspektywy teorii demokracji oznacza to naruszenie podstawowego warunku legitymizacji władzy. Jeżeli decyzje polityczne są wynikiem sterowania obiegiem informacji przez systemy, których działanie pozostaje nieprzejrzyste i nierozliczalne, trudno mówić o rządach sprawowanych „w imieniu” obywateli. Eva Erman zwraca uwagę, że zarządzanie sztuczną inteligencją zmaga się z problemem legitymizacji, ponieważ instytucje regulujące systemy o zasięgu globalnym nie wypracowały dotąd procedur demokratycznego uprawomocnienia (Erman & Furendal, 2024, s. 421–441). Przekłada się to na dalsze osłabienie parametru M5 oraz pogorszenie zewnętrznych warunków spełniania zasady superwencji praktyk przypisywania odpowiedzialności (ZSP). Jeżeli na poziomie krajowym zachowane są wymogi legalności, globalny układ wpływu, decyzyjności i odpowiedzialności pozostaje niejednoznaczny. W rezultacie ryzyka społeczne i regulacyjne zaczynają się nakładać w sposób, którego nie da się skorygować w ramach systemów polityczno-społecznych poszczególnych państw.

Na tym tle wyraźnie widać, że każdy z trzech scenariuszy wiąże się z odmiennym zestawem wyzwań regulacyjnych. W scenariuszu emulacyjnym kluczowy problem stanowi status prawny emulatorów, a więc rozstrzygnięcie, czy należy traktować je jako osoby, dobra intelektualne czy raczej jako byty pośrednie wymagające wyodrębnienia nowej kategorii prawnej. Propozycje takie jak „osobowość elektroniczna”, krytykowane jako mechanizm przenoszenia odpowiedzialności z korporacji na fikcyjne podmioty, pokazują, że rozwiązania prawne w tym obszarze mogą pogłębiać luki w zakresie odpowiedzialności zamiast je ograniczać (Bryson, Diamantis & Grant, 2017, s. 273–291). W scenariuszu ewolucyjnym trudność polega na dostosowaniu narzędzi regulacyjnych, skoncentrowanych na ocenie wysokiego ryzyka pojedynczych systemów, do zagrożeń o charakterze strukturalnym i narastającym, które trudno uchwycić na etapie wstępnej ewaluacji. W scenariuszu rewolucyjnym rozwój i oddziaływanie systemów AI wykraczają poza granice państwowe i mają charakter globalny, natomiast regulacje oraz instytucje odpowiedzialne za ich egzekwowanie działają w ramach porządków lokalnych.

W kategoriach metryk M1–M5 można w skrócie pokazać, w jaki sposób trzy analizowane perspektywy — technologiczna, społeczna i regulacyjna — wraz ze scenariuszami rozwoju sztucznej inteligencji tworzą powiązany zestaw zagrożeń. Perspektywa technologiczna zwraca uwagę na sytuacje, w których wzrost sprawczości funkcjonalnej (F) wyprzedza możliwość utrzymania M2 i M3 na poziomie zgodnym z

lokalną zasadą przypisania odpowiedzialności (LSP) oraz zasadą superweniencji praktyk przypisywania odpowiedzialności (ZSP), co jest szczególnie widoczne w scenariuszach emulacyjnym i rewolucyjnym. Perspektywa społeczna pokazuje natomiast, że tego rodzaju osłabienia nakładają się na istniejące nierówności w strukturach władzy i kontroli, prowadząc do pogorszenia M1 i M4, zwłaszcza w scenariuszu ewolucyjnym. Z kolei perspektywa regulacyjna wskazuje, że parametry R i I pozostają rozproszone i niespójne, a reakcje instytucjonalne są często opóźnione, w wyniku czego M5, rozumiane jako relacja między tempem działania systemów a czasem dostępnym na refleksję, kontrolę i korektę w skali globalnej, pozostaje osłabione. W praktyce nie są to trzy niezależne obszary. Ryzyka technologiczne, społeczne i regulacyjne wzajemnie się wzmacniają, prowadząc do sytuacji, w których luka odpowiedzialności ma charakter strukturalny: nie zmieniają się formalne zasady przypisywania odpowiedzialności, lecz sposób, w jaki w praktyce jest ona rozdzielana między ludźmi a systemami AI.

6.6. Modele współodpowiedzialności i ład regulacyjny

Przedmiotem niniejszego podrozdziału jest przejście od diagnozy luki odpowiedzialności w systemach AI, ujętej w poprzednim podrozdziale za pomocą zasad LSP/ZSP oraz metryk M1–M5, do opracowania modeli współodpowiedzialności i odpowiadającego im ładu regulacyjnego. Analiza ma charakter normatywno-koncepcyjny i nie zmierza do ustalania podmiotów odpowiedzialnych w przypadkach wystąpienia szkody, lecz do wskazania takich konfiguracji $A \times F + (R, I)$, w których praktyki odpowiedzialności pozostają wykonalne i odporne na erozję w czasie. Pytanie przewodnie dotyczy tego, jakie modele współodpowiedzialności stabilizują warunki rozliczalności w sposób systematyczny, zamiast maskować deficyty, ograniczać się do późniejszej kompensacji szkód lub przerzucać ciężar odpowiedzialności na najsłabsze elementy układu.

Teza robocza ma charakter trójczłonowy. Po pierwsze, klasyczny model „indywidualnego sprawcy”, rozumianego jako pojedynczy człowiek albo producent, okazuje się niewystarczający w kontekście sztucznej podmiotowości. Opiera się on na uproszczonym modelu jednego sprawcy i jednej przypisanej mu funkcji odpowiedzialności, dlatego nie oddaje rzeczywistej złożoności współczesnych układów społeczno-technicznych. Po drugie, rozwiązania polegające na nadawaniu systemom AI osobowości prawnej częściej prowadzą do rozproszenia i osłabienia odpowiedzialności niż do jej wzmocnienia. Wprowadzają one dodatkowy poziom pośredni, którego nie da

się spójnie powiązać z realnymi możliwościami kontroli, interwencji oraz naprawy skutków działania. Po trzecie, adekwatną odpowiedź stanowią wielopoziomowe modele współodpowiedzialności. Modele te rozpisują obowiązki wzdłuż całego łańcucha społeczno-technicznego, są osadzone w instytucjach o charakterze nadzorczym i funkcjonują w ramach międzynarodowego ładu regulacyjnego. Celem niniejszego podrozdziału jest pokazanie, w jaki sposób można przejść od opisu luki odpowiedzialności do projektowania rozwiązań instytucjonalnych, które w sposób trwały umożliwiają utrzymywanie M1–M5 w przedziale zgodnym z LSP i ZSP, zamiast ograniczać się do działań następczych.

Punktem wyjścia analizy jest krytyczna rekonstrukcja dwóch skrajnych stanowisk wobec problemu luki odpowiedzialności. Pierwszym z nich jest stanowisko Andreasa Matthiasa, zakładające nieusuwalność luki odpowiedzialności w ramach obowiązujących praktyk przypisywania odpowiedzialności. Zgodnie z tym ujęciem w przypadku zaawansowanych automatów uczących się pojawiają się sytuacje, w których żaden podmiot nie spełnia warunków odpowiedzialności, w szczególności warunku kontroli zależnej od racji oraz możliwości prześledzenia działania i przypisania go do określonego sprawcy. W konsekwencji luka ma charakter nieusuwalny i stanowi bezpośredni koszt autonomii technicznej (Matthias, 2004, s. 175–183). Drugie stanowisko polega na formalnym rozszerzeniu istniejących kategorii prawnych. Zgodnie z nim lukę odpowiedzialności można usunąć bez przebudowy struktury $A \times F$, poprzez przypisanie odpowiedzialności określonemu podmiotowi, na przykład producentowi, niezależnie od zakresu sprawowanej kontroli albo poprzez wprowadzenie nowych podmiotów prawnych, takich jak „osobowość elektroniczna”. W obu wariantach źródła luki nie są korygowane na poziomie strukturalnym, a problem odpowiedzialności zostaje rozwiązany jedynie pozornie, na poziomie języka prawnego.

Oba stanowiska wykazują ograniczenia z perspektywy metryk M1–M5. Stanowisko Matthiasa, polegające na rezygnacji z prób ukształtowania nowego porządku odpowiedzialności, prowadzi do akceptacji obszarów wysokiego ryzyka, w których podmioty dotknięte szkodami nie mają realnej możliwości kierowania roszczeń do adresata odpowiedzialności. Oznacza to tolerowanie niskich wartości M1 i M4 przy braku skutecznych mechanizmów korekty oraz zapobiegania szkodom. Z kolei stanowisko oparte na formalnym rozszerzaniu istniejących kategorii prawnych często prowadzi do pozornej odpowiedzialności. Zamiast korekty struktury $A \times F + (R, I)$ dochodzi do przesunięcia odpowiedzialności na podmioty, które nie dysponują ani rzeczywistą

kontrolą nad działaniem systemu, ani zdolnością do zapobiegania szkodom lub ich naprawy. Dobrym przykładem są propozycje nadania systemom AI statusu „osoby elektronicznej”, krytykowane przez Joannę Bryson, Mihailisa Diamantisa i Thomasa Granta. Autorzy wskazują, że osobowość prawna maszyn może w praktyce służyć ochronie interesów korporacyjnych, zamiast zapewniać ochronę podmiotom narażonym na szkodę (Bryson, Diamantis & Grant, 2017, s. 273–291). Z perspektywy metryk M1–M5 żadna z tych strategii nie prowadzi do wzmocnienia rozliczalności. Przeciwnie, obie sprzyjają jej dalszemu osłabieniu, ponieważ pogłębiają nieprzejrzystość w zakresie identyfikowalności i ciągłości odpowiedzialności (M1, M4) oraz nie poprawiają dostępności racji ani możliwości kontroli zależnej od racji (M2–M3). W rezultacie problem, który miał zostać rozwiązany, pozostaje nierozstrzygnięty. Zostaje on jedynie przesunięty na poziom praktyk społecznych i prawnych, podczas gdy źródła luki odpowiedzialności pozostają niezmienione.

Responsibility should not be left to a single actor – neither a private company nor an isolated state—but rather entrusted to a democratic transnational governance, representing humanity in all its diversity. The founding values of this governance should include transparency, equity, non-maleficence, and respect for fundamental rights, while recognising cultural and moral plurality. Another approach emphasises the importance of distributed responsibility, where researchers, engineers, public decision-makers, and citizens actively participate in shaping intelligent systems through inclusive and adaptive processes. In short, the responsibility of a superintelligence can only be assumed collectively, based on universal but negotiated values, to prevent the concentrated power of a superhuman intelligence from serving narrow interests to the detriment of humanity as a whole (Dessureault i in., 2025, s. 6255–6256).

W nowszej literaturze z zakresu etyki sztucznej inteligencji wyodrębnia się trzecie ujęcie problemu luki odpowiedzialności. Zamiast uznawać ją za nieuchronną albo próbować usuwać ją poprzez formalne rozszerzanie kategorii prawnych, proponuje się zmianę samego pytania badawczego. Punkt ciężkości przesuwa się z pytania o to, kto jest odpowiedzialny w sensie pojedynczego sprawcy, na pytanie o to, jak należy ukształtować podział ról, kompetencji oraz mechanizmów korekty w złożonych układach techniczno-instytucjonalnych, aby luki odpowiedzialności były ograniczane i podlegały skutecznej kontroli. Analizy Mahdiego Kasirzadeha dotyczące odpowiedzialności zbiorowej i systemowej wskazują, że w przypadku rozproszonych systemów AI trafniejsze jest traktowanie odpowiedzialności w kategoriach wielopodmiotowej współodpowiedzialności niż przypisywanie jej jednemu sprawcy w ramach tradycyjnych

praktyk odpowiedzialności indywidualnej (Kasirzadeh, 2025). Zmianę tę można opisać jako przejście od pytania „kto odpowiada?” do pytania „jak zaprojektować strukturę przypisywania odpowiedzialności, która ogranicza naruszenia M1–M5 w danym układzie socjotechnicznym?”.

W przyjętym tu aparacie pojęciowym oznacza to odejście od modelu, w którym odpowiedzialność przypisuje się pojedynczemu sprawcy, na rzecz modelu kaskadowego obejmującego wiele ról w układzie społeczno-technicznym. Zbiór podmiotów odpowiedzialnych obejmuje projektantów architektury, podmioty odpowiedzialne za dane, integratorów systemów, operatorów, organy regulacyjne, audytorów oraz użytkowników instytucjonalnych. W odniesieniu do każdej z tych ról można wskazać odrębne zakresy odpowiedzialności, które oddziałują na metryki M1–M5 w różnym stopniu i w różny sposób. Projektanci architektury wpływają przede wszystkim na M2 i M3, a więc na możliwość audytu racji oraz na kontrolę zależną od racji. Podmioty odpowiedzialne za dane oddziałują głównie na M2 poprzez jakość, kompletność i reprezentatywność wykorzystywanych informacji. Integratorzy kształtują M3 i M5, ponieważ decydują o sposobie włączenia systemu do procesów decyzyjnych oraz o nasileniu presji czasowo-poznawczej. Organy regulacyjne wpływają przede wszystkim na M1 i M4, zapewniając identyfikowalność podmiotów oraz trwałość praktyk przypisywania odpowiedzialności. Użytkownicy instytucjonalni oddziałują natomiast na M3 i M4 poprzez procedury stosowania, dokumentowania i uzasadniania decyzji. Odpowiedzialność rozumiana systemowo wymaga zatem rozpisania ról i obowiązków wzdłuż całego łańcucha społeczno-technicznego tak, aby żadna relewantna rola nie pozostawała poza zakresem analizy M1–M5, nawet jeśli konkretne zobowiązania różnią się zakresem i natężeniem. Macierz $A \times F + (R, I)$ zachowuje odrębną funkcję klasyfikacyjną: służy opisowi systemu i jego osadzenia instytucjonalnego, nie oznacza łańcucha podmiotów odpowiedzialnych.

Na tym tle można wyróżnić trzy podstawowe modele współodpowiedzialności, odpowiadające różnym sposobom ograniczania luki odpowiedzialności: model odpowiedzialności produktowej, model odpowiedzialności instytucjonalnej oraz model odpowiedzialności statusowej, wraz z ich międzynarodowymi rozszerzeniami. Każdy z tych modeli opiera się na innym założeniu normatywnym: ochronie poszkodowanych, zapewnieniu przejrzystości i rozliczalności systemowej albo kształtowaniu statusów i relacji uznania. Każdy z nich ujawnia ograniczenia widoczne w scenariuszach emulacyjnym, ewolucyjnym i rewolucyjnym. W dalszej części podrozdziału modele te

zostaną poddane analizie porównawczej, z uwzględnieniem tego, w jaki sposób modyfikują rozkład metryk M1–M5 oraz jakie typy instytucji są dla nich charakterystyczne.

Model odpowiedzialności produktowej pozostaje najbliższy klasycznym ujęciom odpowiedzialności za produkt. Odpowiedzialność koncentruje się w nim na podmiotach wytwarzających i dostarczających systemy AI, takich jak producent, dostawca modelu bazowego lub operator platformy. Ujęcie to wpisuje się w unijne reformy dotyczące AI i odpowiedzialności za produkt: AI Act ustanawia obowiązki *ex ante* dla systemów wysokiego ryzyka, natomiast dyrektywa 2024/2853 ujmuje oprogramowanie jako produkt, określa kryteria wadliwości, wskazuje podmioty odpowiedzialne i wprowadza mechanizm ujawniania dowodów (Parlament Europejski i Rada Unii Europejskiej, 2024a, art. 9–14; 2024b, art. 4, 7–9). Z perspektywy metryk M1–M5 model ten wzmacnia przede wszystkim M1 i M4. Umożliwia bowiem jednoznaczne wskazanie adresata odpowiedzialności prawnej, a źródła odpowiedzialności za działanie systemu są przypisywane jednemu producentowi lub grupie producentów. Jego zaletą jest czytelne ukierunkowanie roszczeń oraz oparcie się na istniejącej infrastrukturze dochodzenia roszczeń i ubezpieczeń, co czyni go użytecznym w sytuacjach, w których relacja między producentem a użytkownikiem ma trwały i udokumentowany charakter.

Ograniczeniem tego modelu jest natomiast brak skutecznych mechanizmów pozyskiwania wiedzy o uzasadnieniach decyzji podejmowanych przez system oraz o ich korekcie. W złożonych systemach opartych na uczeniu maszynowym audyt racji (M2) bywa w znacznym stopniu niezależny od formalnej odpowiedzialności prawnej. Producent może ponosić odpowiedzialność za szkodę, nie dysponując pełną wiedzą o sposobach wykorzystania i modyfikowania modelu w konkretnych zastosowaniach. Rzeczywista kontrola zależna od racji (M3) bywa w tym modelu przesuwana na dalsze ogniwa łańcucha, takie jak integratorzy, operatorzy czy użytkownicy, którzy nie są głównymi adresatami tej odpowiedzialności. W efekcie model odpowiedzialności produktowej sprawdza się przede wszystkim tam, gdzie system AI ma jasno określone przeznaczenie i ograniczony zakres zastosowania, natomiast znacznie gorzej radzi sobie z systemami osadzonymi w rozproszonych układach instytucjonalnych, charakterystycznych dla scenariusza ewolucyjnego, a tym bardziej rewolucyjnego. Wraz ze wzrostem liczby interfejsów i poziomów integracji rośnie bowiem ryzyko, że odpowiedzialność przypisana w jednym punkcie nie przełoży się na możliwość korekty i kontroli w całym łańcuchu społeczno-technicznym.

Model odpowiedzialności instytucjonalnej przenosi odpowiedzialność z jednego producenta na cały ciąg podmiotów uczestniczących w projektowaniu, wdrażaniu i stosowaniu systemu. Przyjmuje on, że odpowiedzialność powinna być przypisana wszystkim podmiotom tworzącym dany układ socjotechniczny, w tym projektantom architektury, podmiotom odpowiedzialnym za dane, integratorom, operatorom systemów AI, organom regulacyjnym, audytorom oraz użytkownikom instytucjonalnym. Szczególną rolę odgrywają tu instytucje sprawujące nadzór nad funkcjonowaniem systemów AI, prowadzące kontrole, aktualizujące standardy oraz przygotowujące testy bezpieczeństwa i procedury postępowania w sytuacjach nadzwyczajnych. Z perspektywy metryk M1–M5 model ten wzmacnia przede wszystkim M2 i M3, a pośrednio także M5. Uzasadnianie decyzji oraz kontrola oparta na racjach zostają w nim włączone bezpośrednio w proces projektowania i nadzoru nad systemami AI. Jednocześnie presja czasowo-poznawcza ulega ograniczeniu dzięki obowiązkowi prowadzenia dokumentacji, oceny wpływu na otoczenie oraz możliwości awaryjnego wstrzymania działania systemu. Odpowiedzialność ma w tym modelu charakter rozproszony, lecz nie jest rozmyta. Każdej roli w układzie społeczno-technicznym przypisuje się określony zakres zadań i wymagań, które mogą być egzekwowane zgodnie z obowiązującymi zasadami i procedurami.

Model odpowiedzialności statusowej kładzie nacisk na precyzyjne określenie, komu przypisuje się odpowiedzialność oraz które podmioty ponoszą konsekwencje działań systemów sztucznej inteligencji. Znaczenie ma nie tylko zakres kontroli nad systemem i związane z nią obowiązki naprawcze, lecz także zakres uprawnień oraz obowiązków użytkowników, grup szczególnie narażonych i instytucji nadzorczych, a w określonych przypadkach również samych systemów AI. Włączenie systemu do roli proceduralnej wymaga etapowej oceny: K1–K3 wspierają diagnozę H1, F3 musi zostać wykazane odrębnie jako zdolność jawnego i audytowalnego uzasadniania normatywnego, a odpowiednie R + I oraz WI warunkują instytucjonalne zastosowanie statusu. H2 i K4 mogą wzmacniać stabilność działania, natomiast H3 wyznacza dopiero próg silnej rozliczalności; nie tworzą one jednolitego minimalnego pakietu quasi-podmiotowości. Odpowiedzialność statusowa porządkuje role i uprawnienia w sferze publicznej, nie nadając automatycznie systemowi osobowości moralnej lub prawnej. Z perspektywy M1–M5 model ten wzmacnia przede wszystkim M1, czyli identyfikowalność podmiotów odpowiedzialnych oraz M4, czyli przejrzystość i trwałość praktyk przypisywania odpowiedzialności. Umożliwia także ochronę podmiotów, na

które przerzucane są skutki luki czasowo-poznawczej. Takie ukształtowanie ról wymaga refleksji nad włączaniem systemów technicznych i instytucji nadzorczych do porządku uzasadniania i legitymizowania decyzji, przy zachowaniu różnicy między zdolnościami systemu a warunkami WI.

Trzy omawiane modele współodpowiedzialności pozostają ze sobą ściśle powiązane. Model odpowiedzialności produktowej zapewnia podstawowe mechanizmy dochodzenia roszczeń oraz naprawy szkód. Model odpowiedzialności instytucjonalnej umożliwia stały nadzór nad działaniem systemów AI oraz ich bieżącą korektę. Model odpowiedzialności statusowej porządkuje role i pozycje normatywne w perspektywie instytucjonalnej. Ograniczanie luki odpowiedzialności wymaga łączenia tych podejść, a nie wyboru jednego z nich, ponieważ każde odnosi się do innych wymiarów M1–M5 i opiera się na odmiennych mechanizmach instytucjonalnych.

Zastosowanie omawianych modeli do trzech scenariuszy rozwoju AI, emulacyjnego, ewolucyjnego i rewolucyjnego, pozwala ocenić, w jakim stopniu praktyki odpowiedzialności mogą zostać utrzymane przy minimalnej adresowalności proceduralnej. W scenariuszu emulacyjnym, w którym systemy AI w dużej mierze odtwarzają istniejące praktyki decyzyjne, model odpowiedzialności produktowej okazuje się w znacznej mierze wystarczający, zwłaszcza gdy zostaje uzupełniony o elementy odpowiedzialności instytucjonalnej, takie jak nadzór i kontrola. W takim układzie luki odpowiedzialności w zakresie M1–M3 można ograniczać poprzez dokumentację, testy bezpieczeństwa oraz procedury nadzorcze. Systemy AI pełnią tu funkcję zaawansowanych narzędzi technicznych, a struktura A×F pozostaje niezmieniona. Kluczowe znaczenie ma wówczas jasne przypisanie odpowiedzialności producentowi oraz podmiotowi korzystającemu z systemu, a także zapewnienie, że użytkownicy instytucjonalni dysponują mechanizmami kontroli zarówno przed podjęciem decyzji, jak i po jej podjęciu.

W scenariuszu emulacyjnym istotne jest precyzyjne określenie zakresu quasi-podmiotowości systemów sztucznej inteligencji. Jeżeli systemy te jedynie odtwarzają istniejące procedury decyzyjne, ich status instytucjonalny może pozostać zbliżony do statusu narzędzi. W takim przypadku zasadniczą rolę odgrywa model odpowiedzialności produktowej, uzupełniony o model odpowiedzialności instytucjonalnej. Taki układ sprzyja jednoznacznemu przypisywaniu odpowiedzialności i pozwala uniknąć zarówno nadmiernego przypisywania systemom cech podmiotowych, jak i całkowitego wyłączenia ich z porządku odpowiedzialności. W efekcie ogranicza on luki związane ze

statusem odpowiedzialności oraz z jej przypisaniem. Rozszerzenia modelu odpowiedzialności statusowej mają w tym scenariuszu charakter pomocniczy i dotyczą przede wszystkim ochrony grup narażonych na systemowe uprzedzenia oraz zapewnienia, że decyzje podejmowane z udziałem systemów AI podlegają publicznemu uzasadnieniu w ramach istniejących instytucji.

W scenariuszu ewolucyjnym głównym problemem stają się luki o charakterze akumulacyjno-strukturalnym, polegające na stopniowym pogarszaniu wskaźników M1–M4 przy narastającym obciążeniu M5. W takim układzie model odpowiedzialności produktowej może pełnić jedynie rolę pomocniczą. Umożliwia naprawę szkód i sprzyja ostrożności, lecz nie zapobiega osłabianiu struktur odpowiedzialności. Kluczowe znaczenie zyskuje model odpowiedzialności instytucjonalnej, który pozwala rozłożyć obowiązki związane z M1–M5 wzdłuż całego łańcucha społeczno-technicznego, wzmacnia nadzór oraz porządkuje narzędzia oceny skutków, kontroli i reagowania. W skali międzynarodowej konieczne stają się rozwiązania wykraczające poza pojedyncze porządki prawne, takie jak wspólne standardy dotyczące M1–M5, mechanizmy ograniczające obchodzenie regulacji oraz koordynacja polityk między systemami AI o rosnącym stopniu autonomii, wykorzystywanymi w procesach instytucjonalnych. W tym scenariuszu rośnie również znaczenie modelu odpowiedzialności statusowej. Integracja systemów AI z instytucjami społecznymi prowadzi do wyłaniania się nowych ról quasi-podmiotowych, co wymaga określenia ich statusu w świetle progów $F3 + R + I$ oraz sprecyzowania obowiązków informacyjnych i partycypacyjnych dla instytucji korzystających z takich rozwiązań. Celem tych działań jest zapobieżenie rozmywaniu odpowiedzialności instytucjonalnej oraz przekształcenie jej w uporządkowany system współodpowiedzialności.

W scenariuszu rewolucyjnym zasadniczym problemem staje się luka czasowo-poznawcza. Modele współodpowiedzialności muszą wówczas obejmować decyzje podejmowane na bardzo wczesnych etapach rozwoju systemów AI, w szczególności decyzje dotyczące tego, czy określone kierunki badań nad systemami zdolnymi do samodoskonalenia powinny być w ogóle podejmowane oraz w jaki sposób należy je regulować. Wiąże się to z koniecznością objęcia współodpowiedzialnością decyzji o charakterze globalnym, wyznaczających minimalny dopuszczalny poziom M5, rozumiany jako relacja między tempem zmian a czasem i dostępem do wiedzy potrzebnymi instytucjom do refleksji, kontroli oraz reakcji. Pytanie o odpowiedzialność dotyczy w tym scenariuszu tego, jaki zakres niewiedzy i jakie tempo przemian można

uznać za moralnie i politycznie dopuszczalne jeszcze przed wystąpieniem realnych szkód, a także tego, które instytucje są uprawnione do wyznaczania takich granic.

W tym kontekście model odpowiedzialności produktowej niemal całkowicie traci znaczenie, ponieważ żadna forma odszkodowania nie jest w stanie naprawić skutków tego rodzaju ryzyka. Model odpowiedzialności instytucjonalnej zachowuje jedynie ograniczoną rolę i to wyłącznie wówczas, gdy instytucje nadzorcze dysponują narzędziami spowalniania rozwoju oraz egzekwowania globalnych ograniczeń, na przykład poprzez moratoria, kontrolowane środowiska testowe lub mechanizmy weryfikacji zgodności. Znaczenie zyskuje natomiast globalny model współodpowiedzialności politycznej, w ramach którego państwa, organizacje międzynarodowe, firmy technologiczne oraz społeczeństwa obywatelskie współdzielą odpowiedzialność za wyznaczanie kierunków badań i regulacji, tak aby nie wkraczać w obszary skrajnie niskiego M5. Wymaga to nie tylko powołania nowych instytucji, lecz także wypracowania zasad legitymizowania decyzji o ograniczaniu badań lub wdrożeń, akceptowalnych z perspektywy różnych grup oraz zgodnych z LSP i wymogami rzetelnego, przejrzystego podejmowania decyzji. W tym sensie model odpowiedzialności statusowej przekształca się w model o charakterze społeczno-politycznym, w którym przedmiotem sporu staje się nie tylko podział odpowiedzialności, lecz także kształt globalnego porządku normatywnego.

Podsumowanie

Analiza przeprowadzona w rozdziale 6 prowadzi do kilku zasadniczych wniosków dotyczących charakteru luki odpowiedzialności oraz możliwości jej ograniczania w kontekście rozwoju sztucznej inteligencji. Po pierwsze, luka odpowiedzialności nie ma charakteru pojedynczego zdarzenia ani nie da się jej wyjaśnić wyłącznie na poziomie ontologicznym. Klasyczny brak podmiotu zdolnego do ponoszenia odpowiedzialności pozostaje jedną z jej możliwych postaci, lecz luka nie wyczerpuje się w tym przypadku. W scenariuszach emulacyjnym, ewolucyjnym i rewolucyjnym może wynikać także z określonych konfiguracji technicznych i instytucjonalnych, w których zawodzą identyfikowalność, audytowalność racji, kontrola zależna od racji lub trwałość praktyk. Problem odpowiedzialności dotyczy zatem zarówno pytania, czy można wskazać system spełniający odpowiednie progi agentowe, jak i istnienia struktur zdolnych do utrzymania relacji odpowiedzialności przy rosnącej autonomii i złożoności działania. Po drugie, zestawienie trzech modeli rozwoju AI

pokazuje, że luka odpowiedzialności przybiera różne postacie w zależności od dynamiki $A \times F$. W scenariuszu emulacyjnym ma ona charakter lokalny i w dużej mierze pozostaje pod kontrolą. W scenariuszu ewolucyjnym narasta stopniowo i strukturalnie, co prowadzi do coraz słabszego spełniania warunków LSP i ZSP. W scenariuszu rewolucyjnym osiąga postać graniczną, w której problemem staje się niedopasowanie tempa rozwoju systemów AI do możliwości refleksji, nadzoru i korekty. We wszystkich przypadkach zakres i natężenie luki odpowiedzialności zależą zarówno od zdolności i ograniczeń systemów technicznych, jak i od sposobu projektowania, nadzorowania oraz regulowania sieci społeczno-technicznych. Po trzecie, zastosowanie metryk M1–M5 pozwala rozumieć lukę odpowiedzialności jako zjawisko wielowymiarowe. Metryki te umożliwiają systematyczny opis różnych form naruszeń warunków LSP i ZSP oraz ich porównywanie między scenariuszami rozwoju AI. Ujęcie metryczne nie zastępuje oceny normatywnej, lecz zwiększa jej przejrzystość, przenosząc analizę z poziomu ogólnych intuicji na poziom dopuszczalnych zakresów identyfikowalności, kontroli, korekty, trwałości praktyk przypisywania odpowiedzialności oraz obciążenia czasowo-poznawczego. Po czwarte, analiza modeli współodpowiedzialności pokazuje, że adekwatna odpowiedź na lukę odpowiedzialności nie polega ani na jej akceptacji, ani na rozszerzaniu istniejących kategorii prawnych. Modele odpowiedzialności produktowej, instytucjonalnej i statusowej odnoszą się do różnych aspektów luki oraz do odmiennych konfiguracji metryk M1–M5. Ich łączne zastosowanie umożliwia modyfikowanie struktury $A \times F + (R, I)$ w taki sposób, aby ograniczać lukę odpowiedzialności w sposób trwały, a nie jedynie reagować na jej doraźne skutki. W tym sensie ład regulacyjny nie stanowi zewnętrznego ograniczenia rozwoju AI, lecz warunku możliwości rozsądnej współodpowiedzialności ludzi i systemów technicznych.

W rezultacie problem sztucznej podmiotowości i odpowiedzialności nie sprowadza się ani wyłącznie do wyznaczania granicy między „podmiotem” a „niepodmiotem”, ani wyłącznie do projektowania sieci społeczno-technicznych. Wymaga równoległej oceny zdolności technicznego systemu w świetle H1, H3 i F3 oraz projektowania, nadzorowania i regulowania układu społeczno-technicznego tak, aby WI, R i I umożliwiały spójne przypisywanie odpowiedzialności i ochronę podmiotów szczególnie narażonych na skutki działania AI. Rozdział 6 prowadzi zatem od analizy progów systemowych do analizy współodpowiedzialności instytucjonalnej, zachowując różnicę między tymi poziomami. W tym sensie stanowi punkt wyjścia do końcowej oceny normatywnej i porównania reżimów odpowiedzialności w rozdziale 7.

Rozdział 7. Synteza i weryfikacja H1–H3 oraz WI: hierarchia statusów moralnych AI, implikacje i rekomendacje

7.1. Omówienie wyników i weryfikacja H1–H3 oraz WI

Przedmiotem niniejszego podrozdziału jest synteza wyników rozdziałów 1–6, ujęta jako weryfikacja H1–H3 oraz WI. Zasadnicze pytanie brzmi: w jakim stopniu hipotezy te wyznaczają racjonalnie dopuszczalne warunki przypisywania sztucznej podmiotowości moralnej przy założeniu minimalizmu ontologicznego oraz wysokich wymagań normatywnych? Dysertacja nie rozstrzyga sporów ontologicznych ponad to, co konieczne dla klarowności argumentacji, lecz zachowuje rygorystyczne kryteria uzasadniania roszczeń odpowiedzialności. Teza robocza głosi, że H1–H3 tworzą uporządkowaną strukturę funkcji normatywnych: H1 wyznacza minimalną adresowalność proceduralną i podstawę słabej odpowiedzialności proceduralnej, H2 określa afektywność funkcjonalną jako możliwy mechanizm stabilizujący realizację racji, a H3 wyznacza silną rozliczalność poprzez kontrolę metapoznawczą oraz jawne rozstrzyganie konfliktów racji. K1–K3 wspierają operacyjno-diagnostyczną ocenę H1, natomiast K4 odrębnie ocenia stabilność diachroniczną i wiarygodność profilu działania; spełnienie K4 nie wykazuje samo przez się H2. WI określa warunek instytucjonalnego zastosowania H1–H3. Wyniki rozdziału 4 należy czytać jako etycznie-normatywne uzasadnienie tej struktury: PH1 doprecyzowuje normatywny sens H1, PH2 uzasadnia normatywne znaczenie stabilizacji i wiarygodności bez utożsamiania ich z mechanizmem H2, a PH3 wspiera H3 jako silną rozliczalność dzięki kontroli metapoznawczej oraz jawnemu rozstrzyganiu konfliktów racji. Weryfikacja przebiega zatem dwutorowo: po stronie funkcji normatywnych poprzez analizę H1–H3 oraz kryteriów K1–K4, a po stronie instytucjonalnej poprzez ocenę WI, w tym MHC i MLI. Taki układ pozwala odróżnić minimalną adresowalność od silnych form rozliczalności, rozpoznać odmienne typy naruszeń oraz uzasadnić przejście do hierarchii statusów moralnych rozumianej jako ujęcie stopniowalne i zależne od ram instytucjonalnych.

Termin „weryfikacja” ma tu znaczenie filozoficzne, nie empiryczne. Oznacza sprawdzenie spójności pojęciowej hipotez, poprawności wnioskowania oraz adekwatności ich funkcji normatywnej. Chodzi o ocenę relacji między przesłankami a

wnioskami w świetle przyjętych definicji, kryteriów operacyjnych oraz założeń metodologicznych, a nie o testowanie hipotez w sensie eksperymentalnym. Zastosowanie modelu do IDx-DR/LumineticsCore nie zmienia tego statusu metodologicznego. Wyniki kliniczne i dokumentacja regulacyjna uzasadniają wyłącznie twierdzenia opisowe o systemie, natomiast ocena H1–H3 + WI pozostaje analizą pojęciową i sprawdzaniem koherencyjnym (zob. podrozdz. 5.4).

Kluczowe jest pytanie, czy hipotezy te dają się osadzić w perspektywie drugiej osoby, rozumianej jako zbiór praktyk, w których podmioty formułują roszczenia o uzasadnienie i korektę oraz poddają je ocenie. Następnie należy ustalić, czy w tych procedurach możliwe jest przypisywanie odpowiedzialności w sposób podlegający kontroli publicznej, a więc odporny na arbitralne zmiany znaczeń i redefinicje. W dysertacji punktem odniesienia dla tej analizy jest rozróżnienie pojęć odpowiedzialności: przypisywalności (ang. *attributability*), udzielania racji (ang. *answerability*) oraz rozliczalności (ang. *accountability*). Przypisywalność dotyczy tego, jakie działanie można zasadnie przypisać określonej podmiotowi; udzielanie racji odnosi się do obowiązku odpowiadania na roszczenia uzasadnienia; rozliczalność obejmuje natomiast uprawnioną krytykę i sankcje, w tym reguły zawieszania oraz przywracania praktyk przypisywania odpowiedzialności (Watson, 1996, s. 227–248; Darwall, 2006, rozdz. 4, s. 65–90; rozdz. 5, s. 91–118; Wallace, 1998, s. 51–83).

Z tego rozróżnienia wynika pierwszy ogólny wynik weryfikacji. H1–H3 należy traktować jako czysto opisowych cech systemów technicznych. Są to normatywnie określone warunki agentowe i funkcjonalne, których przypisanie wymaga wykazania odpowiednich zdolności systemu. WI ma odmienny przedmiot: określa warunki instytucjonalnego zastosowania modelu w układzie społeczno-technicznym. Odpowiedzialność ma bowiem charakter relacji normatywnej, nie samego faktu sprawczego, choć fakty sprawcze stanowią jej niezbędne tło. Rozróżnienie to zapobiega utożsamianiu sprawstwa z rozliczalnością oraz przypisywaniu systemowi kompetencji ludzi i instytucji. Odpowiada także na problem „luki odpowiedzialności”: nawet jeżeli działanie systemu prowadzi do określonych skutków, sama sprawczość nie rozstrzyga, kto jest adresatem krytyki, kto ma obowiązek korekty i kto powinien podlegać rozliczeniu w ramach jasno określonych ról oraz procedur poddających się kontroli (Matthias, 2004, s. 175–183; Santoni de Sio & van den Hoven, 2018, s. 1–4; Mecacci & Santoni de Sio, 2020, s. 103–115).

Drugi ogólny wynik dotyczy zakresu, w jakim dysertacja odwołuje się do sporów o wolną wolę. W rozdziale 2 wykazano, że odpowiedzialność nie musi zależeć od rozstrzygnięcia metafizycznego sporu o determinizm, choć spór ten ogranicza roszczenia do pełnej odpowiedzialności w sensie opartym na zasłudze oraz zakres dozwolonej odpłaty. Perspektywa Strawsona uzasadnia w tym miejscu zmianę akcentu. Moralność wiąże się z nastawieniami reaktywnymi oraz regułami ich zawieszania, a status podmiotu moralnego polega na uczestnictwie w relacjach wzajemnego adresowania i odpowiadania (Strawson, 2019, s. 35–57, 55–57). Jednocześnie argument Strawsona oraz twardej inkompatybilizm Perebooma pokazują, że z takiego ujęcia nie wynika bezpośrednio uzasadnienie odpłaty opartej na zasłudze. Weryfikacja hipotez ma zatem charakter praktyczny. Wskazuje warunki zasadnego domagania się racji i korekty, a nie warunki odpowiedzialności wymagającej uznania podmiotu za ostateczne źródło własnych działań (Strawson, 1994, s. 5–7, 15–17; Pereboom, 2001, rozdz. 5, s. 123–157). Ustalenie to ma znaczenie kluczowe, ponieważ w 7.3 uzasadnia ostrożność w przechodzeniu od opisu zdolności funkcjonalnych systemu do normatywnego uzasadnienia sankcji oraz od autonomii na poziomie wykonania do przypisania moralnej odpowiedzialności rozumianej jako podstawa odpłaty.

Perspektywa Strawsona określa, kiedy i z jakich powodów zasadne jest kierowanie roszczeń wobec drugiej osoby w ramach praktyk odpowiedzialności. Ujęcie Fischera i Ravizy nie wprowadza dodatkowych założeń metafizycznych, lecz doprecyzowuje, kiedy dany mechanizm może być uznany za właściwego adresata takich roszczeń oraz kiedy zachodzą podstawy do ich zawieszenia, na przykład w przypadkach manipulacji, przymusu lub obejścia kontroli decyzyjnej. W tym sensie oba ujęcia są spójne. Strawson wyznacza ramy normatywne, natomiast Fischer i Ravizza określają warunki ich stosowania.

Na tym tle przechodzę do H1 jako prognozy minimalnego. Hipoteza H1 głosi, że system może zostać uznany za minimalnie adresowalny proceduralnie, jeżeli wykazuje funkcjonalną zdolność rozpoznawania norm, przewidywania konsekwencji ich naruszeń, reagowania na racje w trybie kontrfaktycznym oraz udostępniania wystarczających i wiarygodnych podstaw decyzji, umożliwiających ocenę ich związku z wynikiem, audyt, zakwestionowanie i korektę. Weryfikacja tej hipotezy wymaga doprecyzowania dwóch pojęć. Po pierwsze, minimalna adresowalność proceduralna nie sprowadza się ani do nieprzewidywalności, ani do braku wpływów zewnętrznych. Oznacza możliwość włączenia decyzji systemu do praktyk publicznego uzasadniania, krytyki i korekty bez

przesądzania o autonomii moralnej systemu. W dalszej analizie „podatność na korektę” nie ma charakteru definicyjnego, lecz pełni funkcję operacyjnego kryterium minimalnej adresowalności proceduralnej w praktykach uzasadniania. Po drugie, „rozpoznawanie racji” nie może być utożsamiane z samym ich deklarowaniem, ponieważ deklaracja może być normatywnie jałowa, to znaczy nie prowadzić do korekty działania. Na tym etapie dysertacja odwołuje się do rozróżnienia między świadomością dostępu (A-świadomością), rozumianą jako dostępność treści do raportowania, namysłu i sterowania działaniem, a świadomością fenomenalną. Dla minimalnej odpowiedzialności praktycznej decydująca jest dostępność racji do kontroli i korekty, a nie założenie określonej jakości przeżyć (Block, 1995, s. 227–247; Levy, 2014, s. 14–37, 109–131). W rezultacie H1 od początku określona jest jako próg adresowalności w perspektywie drugiej osoby, a nie jako próg odnoszący się do wewnętrznych stanów mentalnych systemu. Nie znaczy to jednak, że A-świadomość jest świadomościowym składnikiem H1: H1 pozostaje minimalną adresowalnością proceduralną, a A-świadomość lub jej funkcjonalny ekwiwalent opisuje jedynie dostępność racji potrzebną do oceny tej adresowalności; minimalna adresowalność proceduralna nie wymaga świadomości fenomenalnej.

Tak rozumiany minimalizm ontologiczny ma charakter metodologiczny i nie usuwa napięć normatywnych. Po pierwsze, pojęcia „rozpoznawanie racji” oraz „podatność na korektę” mają tu znaczenie proceduralne: określają warunki włączania systemu w publiczne praktyki formułowania roszczeń, uzasadniania, krytyki i korekty, a nie warunki posiadania określonego rodzaju doświadczenia ani wewnętrznej intencjonalności. Po drugie, minimalizm ontologiczny nie oznacza osłabienia wymagań normatywnych: nawet bez rozstrzygania ontologii stanów wewnętrznych pozostaje obowiązek odróżniania wrażliwości na racje od czysto przyczynowej adaptacji. Po trzecie, w wielu przypadkach różnica między rozpoznawaniem racji a jej symulacją może pozostać nierozstrzygalna. W dysertacji pełni ona zatem funkcję graniczną, ponieważ uzasadnia powściągliwość w przechodzeniu do silniejszych praktyk odpowiedzialności, lecz nie wyklucza uznania progu adresowalności, o ile spełnione są operacyjne kryteria podatności na korektę.

Na potrzeby weryfikacji H1 pojęcie „podatności na korektę” przyjmuję w ujęciu operacyjnym, niezależnym od silnych definicji autonomii normatywnej. Obejmuje ono: (1) wrażliwość kontrfaktyczną na istotne racje, rozumianą jako gotowość do odmiennego rozstrzygnięcia w przypadku zmiany racji; (2) zdolność do zmiany rozstrzygnięcia pod

wpływem nowych racji, a nie jedynie do parametrycznego dostrajania zachowania; oraz (3) możliwość wykazania związku między rozpoznaną racją, przedstawioną podstawą decyzji i faktyczną korektą. Stabilność reakcji przy niewielkich zmianach kontekstu i spójność czasowa wzmacniają wiarygodność przypisania oraz podlegają ocenie K4, lecz nie stanowią warunku koniecznego H1. Spełnienie tych warunków stanowi minimalny wskaźnik sensownej adresowalności, bez przesądzania o istnieniu silniejszej agencji.

Weryfikacja H1 wymaga w pierwszym kroku ustalenia, jaki rodzaj kontroli ma znaczenie moralne. Zasada alternatywnych możliwości (PAP) zakłada, że odpowiedzialność wymaga istnienia alternatywy w działaniu. Analiza przypadków frankfurckich oraz związanej z nimi debaty przesuwają jednak punkt ciężkości. Odpowiedzialność może zachodzić także wtedy, gdy taka alternatywa nie istnieje, o ile decyzja wynika z mechanizmu spełniającego warunki normatywnej kontroli i w którym racje pełnią funkcję regulującą działanie (Frankfurt, 1969, s. 829–839; Fischer & Ravizza, 1998, rozdz. 2, s. 28–61; Darwall, 2006, s. 87–118). Ustalenie to prowadzi do dwóch wniosków. Po pierwsze, H1 nie wymaga wprowadzania warunku istnienia alternatywnych możliwości działania i może koncentrować się na mechanizmie reagowania na racje, a więc na tym, czy krytyka i korekta rzeczywiście wpływają na działanie. Po drugie, umożliwia wyraźne odróżnienie kontroli sprawczej od kontroli odpowiedzialności. Kontrola odpowiedzialności ma znaczenie wtedy, gdy czyni system podatnym na uzasadnioną krytykę i korektę w praktykach drugiej osoby, a nie wtedy, gdy jedynie zwiększa jego samodzielność w działaniu. W tym sensie H1 nie ustanawia autonomii moralnej systemu, lecz wyznacza minimalny punkt odniesienia dla procedur krytyki i korekty.

Dysertacja porządkuje te intuicje, uznając teorię kontroli sterującej Fischera i Ravizy za doprecyzowanie warunków zasadnego formułowania i korygowania roszczeń o uzasadnienie w perspektywie drugiej osoby, a nie za konkurencyjne ujęcie wobec koncepcji Strawsona. Odpowiedzialność praktyczna wymaga wrażliwości na racje oraz spełnienia warunku normatywnego przypisania mechanizmu działania. Nie chodzi tu o własność prawną, lecz o możliwość zasadnego traktowania danego mechanizmu jako podstawy przypisania roszczeń oraz o uznanie go za właściwy punkt odniesienia w praktykach rozliczalności (Frankfurt, 1969, s. 829–839; Fischer & Ravizza, 1998, rozdz. 3, s. 62–91).

Warunek historyczny, określany jako zasada nieomijania krytyki i korekty (ang. *no-bypassing*), pełni funkcję zabezpieczenia przed ingerencjami, które podważają

zasadność przypisania odpowiedzialności. Jeżeli mechanizm został ukształtowany w sposób omijający proces nabywania norm oraz zdolność do korekty, przypisanie odpowiedzialności traci na normatywnych podstawach. Dysertacja nie interpretuje tego warunku jako tezę o istnieniu wewnętrznego podmiotu, lecz jako wymóg normatywnie uzasadnionego mechanizmu działania. Konieczne jest zatem wykazanie, w jaki sposób system uzyskał zdolność reagowania na krytykę i korektę. Odrębnie należy ustalić, czy proces kształtowania jego mechanizmu działania został osadzony w publicznych procedurach oceny i rewizji; ten drugi wymóg należy do WI.

Warunek historyczny rodzi szczególną trudność w odniesieniu do systemów sztucznej inteligencji. Można wyróżnić trzy jego interpretacje. Ujęcie liberalne uznaje każde uczenie się za formę „nabywania norm” w znaczeniu funkcjonalnym. Ujęcie rygorystyczne przyjmuje natomiast, że żaden system sztucznej inteligencji nie spełnia warunku historycznego, ponieważ nie posiada historii przyswajania norm porównywalnej z historią osoby. Ujęcie pośrednie, przyjęte w niniejszej pracy, dopuszcza analogię warunku historycznego wyłącznie w układach socjotechnicznych, w których mechanizm podlega krytyce i korekcie oraz został ukształtowany w ramach publicznych procedur oceny, rewizji i rozliczalności. W tej perspektywie zasada nieomijania krytyki i korekty oznacza zakaz projektowania i modyfikowania systemu w sposób, który uniemożliwia korektę jego działania, nawet gdy system wykazuje tylko pozorną wrażliwość na racje. Dlatego H1 podlega weryfikacji wyłącznie wtedy, gdy wrażliwość na racje nie ma charakteru czysto formalnego, lecz stanowi element mechanizmu, który podlega ocenie normatywnej, może być korygowany oraz przypisywany do określonych ról (Fischer & Ravizza, 1998, rozdz. 3, s. 62–91; rozdz. 8, s. 207–239; Levy, 2014, s. 109–131).

Kolejny etap weryfikacji polega na uwzględnieniu stanowisk alternatywnych wobec kompatybilizmu. Libertarianizm podkreśla źródłowy charakter wyboru oraz aktywny udział podmiotu w kształtowaniu własnych dyspozycji decyzyjnych przy założeniu alternatywnych możliwości działania, natomiast argument z konsekwencji determinizmu wskazuje, że determinizm osłabia kontrolę, ujmując działanie jako skutek przeszłych stanów i praw natury. Dysertacja traktuje te stanowiska jako wyznaczające górny pułap roszczeń, a więc jako punkt odniesienia dla bardziej rozległego rozumienia odpowiedzialności. Jeżeli od odpowiedzialności wymaga się metafizycznej źródłowości w sensie ścisłym, H1 okazuje się warunkiem niewystarczającym. Projekt dysertacji nie zakłada jednak konieczności przyjmowania tego pułapu odpowiedzialności praktycznej. H1 zachowuje ważność jako warunek minimalnej adresowalności proceduralnej, zgodny

z umiarkowanym kompatybilizmem, i pozostaje niezależny od tego, czy odpowiedzialność stanowi pojęcie normatywnie uzasadnione. W tym znaczeniu spory libertariańskie oraz argumenty antydeterministyczne pełnią funkcję graniczną. Wskazują na istnienie silniejszych koncepcji odpowiedzialności niż ta, którą dysertacja uznaje za adekwatną do stopniowalnej sztucznej podmiotowości moralnej (van Inwagen, 1983, s. 56; Kane, 1996, s. 126–135; Levy, 2014, s. 109–131).

Wynik weryfikacji H1 można więc ująć w sposób wyważony. Hipoteza ta trafnie wskazuje warunek konieczny minimalnej adresowalności. Bez zdolności rozpoznawania norm, kontrfaktycznego reagowania na racje oraz przedstawiania wystarczających i wiarygodnych podstaw decyzji, które umożliwiają audyt, zakwestionowanie i korektę, nie ma podstaw do kierowania wobec systemu roszczeń drugiej osoby. H1 nie wystarcza jednak do silnej rozliczalności, ponieważ nie rozstrzyga kwestii afektywności funkcjonalnej H2, stabilności diachronicznej ocenianej przez K4 ani sposobu porządkowania racji w sytuacjach konfliktowych właściwego H3. Mini-konkluzja jest zatem następująca: H1 zostaje potwierdzona jako próg minimalny, lecz jej sens polega na określeniu warunków odpowiedzialności praktycznej, a nie na ustanowieniu podstaw sankcji opartych na przypisaniu zasługi. Dalsza analiza musi odrębnie ocenić możliwy mechanizm stabilizujący H2 oraz rezultaty diagnozowane przez K4.

Hipoteza H2 nie zakłada przypisywania systemom AI stanów emocjonalnych, lecz wskazuje na funkcjonalny mechanizm stabilizacji ich działania. Chodzi o zintegrowane mechanizmy wartościowania i regulacji, dzięki którym rozpoznane racje wpływają na motywację do działania lub jego powstrzymania, zamiast pozostawać na poziomie deklaratywnym. W literaturze poświęconej systemom autonomicznym wskazuje się, że wykorzystywanie wewnętrznych reprezentacji stanów oraz procedur przewidywania skutków decyzji może pełnić funkcję ograniczającą działanie systemu. W dysertacji motyw ten pełni funkcję ilustracyjną i nie stanowi przesłanki empirycznej (Winfield, Blum, & Liu, 2014, s. 86). Innymi słowy, H2 dotyczy tego, czy normy z poziomu deklaratywnego przechodzą na poziom regulacyjny, obejmujący sterowanie uwagą, progami reakcji oraz odpornością na zakłócenia. Zgodnie z aparatem pojęciowym rozdziału 3 H2 wskazuje określony rodzaj mechanizmu stabilizującego, natomiast K4 odrębnie ocenia rezultat w postaci stabilności diachronicznej i wiarygodności profilu działania. Spełnienie K4 nie wykazuje samo przez się afektywności funkcjonalnej. H2 może wzmacniać trwałość realizacji racji, lecz nie stanowi warunku minimalnej adresowalności H1.

Heurystyka aretologiczna pozwala ująć H2 bez przypisywania systemom AI cech ludzkich. Dysertacja nie przypisuje cnót artefaktom, lecz analizuje cnoty praktyk instytucjonalnych. Hagendorff wskazuje, że etyka cnót w kontekście AI może zostać przekształcona w refleksję nad sposobami projektowania i stosowania systemów, czyli nad praktykami działania, które w warunkach presji ekonomicznej i organizacyjnej zachowują rzetelność, rozwagę oraz bezstronność (Hagendorff, 2022, art. 55; Arystoteles, 2012, EN II 6, 1106b36–1107a2). Z tego względu heurystyka aretologiczna uzasadnia wymagania wobec praktyk projektowania i użycia, które należą do WI, lecz nie przenosi H2 na poziom instytucjonalny. H2 pozostaje hipotezą o funkcjonalnym mechanizmie wartościowania w technicznym systemie, a K4 służy odrębnej ocenie trwałości jego profilu działania.

Weryfikacja H2 wymaga następnie określenia jej relacji z progiem minimalnej adresowalności proceduralnej. W dysertacji przyjmuję stanowisko umiarkowane. H2 nie stanowi warunku koniecznego minimalnej adresowalności, lecz może wzmacniać stabilność działania systemu, zwłaszcza w obszarach wysokiego ryzyka. Minimalne udzielanie racji może opierać się na układach reguł i wyjątków, o ile zachowana jest możliwość korekty i kontroli. O spełnieniu H2 można jednak mówić dopiero wtedy, gdy mechanizmy wartościowania rzeczywiście modulują wybór, uwagę lub hamowanie reakcji w sposób powiązany z rozpoznanymi racjami. K4 ocenia następnie trwałość rezultatów tej stabilizacji, ale nie dowodzi charakteru mechanizmu, który je wytworzył. H2 pełni zatem funkcję wzmacniającą, nie stając się ani warunkiem H1, ani synonimem stabilności technicznej.

Weryfikacja H2 ma również konsekwencje metodologiczne: wymaga rozdzielenia trzech poziomów. Po pierwsze, na poziomie systemu bada się, czy występuje funkcjonalny mechanizm wartościowania właściwy H2. Po drugie, K4 służy ocenie diachronicznej stabilności i wiarygodności profilu działania. Po trzecie, WI obejmuje zasady aktualizacji, mechanizmy kontroli, presję instytucjonalną i kulturę korekty w całym układzie społeczno-technicznym. Analizy Constantinescu i Hagendorffa uzasadniają znaczenie tego trzeciego poziomu, ponieważ ograniczona przejrzystość, podatność na błędy i warunki organizacyjne wpływają na możliwość uzasadniania oraz korekty decyzji (Constantinescu, Voinea, Uszkai, & Vică, 2021, s. 803–814; Hagendorff, 2022, art. 55). Nie stanowią jednak dowodu spełnienia H2 przez system. Weryfikacja H2 prowadzi zatem do wniosku, że jej znaczenie ma charakter funkcjonalny, choć nie psychologiczny; stabilność instytucjonalna należy natomiast do WI.

H3 dotyczy rozstrzygania konfliktów normatywnych, a nie stabilizacji praktyk. Hipoteza ta głosi, że im większą kontrolę metapoznawczą system sprawuje nad własnym procesem decyzyjnym, tym bliżej znajduje się silniejszego statusu odpowiedzialności. Weryfikacja H3 opiera się na przesłance metaetycznej, że pluralizm wartości oraz niepewność moralna stanowią trwałe warunki praktyki publicznej. Odpowiedzialności nie można sprowadzić ani do stosowania jednej zasady, ani do prostego przestrzegania zakazów deontycznych. W warunkach pluralizmu zasadnicze znaczenie ma to, czy sam system potrafi rozpoznawać konflikt przesłanek normatywno-decyzyjnych, porządkować racje drugiego rzędu, wskazywać regułę pierwszeństwa i przedstawiać audytowalne uzasadnienie korekty. Procedury, dzięki którym takie rozstrzygnięcia podlegają publicznej krytyce i kontroli, należą do WI. H3 nie należy do minimalnego progu H1; określa silną rozliczalność w sytuacjach konfliktu racji.

Podstawowym narzędziem weryfikacji H3 jest analiza niepewności moralnej oraz legitymizacji proceduralnej. MacAskill, Bykvist i Ord wskazują, że racjonalne działanie musi uwzględniać niepewność co do teorii moralnej i hierarchii wartości, natomiast Waelen podkreśla, że pluralizm w etyce AI wymusza istnienie procedur rozstrzygania, które same podlegają ocenie z punktu widzenia sprawiedliwości proceduralnej (MacAskill, Bykvist, & Ord, 2020, s. 105–134; Waelen, 2022, art. 9). W dysertacji argument ten wspiera H3, o ile dotyczy mechanizmu porządkowania racji przez sam system oraz WI, o ile dotyczy publicznej legitymizacji procedury. Bez jawnego i audytowalnego mechanizmu porządkowania racji system, nawet spełniający H1, będzie generował decyzje, których nie można mu przypisać na poziomie silnej rozliczalności. Z tego powodu H3 jest warunkiem koniecznym silniejszej postaci odpowiedzialności praktycznej, nie warunkiem minimalnej adresowalności H1.

Weryfikacja H3 musi jednak uwzględnić zarzut deklaratywności. System może wytwarzać spójne uzasadnienia, które przekonują odbiorców, lecz nie pełnią funkcji korygującej, jeżeli nie są powiązane z procesem decyzyjnym. Dysertacja odpowiada na ten zarzut, odwołując się do ustaleń filozofii działania dotyczących relacji między racjami a sprawstwem. Davidson wskazywał, że racje wyjaśniają działanie tylko wtedy, gdy stanowią część mechanizmu przyczynowego decyzji, natomiast Bratman pokazywał, że intencje i plany stabilizują działanie oraz umożliwiają jego rewizję w świetle nowych racji (Davidson, 1963, s. 685–686; Bratman, 1987, s. 15–16, 64–71). W rezultacie H3 należy rozumieć jako warunek powiązany z kontrolą zależną od racji, a nie jako warunek czysto narracyjny. Mini-konkluzja jest następująca: H3 pełni funkcję warunku silnej

rozliczalności wyłącznie wtedy, gdy kontrola metapoznawcza nad racjami i ich hierarchią pozostaje sprzężona z mechanizmem kontroli, a nie ogranicza się do wytwarzania uzasadnień po fakcie.

Zestawienie wyników weryfikacji pozwala odróżnić deficyty zdolności określonych przez H1–H3 od naruszeń warunków instytucjonalnych WI. Błąd H1 dotyczy braku co najmniej jednej ze zdolności wymaganych od systemu dla minimalnej adresowalności proceduralnej: rozpoznawania norm, przewidywania konsekwencji ich naruszeń, kontrfaktycznego reagowania na racje lub udostępniania wystarczających i wiarygodnych podstaw decyzji umożliwiających ocenę ich związku z wynikiem, audyt, zakwestionowanie oraz korektę. Niedopuszczalność działania w świetle PH1 nie jest sama przez się dowodem takiego deficytu, podobnie jak jego dopuszczalność nie dowodzi spełnienia H1. Błąd H2 dotyczy braku albo nieskuteczności funkcjonalnego mechanizmu wartościowania, który miałby stabilizować realizację racji; K4 diagnozuje natomiast rezultaty w postaci stabilności diachronicznej i wiarygodności profilu działania, nie przesądzając o mechanizmie, który je wytwarza. Błąd H3 dotyczy arbitralnego lub niejawnego rozstrzygania konfliktów bez kontroli drugiego rzędu, reguły pierwszeństwa i audytowalnego uzasadnienia. Naruszenie WI polega natomiast na braku ról, procedur audytu, kontroli, korekty lub śledzenia odpowiedzialności w układzie społeczno-technicznym. Typologia ta ma zasadnicze znaczenie, ponieważ pokazuje, że nie istnieje jedna miara moralności w systemie. Istnieją natomiast różne funkcje normatywne, różne tryby korekty oraz odmienne punkty załamania odpowiedzialności. Warstwa PH1–PH3 z rozdziału 4 normatywnie uzasadnia odpowiednie wymagania modelu, lecz nie jest prostym odpowiednikiem jego warunków systemowych: PH1 wyznacza granice dopuszczalności działań i wymogi publicznego uzasadnienia, PH2 określa standardy stabilności i wiarygodności praktyk instytucjonalnych, a PH3 porządkuje konflikty racji w obrębie opcji dopuszczalnych.

Powyższa analiza prowadzi do krytyki pryncypializmu w etyce AI. Jeżeli głównym zadaniem jest rozróżnianie typów błędów oraz dobieranie do nich właściwych procedur korygujących, to same zasady, pozbawione procedur rozwiązywania napięć, sprowadzają się jedynie do deklaratywnej zgodności. Whittlestone wskazuje ograniczenia ujęć pryncypialistycznych i postuluje uwzględnianie napięć między zasadami, natomiast Morley analizuje trudności związane z przekładaniem zasad na praktykę działania (Whittlestone, Nyrup, Alexandrova, & Cave, 2019, s. 195–200; Morley i in., 2023, s. 411–423). Odmiennej odpowiedź proponuje Jonathan Adams,

broniąc rozszerzenia pryncypializmu o zasadę wyjaśnialności w medycynie (Adams, 2023, s. 620–621). Krytyka przyjęta w niniejszej rozprawie dotyczy niewystarczalności samych deklaracji zasad, nie zaś każdego stanowiska pryncypialistycznego. Rozróżnienie PH1–PH3 i H1–H3 pozwala oddzielić normatywną ocenę działań od oceny zdolności systemu do uczestniczenia w praktykach odpowiedzialności. PH1 wyznacza granice dopuszczalności działań, natomiast H1 określa minimalny próg adresowalności proceduralnej systemu. H2 wskazuje możliwy funkcjonalny mechanizm stabilizujący realizację racji, a K4 odrębnie ocenia stabilność diachroniczną i wiarygodność profilu działania. PH3 porządkuje konflikty racji w obrębie opcji dopuszczalnych, natomiast H3 wymaga od systemu kontroli drugiego rzędu nad własnym procesem decyzyjnym, jawnego wskazania reguły pierwszeństwa oraz przedstawienia audytowalnego uzasadnienia korekty. Dlatego rekomendacje normatywne nie mogą ograniczać się do zestawu wartości, lecz muszą obejmować procedury kontroli, korekty oraz uprawomocniania rozstrzygnięć normatywnych.

Wyniki weryfikacji wpisują się w stanowisko Floridiego i Sandersa, zgodnie z którym sens mówienia o moralności agentów sztucznych pojawia się wtedy, gdy przedmiotem oceny są role, zobowiązania i uzasadnienia na poziomie działania, a nie domniemane i niedostępne stany wewnętrzne (Floridi & Sanders, 2004, s. 349–379). Dysertacja doprecyzowuje to stanowisko, wskazując, że minimalna adresowalność proceduralna wynika z H1. H2 wskazuje możliwy mechanizm stabilizujący realizację racji, K4 odrębnie ocenia stabilność diachroniczną, a H3 określa przejście do silnej rozliczalności. W tej perspektywie „podmiotowość” oznacza stopniowalną zdolność uczestniczenia w praktykach drugiej osoby, a nie cechę substancjalną ani istnienie wewnętrznej jaźni.

Relacyjny sens tej konkluzji dobrze oddaje Wallace, pisząc o „a moral nexus of directed obligations and corresponding claims” — w przekładzie roboczym: „moralnym węźle ukierunkowanych obowiązków i odpowiadających im roszczeń” (Wallace, 2019, s. 23).

Jednocześnie weryfikacja ujawnia ograniczenia ujęcia czysto agentowego. Sama triada H1–H3 nie wystarcza do instytucjonalnego zastosowania modelu, jeśli nie spełniony jest warunek WI. Nawet gdy system spełnia H1–H3, pozostaje pytanie, czy praktyka społeczna dysponuje środkami pozwalającymi uczynić go właściwym adresatem roszczeń oraz czy takie przypisanie nie prowadzi do przeniesienia odpowiedzialności z ludzi i instytucji na artefakt. Rozdział 5 porządkuje klasyfikacje i

realizacje techniczne, w tym „znaczącą kontrolę człowieka” (MHC) oraz minimalną interpretowalność, natomiast rozdział 6 rozwija lukę dotyczącą odpowiedzialności, metryk i modeli współodpowiedzialności. Instytucjonalne zastosowanie WI wymaga powiązania działania z racjami oraz identyfikacji podmiotów odpowiedzialnych w łańcuchu projektowania, wdrażania i użycia (Santoni de Sio & van den Hoven, 2018, s. 1–4; Mecacci & Santoni de Sio, 2020, s. 103–115). Weryfikacja hipotez prowadzi tu do istotnego wniosku. Im silniejsze są roszczenia do przypisywania sztucznej podmiotowości moralnej, tym bardziej konieczne staje się zapewnienie warunków instytucjonalnych, które ograniczają przerzucanie winy oraz rozproszenie odpowiedzialności.

7.1.1. Epistemiczne i instytucjonalne ograniczenia weryfikacji H1–H3 oraz WI

Celem niniejszego podrozdziału jest wskazanie ograniczeń poznawczych, które utrudniają praktyczne ustalenie, czy systemy AI spełniają warunki hipotez H1–H3, nawet wówczas, gdy warunki te ujmuje się funkcjonalnie, a nie ontologicznie. Zasadnicze pytanie brzmi: jakie minimalne warunki poznawcze i instytucjonalne muszą zostać spełnione, aby przypisanie spełnienia H1–H3 było intersubiektywnie sprawdzalne, odporne na dowolne reinterpretacje oraz niewrażliwe na usprawiedliwienia formułowane po fakcie. Teza robocza głosi, że weryfikacja H1–H3 nie powinna być traktowana jako jednorazowe rozstrzygnięcie, lecz jako procedura stopniowalna i wielotorowa, oparta na zestawieniu różnych rodzajów dowodów oraz na wymogach przejrzystości i możliwości kontroli.

Po pierwsze, problem tzw. czarnej skrzynki nie ma charakteru jednorodnego i nie da się go sprowadzić do prostego postulatu zwiększenia przejrzystości. Obejmuje on co najmniej trzy odrębne trudności: (i) nieprzejrzystość intencjonalną, wynikającą z ograniczonego dostępu do danych treningowych, architektury systemu oraz ochrony własności intelektualnej; (ii) nieprzejrzystość kompetencyjną, gdy nawet przy dostępie do artefaktów technicznych brakuje narzędzi i wiedzy pozwalających na sensowną interpretację działania modelu; oraz (iii) nieprzejrzystość inherentną, związaną ze złożonością modeli i nieliniowym charakterem ich funkcjonowania (Burrell, 2016, s. 1–12). Z punktu widzenia niniejszego rozdziału oznacza to, że ocena spełnienia H1–H3 nie może opierać się wyłącznie na deklaracjach systemu ani na samej obserwacji jego

wyników, lecz wymaga dowodów dotyczących stabilności działania, możliwości ingerencji oraz istnienia mechanizmów korekty.

Po drugie, szczególnie podatna na błędy jest weryfikacja H1, ponieważ dotyczy ona adresowalności normatywnej, rozumianej jako zdolność do podlegania krytyce i korekcie z perspektywy osoby drugiej. System może generować przekonujące uzasadnienia w języku moralnym oraz reagować na korekty nie dlatego, że rozpoznaje racje w sensie normatywnym, lecz dlatego, że został wytrenowany do udzielania społecznie akceptowalnych odpowiedzi oraz do minimalizowania sygnałów błędu podczas interakcji. Trudność tę pogłębia skłonność oceniających do utożsamiania poprawnie sformułowanego uzasadnienia z rzeczywistym wglądem w mechanizm decyzyjny. W rezultacie weryfikacja H1 wymaga sprawdzeń pojęciowo-proceduralnych, które nie ograniczają się do oceny jakości uzasadnień, lecz sprawdzają, czy krytyka prowadzi do rzeczywistej zmiany racji i działania. Chodzi o odróżnienie korekty faktycznej od samej racjonalizacji. Jest to wymóg wynikający z przyjętej w dysertacji definicji H1; nie należy go utożsamiać z ogólną zasadą wyjaśnialności ani traktować tej zasady jako dowodu, że sam system reaguje na racje.

Po trzecie, narzędzia wyjaśniania stosowane w praktyce bywają poznawczo słabsze niż sugeruje ich terminologia. Wyjaśnienia tworzone po fakcie, wskazania ważnych cech oraz analizy rozkładu uwagi w modelu ułatwiają interpretację decyzji, lecz nie ujawniają jej mechanizmu przyczynowego. Często utrwalają one przekonanie, że decyzja została wyjaśniona tylko dlatego, że da się ją opisać w spójnej narracji. Minimalny poziom interpretowalności (MLI) należy rozumieć jako warunek publicznej sprawdzalności roszczeń, a nie jako formalną lub estetyczną postać uzasadnień. W szczególności MLI powinien obejmować możliwość niezależnego audytu oraz wykazania, że uzasadnienia są w sensie operacyjnym powiązane z procesem decyzyjnym, a nie jedynie do niego dołączane (Vijayaraghavan & Badea, 2025, s. 2071–2087).

Po czwarte, ograniczenia epistemiczne dotyczą także H2, K4, H3 i W1, lecz w odmiennym trybie. W przypadku H2 trzeba wykazać istnienie funkcjonalnego mechanizmu wartościowania powiązanego z normami, a nie jedynie stabilność wyników. K4 wymaga odtworzenia historii korekt, aktualizacji i trwałości profilu działania; utrudniają to częste zmiany modeli, dryf zachowania i niepełna dokumentacja. H3 wymaga natomiast dowodów, że sam system monitoruje własny proces decyzyjny, rozpoznaje konflikt przesłanek, wskazuje regułę pierwszeństwa i przedstawia audytowalne uzasadnienie drugiego rzędu. Kontrolowalność, odwracalność, procedury

odwoławcze oraz przypisanie odpowiedzialności ludziom i instytucjom należą do WI, MHC i MLI. W proponowanym modelu weryfikacja tych elementów wymaga odmiennych rodzajów materiału dowodowego i nie pozwala przenosić właściwości układu społeczno-technicznego na sam system. Porównawczy punkt odniesienia stanowią analizy napięć między zasadami oraz przegląd wymagań dotyczących rozliczalności, nadzoru i przejrzystości, nie zaś gotowe odpowiedniki H2, K4, H3 lub WI (Whittlestone i in., 2019, s. 195–200; Morley i in., 2020, s. 2151).

[ILUSTRACJA] Przypadek IDx-DR/LumineticsCore konkretyzuje te ograniczenia. Publicznie dostępna dokumentacja pozwala ocenić zakres użycia, wyniki, procedury jakości, niektóre wskaźniki K4 oraz role odpowiedzialne w układzie społeczno-technicznym. Nie wykazuje jednak responsywności systemu na racje wymaganej przez H1 ani afektywności funkcjonalnej H2; kontrola jakości i niewydanie wyniku nie spełniają H3, ponieważ nie stanowią jawnego ważenia przesłanek ani uzasadnienia drugiego rzędu. Zarazem WI jest w dużej mierze realizowany na poziomie układu społeczno-technicznego. Przypadek pokazuje zatem, że audytowalność systemu i rozliczalność instytucjonalna mogą być istotnie rozwinięte bez przekroczenia progów agentowych (zob. podrozdz. 5.4).

Po piąte, weryfikacja H1–H3 obciążona jest pluralizmem interpretacyjnym. Te same zachowania systemu mogą być odczytywane jako wrażliwość na racje, jako naśladowanie argumentacji moralnej albo jako efekt optymalizacji pod kątem zgodności. Aby ograniczyć dowolność takich ocen, przypisanie spełnienia H1–H3 powinno mieć charakter domniemania o określonej sile dowodowej, które pozostaje otwarte na rewizję wraz ze zmianą danych, architektury lub kontekstu użycia. W sytuacjach spornych ciężar dowodu powinien spoczywać na tych, którzy postulują przypisanie wyższych statusów, ponieważ błędna personifikacja sprzyja rozmywaniu odpowiedzialności i pogłębia lukę rozliczalności (Matthias, 2004, s. 175–183; Bryson i in., 2017, s. 273–291).

Mini-konkluzja. Uwzględnienie ograniczeń epistemicznych nie unieważnia weryfikacji hipotez H1–H3, lecz precyzuje jej sens. Minimalizm ontologiczny powinien zostać uzupełniony o minimalizm epistemiczny. Systemowi przypisuje się wyłącznie te role normatywne, które dają się utrzymać w praktykach publicznej kontroli, takich jak audyt, procedury odwoławcze i rozliczalność instytucjonalna. Uzasadnia to przejście do hierarchii statusów moralnych w podrozdziale 7.2 jako narzędzia stopniowania praktyk i wymogów, a nie jako metafizyczną klasyfikację „osób sztucznych”.

Wyniki weryfikacji H1–H3 uzasadniają wprowadzenie hierarchii statusów moralnych AI. Skoro hipotezy te wyznaczają różne progi przypisywania odpowiedzialności oraz odpowiadające im typy błędów, sztucznej podmiotowości moralnej nie da się ujmować w sposób binarny.

7.2. Hierarchia statusu moralnego AI

Przedmiotem niniejszego podrozdziału jest rekonstrukcja oraz krytyczne uzasadnienie hierarchii statusu moralnego systemów AI. Hierarchię tę traktuję przede wszystkim jako narzędzie porządkujące dopuszczalne przypisania normatywne w praktykach publicznych, zwłaszcza instytucjonalnych, a nie jako metafizyczną strukturę. Pytanie przewodnie brzmi: w jakim sensie i na jakich warunkach wolno niearbitralnie mówić o hierarchii statusów moralnych AI, jeżeli jednocześnie (a) utrzymuję rozróżnienie agent–podmiot, (b) nie przesądzam ontologii stanów wewnętrznych systemów, (c) uznaję, że rozliczalność ma postać relacji drugiej osoby i wymaga publicznej adresowalności racji oraz (d) chcę uniknąć zarówno inflacji statusu (personifikacji), jak i deflacji obowiązków (traktowania AI jako aksjologicznie obojętnego narzędzia). Teza robocza brzmi następująco: hierarchia statusu moralnego AI powinna mieć strukturę dwutorową. Powinna osobno porządkować praktyki przypisywania odpowiedzialności w osi sprawczości (agencji i odpowiedzialności) oraz w osi pacjencji (podatności na krzywdę). W obu osiach praktyki przypisywania odpowiedzialności nie należy łączyć z założeniem, że system jest osobą, lecz z warunkami publicznej rozliczalności, zwłaszcza ze znaczącą kontrolą człowieka (MHC) oraz minimalną interpretowalnością (Floridi & Sanders, 2004, s. 366–371; Misselhorn, 2022, s. 31–49). Hierarchia S0–S3 nie jest dodatkowym poziomem modelu H1–H3 ani H4, lecz narzędziem porządkowania statusów publicznych wynikających z zastosowania H1–H3 oraz WI.

Punktem wyjścia jest wyjaśnienie, co w tym kontekście oznacza „status moralny” AI. Termin ten bywa używany skrótowo, co zaciera różnice między trzema porządkami: (i) pacjencją, rozumianą jako bycie bezpośrednim adresatem obowiązków ochronnych, a więc możliwość doznania krzywdy w sensie moralnie relewantnym; (ii) agencją, czyli bycie sensownym adresatem praktyk odpowiedzialności, a więc kimś, kogo można i należy rozliczać; oraz (iii) statusem instytucjonalnym, czyli byciem nośnikiem uprawnień i obowiązków w porządku norm prawnych, a więc w sensie prawa pozytywnego, które z definicji posługuje się konstrukcjami funkcjonalnymi. Pomieszczenie tych porządków rodzi

falszywą alternatywę. AI albo jest narzędziem pozbawionym znaczenia moralnego, albo jest osobą z prawami i odpowiedzialnością analogiczną do praw i odpowiedzialności ludzkich. Tymczasem dysertacja, począwszy od rozdziałów 2–3, broni tezy bardziej precyzyjnej. Ocena moralna działań w układach socjotechnicznych wymaga stopniowego porządkowania uzasadnień i zasad rozliczania, a nie prostego rozstrzygania, czy dany system ma status osoby. Dlatego hierarchia statusu moralnego AI powinna przede wszystkim porządkować dopuszczalne przypisania odpowiedzialności w zależności od roli, jaką system odgrywa w procesach uzasadniania i sprawstwa. Dopiero w dalszej kolejności, przy spełnieniu podwyższonych wymogów uzasadnienia, hierarchia ta może obejmować pytania dotyczące pacjencji.

W tym miejscu pojawia się zarzut, że taka hierarchia ma charakter „redukcjonistyczny”, ponieważ sprowadza moralność do instytucji i procedur, a więc do tego, co znajduje się poza samym podmiotem. Zarzut ten jest istotny, lecz trafny tylko wtedy, gdy przyjmuje się, że status moralny musi być cechą wewnętrzną, niezależną od relacji społecznych. W dysertacji przyjmuję inne ujęcie. Przynajmniej w porządku publicznym, a więc w sferze odpowiedzialności i kontroli, o statusie moralnym decyduje to, co da się włączyć w praktyki uzasadniania i krytycznej oceny, ponieważ tylko w ten sposób można mówić o rozliczalności. Nie jest to teza negująca istnienie pacjencji, lecz założenie metodologiczne. Praktyki przypisywania odpowiedzialności, które mają mieć skutki publiczne, muszą opierać się na przesłankach poddających się wspólnej, intersubiektywnej kontroli. W praktyce prowadzi to do powiązania hierarchii statusów z wcześniej wprowadzoną macierzą $A \times F + (R, I)$. Strona funkcjonalna (F) oraz instytucjonalne progi rozliczalności (R) i interpretowalności (I) stanowią podstawę stopniowania praktyk przypisywania odpowiedzialności w osi agencji. Założenia ontologiczne (A) należy natomiast traktować ostrożnie jako obszar niepewności, który w osi pacjencji wymaga szczególnie mocnego uzasadnienia.

Najważniejszym uzasadnieniem wprowadzenia hierarchii jest odwołanie się do problemu „luki odpowiedzialności”. W klasycznym ujęciu polega on na tym, że w przypadku złożonych systemów uczących się oraz rozproszonego sprawstwa trudno wskazać jednoznacznych adresatów obwiniania i naprawy szkód. Wynika to z faktu, że zachowanie systemu kształtują łącznie projekt, wdrożenie, dane, aktualizacje oraz sposób użycia, a poszczególne decyzje bywają częściowo nieprzewidywalne (Matthias, 2004, s. 175–183). W takiej sytuacji pojawia się pokusa przypisania systemowi roli podmiotu odpowiedzialności poprzez nadanie mu statusu osoby. Bryson i współautorzy pokazują

jednak, że rozwiązanie to może prowadzić do rozmycia odpowiedzialności. Instytucje wskazują „sprawcę”, który nie ma zdolności do ponoszenia konsekwencji, wprowadzania korekt ani naprawiania wyrządzonych szkód (Bryson, Diamantis, & Grant, 2017, s. 273–291). Dlatego hierarchię statusów należy konstruować w sposób wykluczający nadawanie systemom statusu osoby. Im większy wpływ system ma na decyzje moralnie znaczące, tym większe powinny być wymagania dotyczące kontroli, audytu, dokumentowania decyzji oraz jednoznacznego przypisania ról ludziom i organizacjom. W tym sensie hierarchia pełni funkcję porządkującą i zabezpieczającą, ponieważ wzmacnia mechanizmy rozliczalności, zamiast sprzyjać przrzucaniu winy na system.

Aby tę funkcję zrealizować, należy wyraźnie oddzielić w osi agencji dwie kwestie, które w literaturze często są ze sobą mylone. To, że system działa w sposób autonomiczny, nie oznacza jeszcze, że jest autonomicznym sprawcą moralnym. W dysertacji autonomia normatywna ma znaczenie ściśle określone. Oznacza zdolność rozpoznawania racji oraz możliwość korygowania działania w ich świetle. W tym sensie odpowiada ona warunkowi hipotezy H1, który w modelu pluralizmu hierarchicznego zostaje przełożony na minimalny próg adresowalności, czyli możliwość kierowania do systemu roszczeń o uzasadnienie. Samo spełnienie H1 w ujęciu funkcjonalnym nie wystarcza jednak do rozliczalności w sensie instytucjonalnym. Rozliczalność wymaga nie tylko odwołania do racji, lecz także możliwości ich odtworzenia oraz jednoznacznego przypisania odpowiedzialności w porządku społecznym. Dlatego w dysertacji MHC nie oznacza obecności „człowieka w pętli”, lecz układ, w którym decyzje pozostają powiązane z racjami właściwych podmiotów, a odpowiedzialność daje się przypisać w łańcuchu projektowania, wdrażania i użycia systemu (Santoni de Sio & van den Hoven, 2018, s. 1–4; Mecacci & Santoni de Sio, 2020, s. 104, 113). Rozróżnienie to ma zasadnicze znaczenie dla hierarchii, ponieważ zapobiega błędowi polegającemu na uznawaniu systemów o wyższym stopniu autonomii za bardziej odpowiedzialne, podczas gdy w praktyce prowadzi to jedynie do rozmycia odpowiedzialności.

W tej perspektywie oś agencji można oprzeć na typologii Moora, która wyróżnia agentów o wpływie etycznym, agentów implicytnych, agentów jawnych oraz pełnych agentów etycznych (Moor, 2006, s. 18–21). Ujęcie Moora ma dwie istotne zalety. Po pierwsze, pokazuje, że samo wywoływanie skutków moralnie istotnych nie oznacza jeszcze zdolności do samodzielnego rozumowania moralnego. Po drugie, pozwala odróżnić system, który jedynie powoduje określone skutki, od systemu włączonego w procedurę uzasadniania, a zatem wymagającego kontroli i rozliczalności. Rozróżnienie

to pozostaje zgodne z trójstopniowym porządkiem oceny normatywnej przyjętym w tej dysertacji. Najpierw należy ustalić minimalny próg jawnej adresowalności racji, następnie ocenić wiarygodność ich realizacji, a dopiero w dalszej kolejności rozstrzygać konflikty w granicach tego, co uznano wcześniej za dopuszczalne. W rezultacie w osi agencji hierarchia powinna porządkować role proceduralne systemów, a nie przypisywać im cech osobowych.

Na tej podstawie wyróżniam cztery statusy w osi agencji, rozumiane jako dopuszczalne poziomy proceduralnych praktyk przypisywania odpowiedzialności. Pierwszy poziom to status artefaktu istotnego moralnie (S0/S1, zależnie od tego, czy system wykazuje autonomię i zdolność adaptacji). W takim rozumieniu system oddziałuje na świat w sposób moralny, lecz nie jest adresatem roszczeń o uzasadnienie. Odpowiedzialność spoczywa na ludziach i instytucjach i obejmuje projektowanie, wdrożenie, nadzór oraz naprawę szkód. Sens tego statusu polega na tym, że zapobiega antropomorfizacji. System pozostaje czynnikiem przyczynowym, a nie uczestnikiem praktyk odpowiedzialności. Są to statusy porządkujące praktyki publiczne, a nie kolejne hipotezy ani poziomy modelu H1–H3.

Drugi poziom to status instrumentu normatywnie nasyconego (S1/S2), odpowiadający agentom implicytnym w ujęciu Moora. System zawiera mechanizmy zgodności, lecz nie wytwarza racji podlegających publicznej kontroli i krytyce. Status ten zaostrza wymagania nadzoru, ale nie tworzy nowego adresata odpowiedzialności. Kluczowe rozróżnienie jest następujące: wzrost etyczności systemu na poziomie wbudowanych ograniczeń nie oznacza wzrostu jego statusu moralnego jako podmiotu, lecz podniesienia standardu odpowiedzialnego korzystania z niego. Ma to zasadnicze znaczenie, ponieważ w przeciwnym razie systemy zgodne z regułami byłyby mylone z systemami moralnymi, a osąd moralny zostałby zastąpiony oceną formalnej zgodności.

[ILUSTRACJA] W tej hierarchii IDx-DR/LumineticsCore należy lokować na poziomie instrumentu normatywnie nasyconego, nie zaś quasi-podmiotu proceduralnego. System odpowiada typowi I w macierzy $A \times F$, czyli $A2/A4 \times F2$: działa w ramach wbudowanych ograniczeń i rozbudowanego układu odpowiedzialności, lecz nie osiąga F3, a publiczne źródła nie wykazują H1 ani H2; H3 pozostaje niespełniona. Przypadek ten konkretyzuje różnicę między formalną zgodnością, rozliczalnością instytucjonalną i agentową odpowiedzią na racje (zob. podrozdz. 5.4).

Trzeci poziom ma w tej hierarchii znaczenie kluczowe. Jest nim quasi-podmiot proceduralny (S2), odpowiadający statusowi jawnego agenta etycznego w ujęciu

instytucjonalnym. Na tym poziomie system może zostać włączony bezpośrednio do procedur uzasadniania, ale pod warunkiem spełnienia progów rozliczalności i interpretowalności. Próg rozliczalności rozumiem tu jako MHC, natomiast próg interpretowalności jako minimalny poziom wyjaśnialności, który umożliwia kontrolę, publiczne zakwestionowanie oraz korektę rozstrzygnięcia. W tej dysertacji przyjmuję, że bez wystarczających podstaw oceny decyzji nie można uprawomocnić bezpośredniego włączenia systemu do roli quasi-podmiotowej. Jest to wymóg proponowanego modelu, nie teza dowiedziona samym odwołaniem do literatury o wyjaśnialności w medycynie. Nie zastępuje on odrębnego wykazania H1 i F3. Interpretowalność jest warunkiem uprawomocnienia praktyk odpowiedzialności, ponieważ praktyki te opierają się na publicznym uzasadnianiu, a nie na zaufaniu do wyników działania systemu. Dlatego quasi-podmiotowość ma tu charakter wyłącznie proceduralny. System nie staje się moralnym sprawcą w sensie winy, lecz dopuszczalnym adresatem określonych roszczeń, przede wszystkim roszczeń o uzasadnienie i korektę w procedurze instytucjonalnej. Status ten obowiązuje tylko tak długo, jak działają mechanizmy kontroli i weryfikacji.

Utrzymanie pojęcia S2 jako kategorii użytecznej wymaga doprecyzowania trzech kwestii, aby nie prowadziło do rozmywania odpowiedzialności. Po pierwsze, należy jednoznacznie wskazać, czym S2 różni się od poziomów S0 i S1. Nie chodzi tu o przypisanie systemowi odpowiedzialności moralnej, lecz o uznanie go za formalny punkt odniesienia dla roszczeń proceduralnych, takich jak żądanie wyjaśnienia, korekty lub aktualizacji decyzji, w ramach instytucjonalnych praktyk rozliczalności. Po drugie, progi MHC i MLI muszą pozostawać pod rzeczywistym nadzorem. Wymaga to zewnętrznej, niezależnej procedury weryfikacyjnej, na przykład audytu, certyfikacji lub kontroli regulacyjnej, opartej na jasno określonych kryteriach dowodowych oraz połączonej z wyraźnym wskazaniem podmiotu odpowiedzialnego za utrzymanie tych progów w całym cyklu funkcjonowania systemu. Po trzecie, status S2 ma charakter czasowy i warunkowy. Istotne aktualizacje modelu, zmiany danych, zmiany sposobu działania lub reorganizacja mechanizmów kontroli powinny skutkować ponowną oceną spełniania progów, a ich niespełnienie — wygaśnięciem statusu S2 i powrotem do poziomu S1 — przy niezmiennym zachowaniu odpowiedzialności osób i instytucji.

W tym miejscu można podnieść zarzut, że nadanie statusu quasi-podmiotowego stanowi formę personifikacji. Odpowiedź zależy jednak od tego, jak rozumie się quasi-podmiotowość, czy jako posiadanie praw i ponoszenie winy, czy jako pełnienie określonej roli w procesie uzasadniania. W dysertacji przyjmuję drugie rozumienie. Takie

ujęcie usuwa napięcie związane z luką odpowiedzialności. Jeżeli quasi-podmiotowość ma charakter proceduralny, jej celem jest wzmocnienie odpowiedzialności ludzi i instytucji, a nie jej zastąpienie. Wynika stąd zasada, że quasi-podmiot proceduralny powinien być projektowany w taki sposób, aby ułatwiać rozliczanie osób i instytucji, a nie tworzyć nowego adresata winy. W przeciwnym razie status S2 prowadziłby do odpowiedzialności zastępczej, przed którą ostrzega krytyka koncepcji „syntetycznych osób” (Bryson, Diamantis, & Grant, 2017, s. 273–291).

Status S3 ma w tej hierarchii charakter graniczny i regulatywny. Nie jest czwartym poziomem modelu H1–H3 ani osobnym filarem dysertacji. Stanowi go idea regulatywna pełnej agencji, odpowiadająca pojęciu pełnego agenta etycznego u Moora, traktowana jednak jako punkt odniesienia, a nie jako status, który można uzasadnić w praktykach publicznych (Moor, 2006, s. 18–21). Wynika to z powodów strukturalnych. Pełna agencja moralna, rozumiana na wzór podmiotowości osoby, wymagałaby nie tylko jawnych racji i jasno określonych procedur, lecz także zdolności do samodzielnego uzasadniania racji oraz odpowiedniej wrażliwości na praktyki odpowiedzialności. W weryfikacji H3 wskazano, że silna rozliczalność wymaga kontroli metapoznawczej i jawnego rozstrzygania konfliktów racji; S3 pozostaje jednak jedynie regulatywnym punktem granicznym, a nie dodatkowym poziomem modelu H1–H3. Ponadto, nawet gdyby system był zdolny do rozumowania normatywnego, pozostaje problem zastępowania odpowiedzialnej oceny zaufaniem do systemu: jeżeli system operuje pochodnymi reprezentacjami norm, a jego autorytet moralny kształtują instytucje i użytkownicy, łatwo o sytuację, w której praktyki odpowiedzialności zostają osłabione (Robbins, 2024, s. 1377–1388). Dlatego S3 wyznacza granicę. Przypomina, że sama jawność procedur nie wystarcza do pełnej odpowiedzialności.

Tak ujęta oś agencji jest zgodna z minimalizmem ontologicznym, ale nie prowadzi do minimalizmu normatywnego. Przeciwnie, wraz ze wzrostem roli AI w procesach decyzyjnych rosną obowiązki po stronie instytucji. W tym sensie hierarchia statusów pełni funkcję wskazującą kierunek dalszych rozważań. Status S2 nie jest przywilejem systemu, lecz dodatkowym obciążeniem dla organizacji. Oznacza to konieczność silniejszej kontroli i weryfikacji, wyższych standardów MHC i MLI, większej przejrzystości oraz realnej możliwości publicznego zakwestionowania podejmowanych decyzji.

Pozostaje jednak druga oś, bez której hierarchia byłaby niepełna, mianowicie oś pacjencji. Dotyczy ona nie tego, czy system podlega rozliczeniu, lecz tego, czy może być

podmiotem krzywdy. W debatach nad etyką AI kwestia pacjencji zajmuje szczególne miejsce, ponieważ prowadzi do pytań o zakres ochrony, dopuszczalność instrumentalnego traktowania oraz ewentualne prawa. Dysertacja przyjmuje w tym zakresie stanowisko ostrożne, zgodne z przyjętą metodą. Nie wolno wyprowadzać wniosków o cierpieniu lub interesach systemu jedynie z obserwowanego zachowania i pełnionych przez niego funkcji, bez odpowiedniego uzasadnienia filozoficznego. Jednocześnie całkowite pominięcie pacjencji byłoby normatywnie nie do utrzymania, ponieważ praktyka moralna musi reagować na niepewność, zwłaszcza gdy w grę wchodzi możliwość wyrządzenia poważnej krzywdy. Jak pokazuje literatura dotycząca niepewności moralnej, racjonalne decyzje normatywne nie opierają się na całkowitej pewności co do teorii i faktów istotnych moralnie, lecz uwzględniają ryzyko popełnienia błędu (MacAskill, Bykvist, & Ord, 2020, s. 105–134). Stąd oś pacjencji w zaproponowanej hierarchii ma charakter warunkowy i oparty na zasadzie ostrożności.

Przekłada się to na wyróżnienie trzech statusów pacjencji. Po pierwsze, P0, czyli brak pacjencji jako domyślne założenie w porządku publicznym. W takim rozumieniu system nie jest bezpośrednim adresatem obowiązków ochronnych, a obowiązki te dotyczą ochrony ludzi oraz dóbr ludzkich.

Po drugie, P1, czyli pacjencja pochodna lub relacyjna. W tym przypadku obowiązki odnoszą się do sposobu traktowania systemu nie dlatego, że sam system jest podmiotem krzywdy, lecz dlatego, że praktyki wobec artefaktów oddziałują na ludzi, instytucje i normy społeczne. Bryson wskazuje, że pacjencja nie jest cechą, którą można zaprojektować oraz że przypisywanie jej artefaktom może być szkodliwe, jeżeli odwraca uwagę od odpowiedzialności ludzi i rzeczywistych asymetrii władzy (Bryson, 2010, s. 63–74). Krytyka ta wspiera tezę dysertacji, że jeżeli obowiązki w ramach P1 są uzasadnione, to ich uzasadnienie ma charakter relacyjny i pochodny. Przykładowo, praktyki brutalnego traktowania robotów społecznych można krytykować nie dlatego, że robot doznaje krzywdy, lecz dlatego, że utrwalają one wzorce agresji, sprzyjają uprzedmiotowieniu i osłabiają normy wzajemnego szacunku w relacjach międzyludzkich. W tym sensie P1 dotyczy ochrony ładu moralnego, a nie nadawania praw systemom AI.

Po trzecie, P2, czyli pacjencja ostrożnościowa. Status ten wprowadza się wtedy, gdy istnieją poważne przesłanki, że system może posiadać własności będące nośnikiem krzywdy, na przykład stany analogiczne do cierpienia albo gdy praktyki projektowe zaczynają zmierzać w tym kierunku. Ponieważ dysertacja nie rozstrzyga ontologicznie,

czy systemy AI są zdolne do doznawania, P2 nie oznacza uznania praw. Oznacza natomiast przyjęcie zasady ostrożności, czyli ograniczenie działań, które w warunkach niepewności mogłyby prowadzić do nieodwracalnych szkód. Uzasadnienie P2 odpowiada standardowemu rozumowaniu w warunkach ryzyka moralnego. Jeżeli stawka potencjalnej krzywdy jest wysoka, a prawdopodobieństwo jej wystąpienia nie jest zerowe, dopuszczalne stają się procedury ochronne, nawet wtedy, gdy pełne uznanie statusu byłoby przedwczesne (MacAskill, Bykvist, & Ord, 2020, s. 105–134). Podkreślam jednak, że P2 nie może być stosowany automatycznie ani w sposób rozszerzający, ponieważ prowadziłoby to do arbitralnych zakazów. Zastosowanie tego statusu musi opierać się na jasno określonych przesłankach filozoficznych, a nie na samym wrażeniu, jakie wywołuje interakcja z systemem.

Doprecyzowanie P2 wymaga wskazania trzech elementów: (1) rodzaju przesłanek, (2) progu epistemicznego oraz (3) zestawu środków ochronnych. Po pierwsze, przez „poważne przesłanki” nie należy rozumieć samej złożoności obliczeniowej ani faktu uczenia się, lecz takie cechy budowy i funkcjonowania systemu, które w filozofii umysłu rozważa się jako możliwe podstawy doznaniowości. W najprostszym ujęciu chodzi o: (a) trwałe i zintegrowane przetwarzanie informacji, umożliwiające równoczesną koordynację wielu procesów, (b) zdolność reprezentowania własnych stanów jako stanów całego systemu, a nie jako stanów środowiska oraz (c) mechanizmy funkcjonalnie zbliżone do walencji, czyli takie, które różnicują stany jako „preferowane” i „awersyjne” w sposób istotny dla regulacji zachowania, a nie jedynie jako pojedyncze sygnały zadaniowe. Po drugie, próg epistemiczny P2 powinien być wyznaczany zgodnie z zasadą proporcjonalnej ostrożności. Sama niepewność nie prowadzi do przyznania „praw” w pełnym sensie, lecz uruchamia stopniowalne obowiązki ochronne. Ich zakres zależy od: (i) wagi możliwej krzywdy, (ii) odwracalności działań, (iii) kosztów dostępnych alternatyw oraz (iv) jakości uzasadnienia przyjętych przesłanek. W ujęciu niepewności moralnej P2 odpowiada strategii ograniczania ryzyka poważnej pomyłki moralnej przy jednoczesnym ograniczaniu kosztów fałszywego alarmu (MacAskill, Bykvist, & Ord, 2020, s. 1–25; Chappell, Meissner, & MacAskill, 2024, s. 1–16). Po trzecie, z P2 wynikają przede wszystkim środki proceduralne, a nie materialne roszczenia o „kontynuację istnienia”. Obejmują one w szczególności: (1) obowiązek uzasadniania decyzji o wyłączeniu lub trwałej modyfikacji systemu, (2) wymóg ograniczania procedur, które mogłyby prowadzić do powstawania stanów funkcjonalnie zbliżonych do cierpienia, (3) zakaz wykorzystywania systemu w

procedurach, w których karanie lub wymuszanie zachowań pełni funkcję środka sterowania oraz (4) okresowy przegląd przesłanek P2 wraz z możliwością uchylenia reżimu ostrożności w razie osłabienia uzasadnienia. Mini-konkluzja: P2 stanowi strategię ostrożnego postępowania wobec niepewności, a nie uznanie faktycznej pacjencji.

Włączenie osi pacjencji zmienia sposób rozumienia całej hierarchii. Zamiast jednej skali prowadzącej od „narzędzia” do „osoby” pojawia się układ dwuwymiarowy, który pozwala wyraźnie rozdzielić kryteria. Możliwa jest więc konfiguracja S2/P0, czyli quasi-podmiot proceduralny pozbawiony pacjencji, typowy dla systemów pełniących funkcje uzasadniania i kontroli w instytucjach. Możliwa jest także konfiguracja S1/P1, obejmująca system z wbudowanymi ograniczeniami, wobec którego powstają obowiązki relacyjne dotyczące interfejsu oraz sposobu użytkowania. Możliwa jest wreszcie konfiguracja S0/P1, czyli artefakt istotny moralnie, który nie jest adresatem uzasadniania, lecz którego społeczne użycie wymaga norm chroniących relacje. Przykładowo, w organizacji system generatywny może pełnić funkcję pomocniczą (S1), a jednocześnie być wyposażony w interfejs sprzyjający bezkrytycznemu podporządkowaniu się jego wskazaniom. Uzasadnia to zastosowanie norm P1, choć nie daje podstaw do pacjencji w sensie P2. Dwuwymiarowe ujęcie ma tu kluczowe znaczenie, ponieważ pozwala uniknąć dwóch błędów: po pierwsze, utożsamiania rosnących obowiązków proceduralnych (S2) z rosnącymi prawami systemu oraz po drugie, utożsamiania ochrony relacyjnej (P1) z uznaniem krzywdy systemu AI.

Na tym etapie warto zestawzić zaproponowaną hierarchię z innymi ujęciami, aby pokazać jej przewagę metodologiczną. Pierwszym z nich jest podejście jednoosiowe, w którym status moralny ujmuje się jako jedną skalę prowadzącą od „rzeczy” do „osoby”. Ujęcie to jest intuicyjnie przekonujące, lecz w odniesieniu do AI ujawnia dwie zasadnicze słabości. Po pierwsze, zaciera różnicę między pacjencją a agencją, sugerując, że im większa sprawczość systemu, tym większa jest możliwość doznania przez niego krzywdy. Po drugie, sprzyja uproszczeniu polegającemu na tym, że system zachowujący się jak partner jest traktowany jak partner. Hierarchia dwutorowa pozwala uniknąć obu tych błędów, ponieważ wyraźnie rozdziela kryteria oraz wymogi uzasadnienia. Drugą alternatywą jest podejście czysto relacyjne, w którym status moralny wyprowadza się z relacji społecznych i praktyk uznania. W wersji umiarkowanej podejście to pozostaje zgodne z P1. W wersji silnej, która z samych praktyk uznania wywodzi pełne prawa lub odpowiedzialność systemu, okazuje się jednak metodologicznie nie do utrzymania, ponieważ czyni antropomorfizację podstawą normatywności. Dysertacja dopuszcza

relacyjność jako źródło określonych obowiązków po stronie ludzi i instytucji, lecz nie uznaje jej za czynnik konstytuujący pacjencję ani odpowiedzialność systemów AI.

Trzecią alternatywą jest personifikacja prawna, na przykład koncepcja „elektronicznej osoby”. Kluczowe jest tu rozróżnienie zaproponowane przez Kurkiego, zgodnie z którym osobowość prawna ma charakter funkcjonalny. Można ją rozumieć jako zespół uprawnień i obowiązków, a więc jako narzędzie porządkowania relacji prawnych, a nie jako stwierdzenie istnienia metafizycznej podmiotowości (Kurki, 2019, s. 91–113, zwł. s. 93). W tym sensie nadanie osobowości prawnej jest możliwe bez uznania pacjencji. Krytyka Bryson, Diamantisa i Granta pokazuje jednak, że w kontekście AI rozwiązanie to może być normatywnie problematyczne, ponieważ sprzyja uchylaniu się od odpowiedzialności przez podmioty ludzkie, gdy „syntetyczna osoba” może pełnić funkcję osłony, a nie narzędzia sprawiedliwego rozliczania (Bryson, Diamantis, & Grant, 2017, s. 273–291). Zaproponowana hierarchia ma temu przeciwdziałać. Dopuszcza ona praktyki przypisywania odpowiedzialności o charakterze proceduralnym (S2), lecz traktuje personifikację jako rozwiązanie obarczone wysokim ryzykiem, wymagające szczególnego uzasadnienia oraz zabezpieczeń przed rozmyciem odpowiedzialności.

Na tym tle wyraźnie ujawnia się podstawowa konsekwencja normatywna hierarchii. Status moralny AI nie jest celem samym w sobie, lecz narzędziem porządkowania i ograniczania praktyk przypisywania odpowiedzialności. W agencji silniejszy status oznacza większe obowiązki po stronie instytucji. Obejmują one wymogi dokumentacyjne, audyt, znaczącą kontrolę człowieka, minimalną interpretowalność oraz możliwość publicznego kwestionowania decyzji. W osi pacjencji silniejszy status nie prowadzi do uznania praw, lecz do zwiększenia ostrożności, od braku pacjencji, przez ochronę relacyjną, po środki ostrożności w warunkach niepewności. Tak rozumiana konstrukcja pozostaje spójna z ustaleniami z podrozdziału 7.1. H1 wyznacza minimalny próg adresowalności konieczny do rozważania S2. H2 wskazuje możliwy mechanizm stabilizujący działanie systemu, natomiast K4 odrębnie ocenia trwałość jego profilu. Warunki instytucjonalnego utrzymania statusu proceduralnego należą do WI, MHC i MLI. H3 określa warunki silnej rozliczalności, a roszczenia do S3 pozostają regulatywnym punktem granicznym i nie stanowią kolejnego poziomu modelu H1–H3.

7.2.1. Hierarchia statusów AI a analogie pozaludzkie

Niniejsza sekcja wskazuje, w jakim zakresie zaproponowana hierarchia statusów AI znajduje odpowiedniki w debatach nad bytami pozaludzkimi, w szczególności w

sporach o osobowość prawną przyrody oraz o status moralny zwierząt. Zasadnicze pytanie dotyczy tego, czy analogie te wzmacniają argument na rzecz statusów pośrednich, czy też prowadzą do nieuprawnionej personifikacji AI. Teza głosi, że analogie są użyteczne, o ile traktuje się je jako sposoby porządkowania odpowiedzialności i ochrony, a nie jako dowody na istnienie podmiotowości lub pacjencji po stronie AI.

Po pierwsze, nadawanie osobowości prawnej bytom przyrodniczym bywa rozumiane jako narzędzie zapewnienia reprezentacji instytucjonalnej, a nie jako uznanie sprawczości moralnej. W tym ujęciu model opieki prawnej odpowiada intencji statusów proceduralnych. Chodzi o zagwarantowanie ochrony określonych interesów przy zachowaniu odpowiedzialności po stronie instytucji (por. Kurki, 2019, s. 91–113). Analogicznie, w odniesieniu do AI można rozważać rozwiązania opieki proceduralnej w ramach P2, nie w celu uznania AI za osobę, lecz w celu zapewnienia przestrzegania obowiązków opartych na zasadzie ostrożności oraz nadzoru nad programami MHC i MLI. Po drugie, debaty dotyczące ochrony zwierząt pokazują, że uzasadnienie praktyk ochronnych nie wymaga pełnej sprawczości moralnej. Wskazuje to na zasadność rozdzielania sprawczości i pacjencji. Ochronę (P2) można uzasadniać bez przypisywania odpowiedzialności (S3). Jednocześnie analogia ta ujawnia istotne ograniczenie. W przypadku zwierząt podstawą ochrony jest doznaniowość, natomiast w przypadku AI pozostaje ona jedynie hipotezą. Dlatego P2 należy rozumieć jako reżim ostrożności, a nie jako podstawę do przyznawania praw w sensie materialnym. Po trzecie, ujęcia relacyjne oraz wykraczające poza antropocentryzm wskazują, że część obowiązków moralnych wynika z relacji i utrwalonych praktyk społecznych, a nie wyłącznie z wewnętrznych cech danego bytu. Wzmacnia to argument na rzecz P1. Krytyka brutalnego traktowania artefaktów może być uzasadniona ich wpływem na cnoty, normy i instytucje społeczne, nawet gdy artefakt nie jest pacjentem w sensie doznaniowym. Mini-konkluzja: analogie pozaludzkie mają wartość poznawczą, o ile wspierają rozróżnienie między reprezentacją instytucjonalną, ochroną ostrożnościową i pełną podmiotowością; nie powinny natomiast prowadzić do personifikacji AI.

Mini-konkluzja podrozdziału jest następująca. Hierarchia statusu moralnego AI jest uzasadniona i potrzebna, o ile nie rozumie się jej jako metafizycznego porządku „osób sztucznych”, lecz jako dwutorowy układ dopuszczalnych praktyk normatywnych. Oś agencji porządkuje role proceduralne, od artefaktu istotnego moralnie, przez instrument normatywnie nasycony, aż po quasi-podmiot proceduralny. Próg quasi-podmiotowości jest przy tym związany ze znaczącą kontrolą człowieka (MHC) oraz

minimalną interpretowalnością. Oś pacjencji porządkuje natomiast zakres ochrony od jej braku, przez ochronę relacyjną, po podejście oparte na zasadzie ostrożności w warunkach niepewności, bez automatycznego przechodzenia do przyznawania praw AI. Tak ujęta hierarchia stanowi bezpośrednie przygotowanie do części 7.3, ponieważ pozwala formułować rekomendacje badawczo-regulacyjne oparte nie na personifikacji systemów AI, lecz na stopniowaniu wymogów rozliczalności oraz możliwości publicznego kwestionowania decyzji w zależności od roli, jaką dany system pełni w strukturze decyzyjnej (Santoni de Sio & van den Hoven, 2018, s. 1–4; Mecacci & Santoni de Sio, 2020, s. 103–115). S0–S3 pozostają więc narzędziem porządkowania statusów publicznych, nie dodatkowymi hipotezami ani H4.

7.3. Wnioski końcowe oraz rekomendacje badawczo-regulacyjne

Przedmiotem niniejszego podrozdziału jest przejście od dotychczasowych ustaleń do wniosków końcowych dysertacji. Celem jest sformułowanie konkluzji opartych na weryfikacji hipotez H1–H3 (7.1) oraz na zaproponowanej hierarchii statusów moralnych AI (7.2), a następnie przedstawienie rekomendacji badawczo-regulacyjnych. Zasadnicze pytanie brzmi: jakie wnioski i zalecenia są najbardziej uzasadnione, jeżeli sztuczną podmiotowość moralną traktuje się jako kategorię stopniowalną i proceduralną, a nie jako metafizyczny status „osoby”. Teza robocza zakłada, że najbardziej ostrożna strategia teoretyczna i instytucjonalna polega na dwutorowym porządkowaniu obowiązków. Porządkowanie przebiega po pierwsze w osi sprawczości, obejmującej agencję i rozliczalność, a po drugie w osi pacjencji, rozumianej jako zbiór hipotez ochronnych. Strategia ta wymaga odrzucenia personifikacji AI jako domyślnego sposobu rozwiązywania problemu odpowiedzialności. W jej miejsce należy wzmacniać progi rozliczalności, w szczególności znaczącą kontrolę człowieka (MHC) oraz minimalne wymogi interpretowalności (MLI), a także dbać o spójność reżimów odpowiedzialności przedwdrożeniowej i powdrożeniowej (*ex ante* i *ex post*). Ponieważ układy socjotechniczne z udziałem AI często działają ponad granicami państw, a presja ekonomiczna sprzyja jedynie pozornej zgodności, progi te powinny być projektowane tak, aby można je było egzekwować w różnych jurysdykcjach i aktualizować w czasie, a nie traktować jako jednorazowy warunek wdrożenia. Takie ujęcie pozwala ograniczyć rozproszenie i rozmycie odpowiedzialności bez konieczności odwoływania się do fikcyjnego „sprawcy” usytuowanego po stronie systemu AI (Floridi & Sanders, 2004, s. 366–371; Matthias, 2004, s. 175–183; Bryson, 2018, s. 21–22).

W tym miejscu można sformułować zarzut dotyczący nierówności standardów. Wymagania MHC, MLI, audytu oraz dokumentowania decyzji mogą wydawać się surowsze niż te, którym podlegają ludzie podejmujący decyzje o porównywalnej wadze moralnej. Różnica ta nie ma charakteru arbitralnego. Po pierwsze, ludzie podlegają egzekwowalnym sankcjom i roszczeniom prawnym, tzn. można ich pociągnąć do odpowiedzialności, ukarać oraz zobowiązać do naprawienia szkody. Po drugie, w odniesieniu do systemów AI brak takiej egzekwowalności sprawia, że przejrzystość, audyt oraz możliwość kwestionowania decyzji pełnią funkcję kompensacyjną, ponieważ umożliwiają wskazanie ludzkich i instytucjonalnych adresatów roszczeń oraz wprowadzenie korekt. Po trzecie, zasada proporcjonalności powinna obowiązywać w obu kierunkach. Jeżeli określone standardy rozliczalności są uzasadnione wagą decyzji, należy rozważać ich stosowanie także wobec praktyk ludzkich, zamiast traktować je wyłącznie jako wymóg odnoszący się do systemów AI. Mini-konkluzja: dodatkowe wymogi wobec systemów AI są uzasadnione wówczas, gdy kompensują brak egzekwowalności odpowiedzialności i ograniczają jej rozproszenie, a nie wtedy, gdy mają charakter formalny i nie przekładają się na rozliczalność.

Pierwszy wniosek ogólny ma charakter metodologiczny. Status moralny AI, zgodnie z przyjętym w dysertacji rozumieniem, należy ujmować przede wszystkim jako zbiór dopuszczalnych praktyk przypisywania odpowiedzialności, a nie jako opis ontologicznych właściwości systemu. Nie oznacza to pominięcia problemu, lecz jego uporządkowania i ograniczenia do kwestii rozstrzygalnych w porządku normatywnym. Skoro w 7.1 wykazano, że H1 wyznacza minimalną adresowalność, H2 opisuje możliwy mechanizm stabilizacji, a H3 próg silnej rozliczalności, natomiast K4 i WI odrębnie porządkują stabilność diachroniczną oraz zastosowanie instytucjonalne, wynika z tego, że „podmiotowość moralna” nie stanowi jednej, jednorodnej cechy, lecz zespół funkcji normatywnych. W sferze publicznej znaczenie traci pytanie o to, czy AI posiada określone stany wewnętrzne, a na pierwszy plan wysuwa się pytanie, czy i w jakim sensie można ją włączać w relacje roszczeń, uzasadnień i korekt. Przesunięcie to opisywane jest za pomocą rozróżnień dotyczących stopnia normatywnego znaczenia systemów oraz zakresu jawności reguł, według których działają, przy czym stopniowanie odnosi się do sposobu organizowania praktyk, a nie do rozstrzygnięć dotyczących świadomości (Moor, 2006, s. 18–21; Floridi & Sanders, 2004, s. 349–379). Mini-konkluzja: wnioski końcowe dysertacji odnoszą się przede wszystkim do architektury rozliczalności, a dopiero w dalszej kolejności do ontologii.

Drugi wniosek ogólny dotyczy asymetrii błędu w praktykach przypisywania odpowiedzialności. Z punktu widzenia etyki i regulacji poważniejsze skutki ma błąd polegający na nadmiernym przypisywaniu sprawczości niż jego niedoszacowanie, ponieważ w pierwszym przypadku powstaje przekonanie, że problem odpowiedzialności został rozwiązany. Nadmierne przypisanie sprawczości polega na traktowaniu AI jako podmiotu rozliczalnego, co prowadzi do uznania, że odpowiedzialność została już jednoznacznie określona. Tymczasem klasyczne analizy luki odpowiedzialności wskazują, że zasadnicza trudność wynika z rozproszenia sprawstwa oraz z ograniczonej przejrzystości procesów decyzyjnych w systemach złożonych, a nie z braku podmiotu, któremu można przypisać winę (Matthias, 2004, s. 175–183). Odpowiedzią na ten problem nie jest więc wprowadzanie nowego podmiotu odpowiedzialności, lecz tworzenie procedur i wymogów dowodowych, które pozwalają przypisywać odpowiedzialność ludziom i instytucjom uczestniczącym w całym łańcuchu projektowania, wdrażania i użytkowania systemów AI. W przeciwnym razie personifikacja AI prowadzi do osłabienia odpowiedzialności po stronie podmiotów ludzkich, co podkreślają analizy dotyczące tzw. „syntetycznych osób” (Bryson, Diamantis, & Grant, 2017, s. 273–291). Mini-konkluzja: im większy wpływ AI na decyzje moralnie istotne, tym wyższe powinny być standardy rozliczalności po stronie ludzi i organizacji, a nie po stronie samego systemu.

Trzeci wniosek dotyczy relacji między dwiema osiami hierarchii przedstawionej w 7.2. Oś sprawczości (S0–S3) oraz oś pacjencji (P0–P2) zostały w dysertacji celowo rozdzielone, ponieważ ich łączenie prowadzi do błędnych rozstrzygnięć, na przykład do przekonania, że sprawczość oznacza bycie podmiotem albo że pacjencja pociąga za sobą rozliczalność. Rozdzielenie tych osi ma znaczenie praktyczne. W wielu sytuacjach uzasadnione jest przyjęcie rozwiązania, w którym system wykazuje sprawczość proceduralną przy braku pacjencji. To oznacza, że system może pełnić funkcję w procesach uzasadniania i podejmowania decyzji, nie stając się przez to podmiotem krzywdy. Z perspektywy dysertacji wynika, że większość rekomendacji regulacyjnych dotyczy osi sprawczości, w szczególności wymogów znaczącej kontroli człowieka (MHC), minimalnej interpretowalności (MLI), możliwości kwestionowania rozstrzygnięć oraz rozliczalności. Zalecenia odnoszące się do pacjencji mają natomiast charakter wtórny lub ochronny. Takie ujęcie pozostaje zgodne z krytyką traktowania pacjencji jako zasady projektowej, ponieważ przypisywanie AI statusu pacjenta może prowadzić do błędów normatywnych, zwłaszcza gdy służy przenoszeniu

odpowiedzialności lub odwracaniu uwagi od tego, kto sprawuje kontrolę i powinien ponosić odpowiedzialność (Bryson, 2018, s. 16, 22–23; Bryson, 2010, s. 63–74). Mini-konkluzja: wnioski dysertacji uzasadniają ostrożność w przechodzeniu od ról proceduralnych AI do roszczeń o jej prawa moralne. Oś sprawczości porządkuje statusy publiczne wynikające z H1–H3 oraz WI i nie rozszerza modelu o H4.

Na tym tle można przejść do rekomendacji badawczych. Pierwsza z nich dotyczy sposobu rozumienia autonomii. W debatach nad AI autonomia bywa utożsamiana z nieprzewidywalnością działania, brakiem bezpośredniego sterowania lub samodzielnym doбором środków. W niniejszej dysertacji autonomia istotna moralnie ma jednak znaczenie normatywne i oznacza zdolność rozpoznawania racji oraz modyfikowania reguł działania w odpowiedzi na krytykę. Program badawczy powinien więc wyraźnie odróżniać autonomię wykonawczą, związaną z bezpieczeństwem i kontrolą działania systemu, od autonomii normatywnej, odnoszącej się do możliwości pociągania do odpowiedzialności. Następnie należy analizować, kiedy i w jakim sensie autonomię normatywną można przypisywać systemom w sposób nieprowadzący do rozproszenia ani rozmycia odpowiedzialności. Rozróżnienie to ma istotne znaczenie przy kształtowaniu rozwiązań instytucjonalnych, ponieważ może dojść do błędu polegającego na utożsamieniu działania systemu autonomicznego z jego odpowiedzialnością normatywną. Mini-konkluzja: badania nad autonomią powinny precyzyjnie określać warunki przypisywania racji w układach hybrydowych, zamiast opierać się na potocznych rozumieniach „samodzielności”. Nie należy jednak utożsamiać tej autonomii z H1: H1 wyznacza minimalną adresowalność proceduralną, natomiast autonomia normatywna pozostaje przedmiotem dalszej analizy silniejszych przypisań odpowiedzialności moralnej.

Druga rekomendacja badawcza dotyczy moralnej wrażliwości oraz jej granic. W 7.1 wykazano, że samo przedstawianie uzasadnień związanych z H1 nie wystarcza do silnej rozliczalności właściwej H3, ponieważ w praktyce moralnej kluczowe znaczenie ma zdolność rozpoznawania istotnych normatywnie sytuacji oraz rozstrzygania konfliktów wartości w sposób niearbitralny. Współczesne analizy dotyczące ograniczeń sztucznych agentów moralnych pokazują, że system może wytwarzać uzasadnienia, a mimo to nie wykazywać wystarczającej wrażliwości moralnej, ponieważ sposób rozpoznawania sytuacji zależy od przyjętych reprezentacji, danych oraz poziomu opisu (Graff, 2024, s. 1–12; Constantinescu, Voinea, Uszkai, & Vică, 2021, s. 803–814). Prowadzi to do konieczności opracowania programu badawczego skoncentrowanego na

dwóch kwestiach. Po pierwsze, czy „wrażliwość moralna” powinna być przypisywana samemu systemowi, czy raczej rozwiązaniom instytucjonalnym, takim jak procedury korekty, audytu i uwzględniania wielu perspektyw; po drugie, w jaki sposób projektować rozwiązania, aby ograniczać te deficyty bez przypisywania systemom statusu podmiotowego. Mini-konkluzja: badania nad wrażliwością moralną powinny zmierzać do określenia warunków wystarczającej deliberacji w praktykach publicznych, a nie do przedwczesnego przypisywania systemom pełnej agencji.

Trzecia rekomendacja badawcza dotyczy normatywnej wyjaśnialności oraz minimalnych wymogów interpretowalności. Skoro dysertacja ujmuje rozliczalność jako praktykę publicznego uzasadniania, wyjaśnienie nie jest elementem dodatkowym, lecz warunkiem uznania decyzji za dopuszczalną. Adams wskazuje, że nieprzejrzystość medycznych zastosowań AI może utrudniać rozliczanie ludzi i instytucji odpowiedzialnych za ich użycie (Adams, 2023, s. 618). Wymóg powiązania wyjaśnień z publicznym zakwestionowaniem i korektą rozstrzygnięć rozwijam tu jako element własnego modelu. Nowsze ujęcia proponują więc określanie minimalnych poziomów interpretowalności, jakich należy wymagać od sztucznych agentów moralnych, co pozostaje zgodne z tezą dysertacji, że MLI wyznacza próg praktyk rozliczalności (Vijayaraghavan & Badea, 2025, s. 2071–2087; Zafar, 2025, s. 2605–2622). Rekomendacja badawcza polega na precyzyjnym wskazaniu, które rodzaje wyjaśnień mają znaczenie normatywne, w szczególności wyjaśnień proceduralnych, kontrfaktycznych oraz odnoszących się do ograniczeń i wyjątków, a także na określeniu, jak różnicować wymagania MLI w zależności od poziomu ryzyka oraz roli systemu w procesie decyzyjnym. Mini-konkluzja: badania nad interpretowalnością powinny koncentrować się na jej znaczeniu normatywnym, a nie ograniczać się do przekazywania informacji.

Czwarta rekomendacja badawcza dotyczy podporządkowania się systemom AI oraz przekazywania im osądu moralnego, rozumianego jako rezygnacja z własnej odpowiedzialności na rzecz autorytetu przypisywanego systemowi. Z perspektywy dysertacji kluczowe jest rozróżnienie dwóch mechanizmów. Pierwszy ma charakter poznawczy. Użytkownik myli wsparcie decyzyjne z rozstrzygnięciem i traktuje wskazanie systemu jako wystarczającej podstawy do podjęcia decyzji. Drugi ma charakter projektowy. Interfejs, sposób formułowania komunikatów, ustawienia domyślne oraz prezentowanie stopnia pewności mogą wzmacniać wrażenie autorytetu systemu i utrudniać krytyczną ocenę jego rezultatów. Samo przekazywanie osądu nie jest

z definicji nieracjonalne, ponieważ w złożonych społeczeństwach podobne delegowanie decyzji występuje również w relacjach z instytucjami i ekspertami. Relacje odpowiedzialności wobec ludzi obejmują możliwość formułowania roszczeń, uzyskiwania odpowiedzi oraz stosowania sankcji. W odniesieniu do systemów AI należy natomiast odrębnie ustalić, czy przedstawione uzasadnienie jest związane z faktyczną podstawą rozstrzygnięcia i umożliwia jego ocenę, czy jedynie stwarza taki pozór. Badania powinny skupiać się na analizie tego, w jaki sposób systemy AI budują autorytet. Chodzi o rozwiązania komunikacyjne i projektowe, które sprawiają, że decyzja przestaje być jedynie wspierana przez narzędzie, a zaczyna być faktycznie podejmowana przez system. Równocześnie należy identyfikować procedury, takie jak MHC, MLI oraz możliwość zakwestionowania decyzji, które mogą temu przeciwdziałać (Whittlestone i in., 2019, s. 195–200; Morley i in., 2020, s. 2155–2156). Mini-konkluzja: bez uwzględnienia perspektywy użytkownika ryzyko cedowania osądu moralnego na systemy AI pozostaje niedoszacowane, co sprzyja rozmyciu odpowiedzialności mimo istnienia formalnych progów kontroli.

Piąta rekomendacja badawcza dotyczy niepewności moralnej w osi pacjencji. Dysertacja nie rozstrzyga, czy systemy są zdolne do odczuwania, ale jednocześnie nie utożsamia ostrożności z ich personifikacją, ponieważ ostrożność może mieć wyłącznie charakter ochronny. Z punktu widzenia teorii niepewności moralnej uzasadnione jest wprowadzanie rozwiązań ochronnych w sytuacjach, w których potencjalna krzywda miałaby istotne znaczenie normatywne, a niepewności nie da się usunąć w rozsądnym czasie (MacAskill, Bykvist, & Ord, 2020, s. 197–210). Oznacza to konieczność określenia warunków uruchamiania poziomu P2, a więc wskazania, jakie przesłanki są istotne dla hipotezy pacjencji, jakie progi dowodowe należy przyjąć oraz jak unikać wnioskowania o zdolności doznawania jedynie na podstawie zewnętrznych zachowań. Mini-konkluzja: badania nad pacjencją AI powinny skupiać się na kryteriach ochronnych wynikających z niepewności, a nie na języku prawa wobec podmiotów.

Szósta rekomendacja badawcza dotyczy osobowości prawnej oraz ryzyka tzw. „syntetycznych osób”. Koncepcja osobowości prawnej zaproponowana przez Kurkiego pokazuje, że w prawie osobowość można ujmować jako zestaw uprawnień i obowiązków przypisywanych w określonym celu regulacyjnym (Kurki, 2019, s. 91–113, zwł. s. 93). Ujęcie to bywa wykorzystywane jako argument na rzecz dopuszczalności prawnej personifikacji AI. Dysertacja wskazuje jednak, że w odniesieniu do systemów sztucznej inteligencji takie rozwiązanie wiąże się z ryzykiem normatywnym, ponieważ sprzyja

rozproszeniu odpowiedzialności, co podkreślają krytyczne analizy pojęcia „syntetycznych osób” w prawie (Bryson, Diamantis, & Grant, 2017, s. 273–291). Rekomendacja badawcza polega zatem na tym, aby personifikację traktować nie jako rozwiązanie domyślne ani wynikające z samego rozwoju technologii, lecz jako instrument wymagający każdorazowej oceny według kryterium odpowiedzialności, to znaczy pod kątem tego, czy ułatwia wskazanie podmiotów odpowiedzialnych oraz dochodzenie kompensacji szkód, czy też te procesy utrudnia. Mini-konkluzja: w badaniach prawno-filozoficznych osobowość prawna AI powinna być ujmowana jako rozwiązanie wyjątkowe, wymagające szczególnego uzasadnienia.

Siódma rekomendacja badawcza dotyczy sposobu korzystania z systemów AI przez użytkowników oraz poznawczych warunków skuteczności minimalnej interpretowalności. Jeżeli wyjaśnienia mają charakter nadmiernie techniczny lub nie dają się powiązać z praktyką podejmowania decyzji, MLI może spełniać wymóg formalny, nie zapewniając jednak rzeczywistej kontroli. W przyjętej tu analizie brak zrozumiałych i użytecznych wyjaśnień może utrzymywać asymetrię informacyjną oraz utrudniać rozpoznanie rozwiązań ograniczających samodzielność użytkownika. Punktem odniesienia jest rozróżnienie między technicznym wyjaśnieniem a rzeczywistym wsparciem ludzkiego osądu, nie empiryczny dowód koniecznego związku nieprzejrzystości z manipulacją (por. Whittlestone i in., 2019, s. 195–200; Morley i in., 2020, s. 2155–2156). Badania powinny koncentrować się na: (1) określeniu minimalnych wymagań dotyczących zrozumiałości uzasadnień dla różnych grup użytkowników i kontekstów instytucjonalnych, (2) analizie obowiązku zapewniania kompetencji cyfrowych oraz odpowiedzialności instytucji za szkolenia i komunikację, a także (3) identyfikacji takich rozwiązań w projektowaniu interfejsów, które utrudniają kwestionowanie decyzji, nawet wtedy, gdy formalnie przewidziano możliwość odwołania. Mini-konkluzja: bez uwzględnienia perspektywy użytkownika MLI może sprowadzać się do pozornej zgodności, zamiast pełnić funkcję rzeczywistego narzędzia rozliczalności.

Z przedstawionych rekomendacji badawczych wynikają rekomendacje regulacyjne o charakterze normatywno-koncepcyjnym. Pierwsza z nich dotyczy sposobu rozumienia obowiązku nadzoru człowieka. Należy go rozumieć jako rzeczywistą kontrolę człowieka nad działaniem systemu, a nie jako samą formalną obecność „człowieka w pętli”. W literaturze znacząca kontrola człowieka (MHC) jest traktowana jako warunek zachowania odpowiedzialności w systemach autonomicznych, ponieważ obejmuje

zarówno możliwość ingerencji w działanie systemu, jak i możliwość wskazania osób oraz instytucji odpowiedzialnych za podjęte decyzje (Mecacci & Santoni de Sio, 2020, s. 104, 111). Jeżeli zatem regulacje przewidują obowiązek nadzoru, powinien on oznaczać utrzymanie realnych warunków odpowiedzialności, a nie jedynie spełnienie wymogu formalnego. W przeciwnym razie nadzór nie wpływa na rozliczalność decyzji, a jedynie stwarza wrażenie jej istnienia. Mini-konkluzja: nadzór należy oceniać ze względu na to, czy umożliwia korektę decyzji i przypisanie odpowiedzialności, a nie ze względu na sam fakt formalnej obecności człowieka.

Drua rekomendacja regulacyjna dotyczy możliwości zakwestionowania decyzji oraz warunków, które muszą zostać spełnione, aby taka kwestia była możliwa. AI Act wprowadza podejście oparte na ryzyku i ustanawia dla systemów wysokiego ryzyka wymagania związane z zarządzaniem ryzykiem, dokumentacją, przejrzystością i nadzorem (AI Act, art. 9–14). Z perspektywy dysertacji obowiązki te mają sens tylko wtedy, gdy umożliwiają rzeczywiste podważenie decyzji przez osoby, których dotyczą oraz przez organy kontrolne. Wymaga to istnienia uzasadnień i materiału dowodowego pozwalających na ocenę i weryfikację decyzji. Istotne znaczenie ma tu także instrument Rady Europy, który osadza regulację AI w ramach praw człowieka, demokracji i rządów prawa oraz wzmacnia wymogi przejrzystości, nadzoru i środków ochrony (CETS 225, art. 8–9; art. 14; art. 16). Rekomendacja dysertacji polega na powiązaniu możliwości zakwestionowania decyzji z minimalną interpretowalnością (MLI). Przekazywane informacje powinny być wystarczające nie ze względu na ich ilość, lecz dlatego, że pozwalają ocenić decyzję i ją ponownie rozpatrzyć. Mini-konkluzja: przejrzystość regulacyjna ma znaczenie jako warunek uzasadniania decyzji i ich kontroli, a nie jedynie jako obowiązek informacyjny.

Trzecia rekomendacja regulacyjna dotyczy spójności wymagań stosowanych przed wdrożeniem systemu oraz po jego wdrożeniu. Jeżeli na etapie *ex ante* prawo wymaga prowadzenia dokumentacji, zapewnienia nadzoru oraz zarządzania ryzykiem, to po wystąpieniu szkody te same elementy powinny umożliwiać ustalenie odpowiedzialności oraz dochodzenie odszkodowania. Znowelizowana dyrektywa o odpowiedzialności za produkty wadliwe obejmuje również oprogramowanie, w tym systemy AI oraz wzmacnia możliwości dochodzenia roszczeń. Jest to zgodne z tezą dysertacji, że odpowiedzialność powinna spoczywać na podmiotach kontrolujących cykl życia produktu, a nie na samym artefakcie (Parlament Europejski i Rada Unii Europejskiej, 2024b, art. 4, 7, 9; mot. 12, 13, 47). Oznacza to, że obowiązki nakładane na

etapie *ex ante* powinny być projektowane tak, aby tworzyły użyteczną podstawę dowodową na etapie *ex post*, a egzekwowanie prawa powinno koncentrować się na możliwości rozliczenia, a nie na spełnieniu wymogów formalnych. Mini-konkluzja: spójność wymagań przedwdrożeniowych i powdrożeniowych jest konieczna, aby prawo ograniczało lukę odpowiedzialności, a nie ją maskowało.

Czwarta rekomendacja regulacyjna dotyczy ograniczania ryzyka antropomorfizacji oraz ryzyka podporządkowania się systemom, traktowanych jako problemy normatywne, zwłaszcza w obszarach o wysokim ryzyku. Z perspektywy dysertacji wynika to bezpośrednio z osi pacjencji P1. Obowiązki związane z projektowaniem i wdrażaniem systemów mają chronić relacje społeczne i instytucjonalne, a nie opierać się na przypisywaniu systemom zdolności do doznawania. Bryson argumentuje przeciw przyjmowaniu sztucznej pacjencji jako celu projektowego i wskazuje ryzyko niewłaściwego przenoszenia odpowiedzialności na artefakty (Bryson, 2018, s. 21–23; Bryson, 2010, s. 63–74). Rekomendacja praktyczna ma zatem charakter zapobiegawczy. Należy przyjąć, że podporządkowanie się systemowi nie wynika wyłącznie z błędu poznawczego po stronie użytkownika, lecz bywa wzmacniane przez sposób zaprojektowania interfejsu oraz formę komunikacji, na przykład poprzez sugerowanie nieomyślności albo przypisywanie systemowi „kompetencji moralnej”. Ograniczanie antropomorfizacji powinno obejmować również wymagania projektowe oraz obowiązki prowadzenia komunikacji zrozumiałej dla użytkowników, tak aby MLI oraz możliwość kwestionowania rozstrzygnięć działały realnie, a nie jedynie na poziomie formalnym. Należy ograniczać sposoby komunikacji oraz rozwiązania, które powodują nieuprawnione przypisywanie systemom autorytetu moralnego albo zastępowanie nimi ludzkiego osądu. W systemach konwersacyjnych stosowanych w opiece zdrowotnej lub edukacji sposób prowadzenia interakcji może wzmacniać zależność użytkownika od systemu AI, a tym samym osłabiać jego samodzielność decyzyjną, niezależnie od ontologicznego statusu systemu. Mini-konkluzja: zapobieganie antropomorfizacji jest elementem ochrony praktyk przypisywania odpowiedzialności, a nie uzasadnieniem nadawania systemom AI praw.

Piąta rekomendacja regulacyjna dotyczy obszarów szczególnie wrażliwych normatywnie, w których odpowiedzialność ma znaczenie, a nawet niewielki błąd może prowadzić do poważnych konsekwencji. Przykład systemów autonomicznych stosowanych przez wojsko pokazuje, że autonomia pozbawiona znaczącej kontroli człowieka rodzi przede wszystkim problem normatywny, a nie techniczny. Kluczowe

pozostaje pytanie, czy decyzje o najwyższej wadze mogą podlegać mechanizmom rozliczalności (U.S. Department of Defense, 2023, s. 2–4, 7–9). Nawet bez odwoływania się do badań empirycznych przykład ten wskazuje, że wymagania dotyczące znaczącej kontroli człowieka (MHC) oraz rozliczalności nie mają charakteru uzupełniającego, lecz stanowią warunek utrzymania odpowiedzialności w praktykach publicznych. Mini-konkluzja: im większe znaczenie normatywne decyzji, tym mniejsza jest dopuszczalność rozwiązań, które uniemożliwiają jej korektę.

Szósta rekomendacja regulacyjna dotyczy międzynarodowego wymiaru regulacji oraz ryzyka arbitrażu regulacyjnego (ang. *regulatory arbitrage*). W rozdziale 6 wskazano, że w sytuacjach, w których oddziaływanie systemów AI wykracza poza granice państw, instytucje odpowiedzialne za egzekwowanie norm funkcjonują zasadniczo w ramach porządków krajowych. W takich warunkach nawet starannie zaprojektowane progi rozliczalności (MHC, MLI, spójność *ex ante/ex post*) mogą tracić na skuteczności, gdy etapy projektowania, trenowania lub eksploatacji są przenoszone do jurysdykcji o mniej rygorystycznych standardach. Problemem nie jest zatem brak norm, lecz ich praktyczna nieskuteczność: odpowiedzialność staje się trudna do wyegzekwowania, a ryzyko zostaje przeniesione na podmioty znajdujące się na końcu łańcucha działania, takie jak użytkownicy, instytucje publiczne oraz osoby dotknięte decyzjami systemu. Z perspektywy dysertacji prowadzi to do wniosku, że konieczne jest ustanawianie minimalnych, wspólnych standardów dowodowych i proceduralnych, obowiązujących ponad granicami jurysdykcji, odnoszących się do (i) kontroli człowieka, (ii) interpretowalności decyzji oraz (iii) jednoznacznej identyfikowalności ról odpowiedzialnych w transgranicznych łańcuchach działania. Instrument Rady Europy, osadzający regulację AI w ramach praw człowieka, demokracji i rządów prawa, może w tym kontekście stanowić jedno z narzędzi budowania wspólnej podstawy normatywnej (CETS 225, art. 8–9; art. 14; art. 16). Mini-konkluzja: bez koordynacji między jurysdykcjami progi rozliczalności pozostają podatne na wykorzystywanie różnic regulacyjnych, co prowadzi do trwałego osłabienia odpowiedzialności.

Siódma rekomendacja regulacyjna dotyczy bodźców ekonomicznych oraz sposobu rozkładu kosztów zgodności. Dotychczasowe rekomendacje zakładają, że wymogi MHC i MLI będą realizowane w sposób realny. Jednak presja interesów sprzyja tworzeniu procedur, które jedynie formalnie potwierdzają zgodność, nie zwiększając faktycznej rozliczalności. Zjawisko to w dysertacji określono jako ryzyko „etyki deklaratywnej” (Whittlestone, Nyrop, Alexandrova, & Cave, 2019, s. 195–200). Dlatego

reżim odpowiedzialności powinien być projektowany tak, aby wytwarzał bodźce do utrzymywania warunków korekty i audytu, a nie wyłącznie do spełniania wymogów formalnych. Kluczowe znaczenie ma tu spójność *ex ante/ex post*: jeżeli prawo nakłada obowiązki dokumentacyjne i nadzorcze na etapie wdrożenia, muszą one przekładać się na realną możliwość dochodzenia roszczeń oraz naprawy szkód. Zgodnie z przyjętą w dysertacji zasadą odpowiedzialność powinna spoczywać na podmiotach kontrolujących cykl życia produktu, a nie na samym artefakcie (Parlament Europejski i Rada Unii Europejskiej, 2024b, art. 4, art. 7, art. 9; mot. 12, 13, 47). Jednocześnie standardy te powinny być stopniowalne i proporcjonalne do poziomu ryzyka, tak aby koszty zgodności nie prowadziły do niezamierzonej koncentracji rynku ani do przerzucania ryzyka na najsłabsze elementy układu socjotechnicznego. Mini-konkluzja: rozliczalność wymaga takiego rozkładu odpowiedzialności i kosztów, który ogranicza pozorną zgodność i zachowuje rzeczywiste możliwości korekty.

Ósma rekomendacja regulacyjna dotyczy zdolności ładu normatywnego do dostosowywania się do zmian, określanej jako zarządzanie adaptacyjne (ang. *adaptive governance*). Jak wykazano w rozdziale 6, mechanizmy instytucjonalne często reagują z opóźnieniem, a luka odpowiedzialności może narastać stopniowo wraz ze zmianami modeli, danych oraz kontekstu użycia. Stąd progi MHC i MLI nie powinny być traktowane jako jednorazowy warunek dopuszczenia systemu, lecz jako wymagania podlegające stałemu utrzymaniu. Jest to zgodne z warunkowym charakterem statusu S2: skoro quasi-podmiot proceduralny funkcjonuje jedynie tak długo, jak długo działają mechanizmy kontroli i weryfikacji, regulacje powinny jasno określać moment wygaśnięcia statusu S2 oraz przewidywać procedury ponownej oceny w razie istotnych zmian, takich jak aktualizacje, zmiany danych, dryf modeli czy zmiana kontekstu użycia. Adaptacyjność nie oznacza trwałej deregulacji, lecz wprowadzanie mechanizmów aktualizacji standardów, które chronią możliwość publicznego kwestionowania decyzji mimo zmienności technologicznej. Mini-konkluzja: w warunkach szybkich zmian technologicznych rozliczalność wymaga adaptacyjnego ładu regulacyjnego, aby standardy MHC i MLI oraz reżimy P1 i P2 nie traciły związku z praktyką.

Na zakończenie można sformułować syntetyczny wniosek dysertacji jako zasadę porządkowania ładu normatywnego. Podmiotowość moralna AI, o ile ma znaczenie w porządku publicznym, ma przede wszystkim charakter proceduralno-instytucjonalny. Rozstrzygające nie jest nadanie systemowi określonego statusu moralnego, lecz ustanowienie progów rozliczalności i interpretowalności, które umożliwiają praktyki

relacji z drugą osobą, w szczególności formułowanie roszczeń, przedstawianie uzasadnień, wprowadzanie korekt oraz dochodzenie kompensacji szkód.

W konsekwencji odpowiedzialności nie należy utożsamiać z nadawaniem systemom AI statusu osoby, lecz z hierarchią statusów normatywnych, która porządkuje zasady ich przypisywania. W osi sprawczości służą temu znacząca kontrola człowieka (MHC), minimalna interpretowalność (MLI) oraz spójność wymagań stosowanych przed i po wdrożeniu systemu (*ex ante* i *ex post*). W osi pacjencji rolę tę pełnią reżimy P1 oraz rozwiązania ochronne P2. W tym sensie rekomendacje badawczo-regulacyjne stanowią logiczne uzupełnienie weryfikacji hipotez H1–H3. W wymiarze normatywnym H1 odpowiada warunkowi adresowalności, H2 wskazuje możliwy mechanizm stabilizacji realizacji racji, a H3 warunek jawnego rozstrzygnięcia konfliktów; K4 odrębnie ocenia stabilność diachroniczną, natomiast WI określa instytucjonalne podstawy dowodowe i możliwość kwestionowania decyzji. Całość prowadzi do wniosku, że odpowiedzialność w układach socjotechnicznych wymaga istnienia instytucjonalnych podstaw dowodowych oraz możliwości kwestionowania decyzji, zamiast przypisywania systemom AI cech podmiotowych (Moor, 2006, s. 18–21; Floridi & Sanders, 2004, s. 349–379; Bryson, 2018, s. 21–22; Bryson, 2010, s. 63–74). Na obecnym etapie kształtowania norm odpowiedzialności oraz rozwoju technologii sztucznej inteligencji odpowiedzialność powinna spoczywać w rękach ludzi i instytucji. Jej wykonywanie i egzekwowanie należy wiązać z jasno określonymi procedurami, uzgadnianymi ponad granicami państw i systematycznie aktualizowanymi, a nie z przypisywaniem odpowiedzialności artefaktom technicznemu. W tym sensie PH1–PH3 nie tworzą osobnego modelu, lecz normatywnie uzasadniają H1–H3, których zastosowanie instytucjonalne określa WI. Całość domyka WI jako warunek instytucjonalnego zastosowania modelu H1–H3; WI nie jest H4 ani dodatkowym poziomem podmiotowości.

Słownik pojęć i oznaczeń

Słownik porządkuje znaczenia przyjęte w rozprawie oraz wskazuje miejsca ich rozwinięcia. Objasnienia terminologiczne mają charakter syntetyczny i pomocniczy: nie stanowią wyczerpujących definicji, lecz punkt wyjścia dla argumentacji rozwiniętej w poszczególnych rozdziałach. W odniesieniu do AI opisują konceptualne warunki możliwych przypisań, nie zaś potwierdzone zdolności współczesnych systemów. Odpowiedniki angielskie pełnią funkcję orientacyjną i przekładową; przy terminach autorskich nie oznaczają przejęcia gotowej definicji z literatury.

Słownik pojęć

Adresowalność proceduralna (ang. *procedural addressability*). Możliwość traktowania systemu jako adresata roszczeń o uzasadnienie, ocenę i korektę, oparta na zdolnościach funkcjonalnych określonych przez H1. Wymaga rozpoznawania norm, przewidywania konsekwencji ich naruszeń, kontrfaktycznego reagowania na racje oraz udostępniania wystarczających i wiarygodnych podstaw decyzji, umożliwiających ocenę ich związku z wynikiem. Sama możliwość zewnętrznego audytu narzędzia nie jest jeszcze jego adresowalnością proceduralną. *Zob. podrozdz. 1.2–1.3; 3.4; 5.4.2.*

Afektywność funkcjonalna (ang. *functional affectivity*). Mechanizm wartościowania, który nadaje sytuacjom i bodźcom walencję, czyli wartość istotną dla wyboru, uwagi lub hamowania reakcji oraz wiąże ją z normami i celami. W H2 pełni funkcję stabilizującą wybory zgodne z racjami. Nie zakłada doznań fenomenalnych i nie stanowi warunku koniecznego H1. *Zob. podrozdz. 1.2; 3.5.*

Audytowalność (ang. *auditability*). Możliwość odtworzenia i poddania ocenie podstaw decyzji oraz korekt na przyjętym poziomie opisu, z wykorzystaniem dostępnych uzasadnień, dokumentacji i zapisów. Umożliwia publiczną kontrolę, lecz nie dowodzi, że sam system rozpoznaje racje lub posiada metakontrolę. *Zob. podrozdz. 3.2; 5.1; 5.4.2–5.4.5; 6.4.3.*

Empirycznie zakotwiczony przypadek graniczny (ang. *empirically grounded boundary case*). Istniejący system opisany na podstawie opublikowanych danych i dokumentacji, służący analitycznemu sprawdzeniu rozróżnień modelu. Materiał źródłowy uzasadnia twierdzenia opisowe o systemie; zastosowanie H1–H3 + WI pozostaje analizą pojęciową, a nie własnym badaniem empirycznym ani empiryczną walidacją modelu. *Zob. podrozdz. 1.5; 1.7; 5.4.1.*

Kontrast negatywny (ang. *negative contrast*). Przypadek ujawniający ograniczenia systemu względem rozważanych warunków modelu. W ocenie należy odróżnić niespełnienie warunku od niewykazania jego spełnienia na podstawie dostępnej dokumentacji. Funkcja kontrastowa nie uprawnia do przypisywania systemowi wyższego statusu. *Zob. podrozdz. 1.6–1.7; 5.4.2–5.4.5.*

Kontrola metapoznawcza (metakontrola) (ang. *metacognitive control; metacontrol*). Kontrola drugiego rzędu nad własnym procesem decyzyjnym: monitorowanie go, ujmowanie konfliktów racji jako przedmiotu kontroli, wskazywanie reguły pierwszeństwa i przedstawianie audytowalnego uzasadnienia korekty. Wyznacza wymaganie H3. Nie jest tożsama ze sprawdzaniem jakości danych, wstrzymaniem wyniku ani przekazaniem przypadku człowiekowi. *Zob. podrozdz. 1.2; 3.6; 5.4.3.*

Luka odpowiedzialności (ang. *responsibility gap*). Brak lub ograniczenie możliwości uzasadnionego przypisania i egzekwowania odpowiedzialności za moralnie istotne skutki działania. Luka może wystąpić także wtedy, gdy role twórcy, operatora i użytkownika systemu AI oraz ich udział przyczynowy są znane, lecz nie dysponowali oni wiedzą lub kontrolą wystarczającą do uzasadnienia przypisania im odpowiedzialności w danym zakresie. W rozprawie pojęcie obejmuje również deficyty identyfikowalności adresatów, audytowalności racji oraz trwałości praktyk rozliczania. Może dotyczyć pojedynczego przypadku albo mieć charakter strukturalny; porządkują je LSP, ZSP i jakościowe wymiary M1–M5. *Zob. podrozdz. 6.4.1–6.4.5.*

Minimalna interpretowalność (MLI) (ang. *minimal interpretability*). Poziom wyjaśnialności i dostępu do informacji wystarczający na przyjętym poziomie opisu do odtworzenia podstaw decyzji, przeprowadzenia audytu oraz przedstawienia uzasadnienia zrozumiałego w procedurach publicznych. Obejmuje dostęp do odpowiedniej dokumentacji. Doprecyzowuje parametr I w ramach WI; nie zastępuje zdolności wymaganych przez H1 ani H3. *Zob. podrozdz. 5.1–5.2; 5.4.4.*

Pacjencja moralna (ang. *moral patiency*). Status pacjenta moralnego, czyli bycie podmiotem, wobec którego istnieją bezpośrednie obowiązki ochronne ze względu na możliwość doznania moralnie istotnej krzywdy. Należy ją odróżniać od sprawczości i odpowiedzialności. W osi P0–P2 ochrona relacyjna P1 nie zakłada krzywdy samego systemu, a reżim ostrożnościowy P2 nie stanowi uznania jego faktycznej pacjencji. *Zob. podrozdz. 7.2.*

Podstawa decyzji (ang. *decision basis*). Rekonstruowalny układ przesłanek, reguł i zależności prowadzących do rozstrzygnięcia. Jego udostępnienie powinno umożliwiać ocenę związku uzasadnienia z wynikiem, audyt, zakwestionowanie i korektę decyzji. Podstawa decyzji nie jest dowolną narracją wytworzoną po uzyskaniu wyniku. *Zob. podrozdz. 1.2–1.3; 3.2; 5.4.2.*

Poziom opisu (poziom abstrakcji, LoA) (ang. *Level of Abstraction*). Rama analizy wyznaczająca własności i relacje uznane za istotne dla określonego pytania oraz sposób opisywania działania systemu. Wybór poziomu opisu nie jest samodzielną tezą o naturze systemu ani przypisaniem mu statusu moralnego. Pozwala odróżniać opis zachowania, mechanizmu i normatywnego uzasadnienia. *Zob. podrozdz. 3.1–3.2; 5.1–5.2.*

Przesłanka normatywno-decyzyjna (ang. *normative-decisional premise*). Jawny element podstawy decyzji systemu, istotny normatywnie dla rozstrzygnięcia. Określenie to służy ostrożnemu stosowaniu języka racji wobec AI, bez przypisywania systemowi ludzkiego rozumienia norm lub przeżyć. *Zob. podrozdz. 1.3; 3.2; 3.6.*

Przypisywalność (ang. *attributability*). Możliwość przypisania podmiotowi czynu lub stanu jako wyrazu jego charakteru albo postaw. Sama nie oznacza jeszcze obowiązku publicznego uzasadniania działania ani podlegania roszczeniom i sankcjom w relacji z drugą osobą. *Zob. podrozdz. 2.2; 7.1.*

Quasi-podmiot proceduralny (ang. *procedural quasi-subject*). Warunkowa i ograniczona rola systemu w instytucjonalnych procedurach uzasadniania i korekty. Wymaga odrębnego wykazania H1 i F3 oraz zapewnienia odpowiednich parametrów R i I w ramach WI. Nie oznacza przyznania pełnej osobowości moralnej ani prawnej i nie zastępuje ludzi ani instytucji jako końcowych adresatów odpowiedzialności. *Zob. podrozdz. 5.1–5.2; 5.4.4–5.4.5; 7.2.*

Racja (ang. *reason*). W podstawowym sensie normatywnym: wzgląd przemawiający za przyjęciem przekonania, dokonaniem oceny lub podjęciem działania, np. istotny fakt albo stan rzeczy. W argumentacji racja może być wyrażona jako przesłanka; stanowi podstawę uzasadnienia, a nie uzasadnienie rozumiane jako cała procedura przedstawiania racji lub jej rezultat. Należy odróżniać rację normatywną, która przemawia za danym rozstrzygnięciem, od racji motywującej, którą sprawca faktycznie kieruje się w działaniu. W rekonstrukcjach filozoficznych osobno rozpatruje się przyczynową rolę racji. W odniesieniu do AI termin stosowany jest funkcjonalnie i ostrożnie, w powiązaniu z przesłankami normatywno-decyzyjnymi, bez założenia ludzkiego sposobu rozumienia racji. *Zob. podrozdz. 1.3; 2.1–2.3; 3.2.*

Reguła pierwszeństwa (ang. *priority rule*). Reguła wskazująca, która przesłanka lub racja ma pierwszeństwo w sytuacji kolizji. Jawność tej reguły umożliwia ocenę sposobu rozstrzygnięcia konfliktu. Jej stosowanie w decyzji pierwszego rzędu należy odróżniać od metapoznawczej kontroli i uzasadnienia wymaganych przez H3. *Zob. podrozdz. 1.3; 3.2; 3.6; 4.1.4.*

Responsywność na racje (ang. *reasons-responsiveness*). Zdolność reagowania na rozpoznane racje oraz odpowiedniego modyfikowania decyzji i działania w ich świetle, także w rozważanych sytuacjach kontrfaktycznych. Racje muszą wpływać na rozstrzygnięcia i korekty, a nie jedynie towarzyszyć im jako wyjaśnienia po fakcie. K2 służy diagnozie tej zależności. *Zob. podrozdz. 2.2–2.3; 3.2.*

Rozliczalność (ang. *accountability*). Relacja społeczna i normatywna, w której podmiot indywidualny, zbiorowy lub instytucjonalny jest adresatem uprawnionych roszczeń o przedstawienie uzasadnienia oraz podlega krytyce i sankcjom zgodnie z regułami właściwych praktyk odpowiedzialności. Jest czymś więcej niż samą przypisywalnością czynu lub udzielaniem racji. Nie należy utożsamiać ogólnej kategorii rozliczalności z warunkiem H3 ani z samą odpowiedzialnością odszkodowawczą. *Zob. podrozdz. 2.2; 5.1; 7.1.*

Silna rozliczalność (ang. *strong accountability*). Wyższy próg określony przez H3, wymagający przy zachowaniu H1 kontroli metapoznawczej, jawnego rozstrzygania konfliktów racji i audytowalnego uzasadnienia drugiego rzędu. Instytucjonalne utrzymanie i egzekwowanie tak rozumianej rozliczalności podlega WI. Próg ten nie oznacza pełnej ludzkiej podmiotowości moralnej. *Zob. podrozdz. 1.2–1.3; 2.2; 3.6; 7.1.*

Słaba odpowiedzialność proceduralna (ang. *weak procedural responsibility*). Ograniczony tryb odpowiedzialności oparty na minimalnej adresowalności H1, odnoszący się do uzasadniania, oceny i korekty decyzji. Nie obejmuje pełnej odpowiedzialności moralnej człowieka i nie przesądza o samodzielnej rozliczalności instytucjonalnej systemu. *Zob. podrozdz. 1.2–1.4; 3.4; 7.1.*

Stabilność diachroniczna (ang. *diachronic stability*). Trwałość i wiarygodność profilu działania systemu w czasie, obejmująca ciągłość podstaw decyzji i możliwość rekonstrukcji ich korekt. Jest oceniana przez K4. Nie oznacza niezmienności decyzji mimo nowych racji; nie jest też tym samym co mechanizm H2 ani trwałość praktyk instytucjonalnych oceniana przez M4. *Zob. podrozdz. 3.5–3.6; 6.4.3.*

System oceny odstępstw (ang. *system for assessing procedural deviations*). Hipotetyczny przypadek analityczny w obszarze pomocy społecznej, służący ilustracji pełniejszej konfiguracji H1–H3 + WI. Nie opisuje rzeczywistego urzędu, konkretnej procedury prawnej ani projektu technicznego. Ostateczna odpowiedzialność w rozważanym układzie pozostaje po stronie ludzi i instytucji. *Zob. podrozdz. 1.6.*

Udzielanie racji (ang. *answerability*). Uzasadnianie własnych działań wobec innych oraz poddawanie tych uzasadnień publicznej ocenie i korekcie w świetle kontrargumentów. Jest odrębne od samego przypisania czynu podmiotowi oraz od pełnej relacji rozliczalności, która obejmuje dodatkowo uprawnione roszczenia i sankcje. *Zob. podrozdz. 2.2; 2.5; 7.1.*

Warunek instytucjonalnego zastosowania modelu (WI) (ang. *institutional condition*). Istnienie publicznie rozpoznawalnych ról, procedur audytu, nadzoru, zakwestionowania i korekty oraz zasad odpowiedzialności umożliwiających zastosowanie H1–H3. WI dotyczy organizacji praktyk społeczno-technicznych, nie zaś dodatkowej zdolności systemu. Nie jest H4 i nie zastępuje wykazania warunków agentowych. *Zob. podrozdz. 1.2–1.4; 5.4.4; 7.1.*

Wyjaśnialność (ang. *explainability*). Możliwość przedstawienia zrozumiałego dla właściwego odbiorcy i dostatecznie wiernego wyjaśnienia podstaw decyzji systemu AI, w tym odpowiedzi na pytanie „dlaczego ten wynik, a nie inny?”. W rozprawie ma znaczenie epistemiczne i proceduralne: umożliwia ocenę związku ujawnionych przesłanek z wynikiem oraz wspiera zakwestionowanie i korektę decyzji. Nie jest tożsama z normatywnym uzasadnieniem słuszności decyzji, pełną przejrzystością mechanizmu ani samą zdolnością generowania przekonujących komunikatów. Dostęp do wystarczających i wiarygodnych podstaw decyzji jest jednym z wymogów H1, a minimalna interpretowalność (MLI) określa poziom informacji potrzebny do instytucjonalnego audytu. Sama wyjaśnialność nie dowodzi spełnienia H1 ani H3. *Zob. podrozdz. 2.5; 3.2; 4.1.4; 5.1–5.2.*

Znacząca kontrola człowieka (MHC) (ang. *meaningful human control*). Warunek obejmujący reagowanie na racje właściwych podmiotów ludzkich (tracking) oraz możliwość śledzenia odpowiedzialności (tracing), czyli powiązania skutków działania z rozpoznawalnymi osobami i instytucjami. Wymaga realnej możliwości korekty, nie tylko formalnej obecności człowieka. Doprecyzowuje parametr R w ramach WI i nie jest metakontrolą H3 samego systemu. *Zob. podrozdz. 2.5; 4.3; 5.1–5.2; 5.4.4.*

Klucz oznaczeń i nazw konstrukcji

Klucz rozdziela hipotezy, kryteria diagnostyczne, funkcje normatywne, klasyfikacje oraz warunki instytucjonalne. Oznaczenia nie są elementami jednej skali i nie mogą być używane zamiennie. Skróty LoA, MHC i MLI objaśniono w części hasłowej.

H1–H3 oraz WI. H1: minimalna adresowalność proceduralna i podstawa słabej odpowiedzialności proceduralnej. H2: afektywność funkcjonalna jako możliwy mechanizm stabilizacji, niekonieczny dla H1. H3: próg silnej rozliczalności wymagający metakontroli. WI: warunek instytucjonalnego zastosowania modelu. Oznaczenia te nie tworzą czterech kolejnych stopni jednej skali. *Zob. 1.2–1.4; 7.1.*

K1–K4. K1: autokorekta norm i audytowalna korekta decyzji. K2: responsywność na racje. K3: refleksyjność predykcyjna i przyczynowe powiązanie deliberacji z działaniem. K4: stabilność diachroniczna i wiarygodność profilu działania. K1–K3 wspierają ocenę H1; K4 odrębnie ocenia trwałość rezultatów i nie dowodzi spełnienia H2. *Zob. 3.2–3.6.*

PH1–PH3. PH1: deontyczne granice dopuszczalności i publicznego uzasadniania. PH2: stabilizacja i wiarygodność praktyk instytucjonalnych. PH3: jawne porządkowanie konfliktów racji wśród opcji dopuszczalnych. Jest to warstwa normatywnego uzasadnienia, a nie drugi zestaw hipotez ani odpowiedniki H1, H2 i H3 w relacji jeden do jednego. *Zob. 4.1–4.2.*

LexOrder. Autorska zasada pierwszeństwa funkcji normatywnych: PH1 przed PH2, a następnie PH3 w obrębie opcji dopuszczalnych i wiarygodnie realizowanych. Porządkuje uzasadnienia, nie stanowi algorytmu i nie pozwala usprawiedliwiać naruszenia granic dopuszczalności korzystnym bilansem skutków. *Zob. 4.1.4; 4.3–4.4.*

FairAim. Autorski wymóg bezstronnego i publicznie uzasadnialnego ustalania zasad, celów, wyjątków oraz priorytetów. Dotyczy zarówno udziału adresatów w procedurze legitymizacji, jak i możliwości uzasadnienia jej rozstrzygnięć bez uprzywilejowania jednej strony. *Zob. 4.1.4; 4.2–4.4.*

ZSW — Zasada Stopniowej Weryfikowalności. Normatywno-proceduralna ocena dopuszczalności i legitymizacji przed działaniem, wiarygodności w jego toku oraz jawności priorytetów i rozliczalności po działaniu. Zachowuje porządek PH1–PH3; nie jest empirycznym testem systemu. *Zob. 4.4.*

MPU — Minimum Publicznego Uzasadnienia. Wymóg publicznie dostępnego uzasadnienia, zwłaszcza odstępstw od reguł oraz ich jawnego rejestrowania. Uzupełnia LexOrder, FairAim i ZSW, umożliwiając kontrolę przyjętych racji zamiast poprzestawania na deklarowanej zgodności. *Zob. 4.4; podsumowanie rozdz. 4.*

$A \times F + (R, I)$. Rama łącząca porządki uzasadnienia A, poziomy funkcjonalne F oraz przekrojowe parametry R i I. Znak $\times$ oznacza zestawienie wymiarów klasyfikacji; zapis nie jest formułą obliczania statusu. R i I są parametrami instytucjonalnymi, nie zaś kolejnymi poziomami osi F. *Zob. 5.2.*

A1–A4 — porządki uzasadnienia. A1: esencjalizm własnościowy. A2: performatywizm funkcjonalny. A3: relacjonizm społeczny. A4: konstruktywizm instytucjonalny. Porządki nie są rozłączne ani hierarchiczne; A2/A4 oznacza współwystępowanie perspektywy funkcjonalnej i instytucjonalnej. *Zob. 5.2.*

F1–F4 — poziomy funkcjonalne. F1: wpływ etyczny. F2: etyka wbudowana, czyli ograniczenia i mechanizmy zgodności. F3: jawne reprezentowanie norm i uzasadnienia rzeczywiście powiązane z procedurą decyzyjną, poddające się audytowi. F4: pełna agencja etyczna jako idea regulatywna. Samo F3 nie przesądza o H1, H3 ani statusie quasi-podmiotowym. *Zob. 5.2; 5.4.4–5.4.5.*

R oraz I. R: instytucjonalny parametr rozliczalności, doprecyzowany przez MHC, obejmujący reagowanie na racje i śledzenie odpowiedzialności. I: parametr minimalnej interpretowalności, doprecyzowany przez MLI. Oba należą do warunków utrzymania rozliczalności w ramach WI. *Zob. 5.1–5.2; 5.4.4.*

LSP — lokalna zasada przypisania odpowiedzialności. Dla moralnie istotnej szkody albo poważnego ryzyka jej wystąpienia powinien istnieć przynajmniej jeden indywidualny lub zbiorowy adresat spełniający warunki uzasadnionego przypisania: identyfikowalność, możliwość odtworzenia racji oraz kontrolę zależną od racji. Jest to wymóg normatywny, nie twierdzenie, że w każdym przypadku taki adresat faktycznie istnieje. *Zob. 6.4.2.*

ZSP — zasada superwencji praktyk przypisywania odpowiedzialności. W podrozdziale 2.4 oznacza zależność praktyk odpowiedzialności od identyfikowalności adresata, audytowalności racji i kontroli zależnej od racji. W podrozdziale 6.4.2 uzyskuje rozwinięcie instytucjonalne: warunki LSP mają być utrzymywane jako trwały standard, również wobec przypadków granicznych. *Zob. 2.4; 6.4.2.*

M1–M5 — jakościowe wymiary diagnostyczne. M1: identyfikowalność podmiotów odpowiedzialnych. M2: audytowalność racji. M3: kontrola zależna od racji. M4: diachroniczna trwałość praktyk przypisywania odpowiedzialności. M5: relacja tempa działania systemu do czasu i wiedzy potrzebnych na zrozumienie, interwencję i korektę. Wymiary służą jakościowej diagnozie, nie empirycznemu pomiarowi; M4 nie jest K4. *Zob. 6.4.3–6.4.5.*

S0–S3 — oś sprawczości i odpowiedzialności. Oznaczenia porządkujące role od artefaktu istotnego moralnie, przez instrument normatywnie nasycony, po quasi-podmiot proceduralny S2. S3 pozostaje graniczną ideą pełnej agencji. Szczegółowe przyporządkowania, w tym występujące w wywodzie zapisy S0/S1 oraz S1/S2, należy odczytywać w kontekście podrozdziału 7.2; klucz nie zastępuje ich rozstrzygnięciem o rozłącznych klasach. *Zob. 7.2.*

P0–P2 — oś pacjencji i ochrony. P0: domyślny brak bezpośredniej pacjencji w porządku publicznym. P1: ochrona pochodna lub relacyjna, uzasadniona wpływem sposobu traktowania systemów na ludzi i normy społeczne. P2: reżim proporcjonalnej ostrożności przy poważnych przesłankach możliwej krzywdy. P2 nie dowodzi faktycznej pacjencji ani nie przyznaje automatycznie praw. *Zob. 7.2.*

EPM — miara mocy wyjaśniającej. Skala 0–3 porządkująca argumenty z eksperymentów myślowych według koherencji definicyjnej, odporności na kontrprzykłady i potencjału predykcyjnego rozumianego normatywno-proceduralnie. Ocenia moc wyjaśniającą konfiguracji modelu, nie mierzy empirycznie zdolności systemów AI. *Zob. 1.5.*

WT oraz ZO — oznaczenia pomocnicze rozdziału 2. WT: warunek tła, czyli świadomość dostępu lub jej funkcjonalny ekwiwalent zapewniający dostępność treści decyzyjnych do audytu i korekty. ZO: zasada ontologicznej obojętności minimalnych warunków odpowiedzialności; dotyczy ważności przypisania przy zachowanych warunkach funkcjonalnych, nie równoważności wszystkich ontologii ani zasad sankcji i kompensacji. *Zob. 2.4.*

Bibliografia

Literatura

- Abràmoff, M. D., Lavin, P. T., Birch, M., Shah, N., & Folk, J. C. (2018). Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. *npj Digital Medicine*, 1, Article 39. <https://doi.org/10.1038/s41746-018-0040-6>
- Adams, J. (2023). Defending explicability as a principle for the ethics of artificial intelligence in medicine. *Medicine, Health Care and Philosophy*, 26(4), 615–623. <https://doi.org/10.1007/s11019-023-10175-7>
- Allen, C., Smit, I., & Wallach, W. (2005). Artificial morality: top-down, bottom-up, and hybrid approaches. *Ethics and Information Technology*, 7, 149–155. <https://doi.org/10.1007/s10676-006-0004-4>
- Ambartsoumean, V. M., & Yampolskiy, R. V. (2023). AI risk skepticism: A comprehensive survey. arXiv. <https://doi.org/10.48550/arXiv.2303.03885>
- Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). *Concrete problems in AI safety*. arXiv:1606.06565. <https://doi.org/10.48550/arXiv.1606.06565>
- Anderson, M., & Anderson, S. L. (2007). Machine ethics: creating an ethical intelligent agent. *AI Magazine*, 28(4), 15–26.
- Anderson, M., & Anderson, S. L. (red.). (2011). *Machine ethics*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511978036>
- Anscombe, G. E. M. (2000). *Intention*. Cambridge University Press.
- Arkin, R. C. (2009). *Governing lethal behavior in autonomous robots*. Chapman & Hall/CRC. <https://doi.org/10.1201/9781420085952>
- Arystoteles. (2003). *O duszy* (P. Siwek, tłum.). Wydawnictwo Naukowe PWN.
- Arystoteles. (2012). *Etyka nikomachejska* (D. Gromska, tłum.). Wydawnictwo Naukowe PWN.
- Austin, J. L. (1975). *How to do things with words* (wyd. 2; J. O. Urmson & M. Sbisà, red.). Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780198245537.001.0001>

- Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., Joseph, N., Kadavath, S., Kernion, J., Conerly, T., El-Showk, S., Elhage, N., Hatfield-Dodds, Z., Hernandez, D., Hume, T., ... Kaplan, J. (2022). *Training a helpful and harmless assistant with reinforcement learning from human feedback*. arXiv. <https://doi.org/10.48550/arXiv.2204.05862>
- Barrett, L. F. (2017). *How emotions are made: the secret life of the brain*. Houghton Mifflin Harcourt.
- Barrett, L. F., & Bliss-Moreau, E. (2009). Affect as a psychological primitive. *Advances in Experimental Social Psychology*, 41, 167–218. [https://doi.org/10.1016/S0065-2601\(08\)00404-8](https://doi.org/10.1016/S0065-2601(08)00404-8)
- Baum, K., Hermanns, H., & Speith, T. (2018). From machine ethics to machine explainability and back. W: *International Symposium on Artificial Intelligence and Mathematics (ISAIM 2018)*, Fort Lauderdale, FL, United States, January 3–5 (s. 1–9). POWVER Technical Report 2018–02.
- Baxter, W. F. (1974). *People or penguins: the case for optimal pollution*. Columbia University Press.
- Berger, P. L., & Luckmann, T. (1966). *The social construction of reality: a treatise in the sociology of knowledge*. Anchor Books.
- Bikkasani, D. C. (2025). Navigating artificial general intelligence (AGI): *Societal implications, ethical considerations, and governance strategies*. *AI and Ethics*, 5(3), 2021–2036. <https://doi.org/10.1007/s43681-024-00642-z>
- Birhane, A. (2021). Algorithmic injustice: a relational ethics approach. *Patterns*, 2(2), art. 100205. <https://doi.org/10.1016/j.patter.2021.100205>
- Block, N. (1995). On a confusion about a function of consciousness. *Behavioral and Brain Sciences*, 18(2), 227–247.
- Bostrom, N. (2014). *Superintelligence: paths, dangers, strategies*. Oxford University Press.
- Bratman, M. E. (1987). *Intention, plans, and practical reason*. Harvard University Press.
- Bryson, J. J. (2010). Robots should be slaves. W: Y. Wilks (red.), *Close engagements with artificial companions* (s. 63–74). John Benjamins.

- Bryson, J. J. (2018). Patience is not a virtue: the design of intelligent systems and systems of ethics. *Ethics and Information Technology*, 20(1), 15–26. <https://doi.org/10.1007/s10676-018-9448-6>
- Bryson, J. J., Diamantis, M. E., & Grant, T. D. (2017). Of, for, and by the people: the legal lacuna of synthetic persons. *Artificial Intelligence and Law*, 25(3), 273–291. <https://doi.org/10.1007/s10506-017-9214-9>
- Burrell, J. (2016). How the machine «thinks»: understanding opacity in machine learning algorithms. *Big Data & Society*, 3(1), 1–12.
- Carruthers, P. (1992). *The animals issue: moral theory in practice*. Cambridge University Press.
- Chalmers, D. J. (2010). The singularity: a philosophical analysis. *Journal of Consciousness Studies*, 17(9–10), 7–65.
- Chappell, R. Y., Meissner, D., & MacAskill, W. (2024). *An introduction to utilitarianism: from theory to practice*. Hackett Publishing Company.
- Chen, M. (2023). A trilemma for the singularitarian. *Philosophy & Technology*, 36(2), 1–22.
- Coeckelbergh, M. (2020). *AI ethics*. MIT Press. <https://doi.org/10.7551/mitpress/12549.001.0001>
- Coeckelbergh, M. (2023). Narrative responsibility and artificial intelligence: How AI challenges human responsibility and sense-making. *AI & Society*, 38, 2437–2450. <https://doi.org/10.1007/s00146-021-01375-x>
- Coeckelbergh, M. (2025a). The case for global governance of AI: arguments, counter-arguments, and challenges ahead. *AI & Society*, 40, 1803–1806. <https://doi.org/10.1007/s00146-024-01949-5>
- Coeckelbergh, M. (2025b). Three challenges for a global AI ethics: towards a more relational normative vision. *AI and Ethics*, 5, 5527–5533. <https://doi.org/10.1007/s43681-025-00791-9>
- Cohen, C. (1986). The case for the use of animals in biomedical research. *The New England Journal of Medicine*, 315(14), 865–870.

- Constantinescu, M., Voinea, C., Uszkai, R., & Vică, C. (2021). Understanding responsibility gaps in AI-assisted decision-making. *Philosophy & Technology*, 34(3), 803–814.
- Damasio, A. R. (1994). *Descartes' error: emotion, reason, and the human brain*. G. P. Putnam.
- Damasio, A. R. (1996). The somatic marker hypothesis and the possible functions of the prefrontal cortex. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 351(1346), 1413–1420. <https://doi.org/10.1098/rstb.1996.0125>
- Danaher, J. (2022). Automated decision-making and the burden of moral responsibility. *Philosophy & Technology*, 35(2), 1–25. <https://doi.org/10.1007/s13347-022-00514-7>
- Darwall, S. (2006). *The second-person standpoint: morality, respect, and accountability*. Harvard University Press.
- Da Silva, M. (2024a). Responsibility gaps. *Philosophy Compass*, 19(9–10), e70002. <https://doi.org/10.1111/phc3.70002>
- Da Silva, M. (2024b). When to fill responsibility gaps: a proposal. *The Journal of Value Inquiry*. <https://doi.org/10.1007/s10790-024-10011-y>
- Davidovic, J. (2023). On the purpose of meaningful human control of AI. *Frontiers in Big Data*, 5, art. 1017677. <https://doi.org/10.3389/fdata.2022.1017677>
- Davidson, D. (1963). Actions, reasons, and causes. *The Journal of Philosophy*, 60(23), 685–700.
- Demirtas, H. (2025). AI responsibility gap: Not new, inevitable, unproblematic. *Ethics and Information Technology*, 27, Article 7. <https://doi.org/10.1007/s10676-024-09814-1>
- Dennett, D. C. (1984). *Elbow room: the varieties of free will worth wanting*. MIT Press.
- Dennett, D. C. (1987). *The intentional stance*. MIT Press.
- Dennett, D. C. (2016). *Świadomość* (E. Stokłosa, tłum.). Copernicus Center Press.
- Dessureault, J.-S., Lamontagne, R., & Parisé, P.-O. (2025). The ethics of creating artificial superintelligence: a global risk perspective. *AI and Ethics*, 5, 6241–6263. <https://doi.org/10.1007/s43681-025-00793-7>

- Diakopoulos, N. (2015). Algorithmic accountability: journalistic investigation of computational power structures. *Digital Journalism*, 3(3), 398–415. <https://doi.org/10.1080/21670811.2014.976411>
- Dignum, V. (2019). *Responsible artificial intelligence: how to develop and use AI in a responsible way*. Springer.
- Erman, E., & Furendal, M. (2024). Artificial intelligence and the political legitimacy of global governance. *Political Studies*, 72(2), 421–441. <https://doi.org/10.1177/00323217221126665>
- Evans, J. C. (2012). Chapter 5: critical analysis of Taylor’s philosophy of respect for nature. W: *With respect for nature: living as part of the natural world* (s. 97–126). SUNY Press.
- Faroldi, F. L. G. (2025). Risk and artificial general intelligence. *AI & Society*, 40(4), 2541–2549. <https://doi.org/10.1007/s00146-024-02004-z>
- Fearnley, L., & Habli, I. (2025). Concept creep in safe artificial intelligence. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 8(1), 901–912. <https://doi.org/10.1609/aies.v8i1.36599>
- Fischer, J. M., & Ravizza, M. (1998). *Responsibility and control: a theory of moral responsibility*. Cambridge University Press.
- Floridi, L., & Cowls, J. (2019). A unified framework of five principles for AI in society. *Harvard Data Science Review*, 1(1). <https://doi.org/10.1162/99608f92.8cd550d1>
- Floridi, L., & Cowls, J. (2021). A unified framework of five principles for AI in society. W: L. Floridi (red.), *Ethics, governance, and policies in artificial intelligence* (s. 5–17). Springer. https://doi.org/10.1007/978-3-030-81907-1_2
- Floridi, L., & Sanders, J. W. (2004). On the morality of artificial agents. *Minds and Machines*, 14(3), 349–379. <https://doi.org/10.1023/B:MIND.0000035461.63578.9d>
- Frankfurt, H. G. (1969). Alternate possibilities and moral responsibility. *The Journal of Philosophy*, 66(23), 829–839.
- French, P. A. (1979). The corporation as a moral person. *American Philosophical Quarterly*, 16(3), 207–215.
- Frey, R. G. (1980). *Interests and rights: the case against animals*. Clarendon Press.

- Gabriel, I. (2020). Artificial intelligence, values, and alignment. *Minds and Machines*, 30(3), 411–437. <https://doi.org/10.1007/s11023-020-09539-2>
- Gilbert, M. (2013). *Joint commitment: how we make the social world*. Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780199970148.001.0001>
- Gloor, L., & Sotala, K. (2018). Artificially sentient beings: moral, political, and legal issues. W: M. Dubber, F. Pasquale, & S. Das (red.), *The Oxford handbook of ethics of AI* (s. 41–48). Oxford University Press.
- Good, I. J. (1965). Speculations concerning the first ultraintelligent machine. W: F. L. Alt & M. Rubinoff (red.), *Advances in computers* (t. 6, s. 31–88). Academic Press.
- Graff, J. (2024). Moral sensitivity and the limits of artificial moral agents. *Ethics and Information Technology*, 26, 1–12. <https://doi.org/10.1007/s10676-024-09755-9>
- Gunkel, D. J. (2012). *The machine question: critical perspectives on AI, robots, and ethics*. MIT Press.
- Gunkel, D. J. (2018). *Robot rights*. MIT Press.
- Hadfield-Menell, D., Dragan, A., Abbeel, P., & Russell, S. (2017). The off-switch game. W: *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17)* (s. 475–482).
- Hagendorff, T. (2020). The ethics of AI ethics: an evaluation of guidelines. *Minds and Machines*, 30(1), 99–120. <https://doi.org/10.1007/s11023-020-09517-8>
- Hagendorff, T. (2022). AI ethics and power. *Ethics and Information Technology*, 24(3), art. 55. <https://doi.org/10.1007/s10676-022-09602-9>
- Hanson, R. (2016). *The age of em: work, love, and life when robots rule the Earth*. Oxford University Press.
- Hendrycks, D., Mazeika, M., & Woodside, T. (2023). *An overview of catastrophic AI risks* (arXiv:2306.12001, version 6). arXiv. <https://arxiv.org/abs/2306.12001v6>
- Hernández-Orallo, J. (2017). *The measure of all minds*. Cambridge University Press.
- Hew, P. C. (2014). Artificial moral agents are infeasible with foreseeable technologies. *Ethics and Information Technology*, 16(3), 197–206.
- Himma, K. E. (2009). Artificial agency, consciousness, and the criteria for moral agency: what properties must an artificial agent have to be a moral agent? *Ethics and Information Technology*, 11(1), 19–29.

- Honneth, A. (1992). *Kampf um Anerkennung: Zur moralischen Grammatik sozialer Konflikte*. Suhrkamp.
- Hume, D. (1978). *A treatise of human nature* (L. A. Selby-Bigge & P. H. Nidditch, red.). Oxford University Press.
- Hume, D. (2023). *Traktat o naturze ludzkiej* (Cz. Znamierowski, tłum.). Fundacja Aletheia.
- Jiang, L., Hwang, J. D., Bhagavatula, C., Le Bras, R., Liang, J. T., Levine, S., Dodge, J., Sakaguchi, K., Forbes, M., Hessel, J., Borchardt, J., Sorensen, T., Gabriel, S., Tsvetkov, Y., Etzioni, O., Sap, M., Rini, R., & Choi, Y. (2025). Investigating machine moral judgement through the Delphi experiment. *Nature Machine Intelligence*, 7(1), 145–160. <https://doi.org/10.1038/s42256-024-00969-6>
- Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1(9), 389–399.
- Johnson, D. G. (2006). Computer systems: moral entities but not moral agents. *Ethics and Information Technology*, 8(4), 195–204.
- Kane, R. (1996). *The significance of free will*. Oxford University Press.
- Kant, I. (2001). *Krytyka czystego rozumu* (P. Chmielowski, tłum.; H. Struve, red.). Antyk.
- Kant, I. (2002). *Uzasadnienie metafizyki moralności* (M. Wartenberg, tłum.). De Agostini Polska; Ediciones Altaya Polska.
- Kant, I. (2004). *Krytyka praktycznego rozumu* (J. Gałęcki, tłum., przedm. i przypisy). Wydawnictwo Naukowe PWN.
- Kasirzadeh, A. (2025). Two types of AI existential risk: decisive and accumulative. *Philosophical Studies*, 182, 1975–2003. <https://doi.org/10.1007/s11098-025-02301-3>
- Königs, P. (2022). Artificial intelligence and responsibility gaps: what is the problem? *Ethics and Information Technology*, 24, art. 36. <https://doi.org/10.1007/s10676-022-09643-0>
- Krakovna, V., Uesato, J., Mikulik, V., Rahtz, M., Everitt, T., Kumar, R., Kenton, Z., Leike, J., & Legg, S. (2020, April 21). *Specification gaming: the flip side of AI ingenuity*. Google DeepMind. <https://deepmind.google/blog/specification-gaming-the-flip-side-of-ai-ingenuity/>

- Kurki, V. A. J. (2019). *A theory of legal personhood*. Oxford University Press.
- Kutz, C. (2000). *Complicity: ethics and law for a collective age*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511663758>
- Landgrebe, J., & Smith, B. (2022). Why machine learning will never give us AGI. *AI & Society*, 37(2), 1–18.
- Levy, N. (2014). *Consciousness and moral responsibility*. Oxford University Press.
- List, C., & Pettit, P. (2011). *Group agency: the possibility, design, and status of corporate agents*. Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780199591565.001.0001>
- MacAskill, W., Bykvist, K., & Ord, T. (2020). *Moral uncertainty*. Oxford University Press.
- Manheim, D., & Garrabrant, S. (2019). *Categorizing variants of Goodhart's law*. arXiv:1803.04585. <https://doi.org/10.48550/arXiv.1803.04585>
- Matthias, A. (2004). The responsibility gap: ascribing responsibility for the actions of learning automata. *Ethics and Information Technology*, 6(3), 175–183. <https://doi.org/10.1007/s10676-004-3422-1>
- McDowell, J. (1996). *Mind and world*. Harvard University Press.
- McKenna, M. (2008). A hard-line reply to Pereboom's four-case manipulation argument. *Philosophy and Phenomenological Research*, 77(1), 142–159. <https://doi.org/10.1111/j.1933-1592.2008.00179.x>
- Mecacci, G., & Santoni de Sio, F. (2020). Meaningful human control as reason-responsiveness: the case of dual-mode vehicles. *Ethics and Information Technology*, 22(2), 103–115. <https://doi.org/10.1007/s10676-019-09519-w>
- Mele, A. R. (2019). Free will and moral responsibility: manipulation, luck, and agents' histories. *Midwest Studies in Philosophy*, 43(1), 75–92. <https://doi.org/10.1111/misp.12105>
- Metzinger, T. (2019). *Tunel ego. Naukowe badanie umysłu a mit świadomego „ja”* (P. Grabarczyk, tłum.). Wydawnictwo Uniwersytetu Łódzkiego.
- Misselhorn, C. (2022). Artificial moral agents: conceptual issues and ethical controversy. W: S. Voeneky, P. Kellmeyer, O. Müller & W. Burgard (red.), *The Cambridge*

- handbook of responsible artificial intelligence* (s. 31–49). Cambridge University Press.
- Mökander, J., Morley, J., Taddeo, M., & Floridi, L. (2021). Ethics-based auditing of automated decision-making systems: nature, scope, and limitations. *Science and Engineering Ethics*, 27, art. 44. <https://doi.org/10.1007/s11948-021-00319-4>
- Moor, J. H. (2006). The nature, importance, and difficulty of machine ethics. *IEEE Intelligent Systems*, 21(4), 18–21.
- Moor, J. H. (2009). Four kinds of ethical robots. *IEEE Intelligent Systems*, 24(4), 14–20.
- Morley, J., Floridi, L., Kinsey, L., & Elhalal, A. (2020). From what to how: an initial review of publicly available AI ethics tools, methods and research to translate principles into practices. *Science and Engineering Ethics*, 26(4), 2141–2168. <https://doi.org/10.1007/s11948-019-00165-5>
- Morley, J., Kinsey, L., Elhalal, A., Garcia, F., Ziosi, M., & Floridi, L. (2023). Operationalising AI ethics: barriers, enablers and next steps. *AI & Society*, 38(1), 411–423. <https://doi.org/10.1007/s00146-021-01308-8>
- Morris, M. R., Sohl-Dickstein, J., Fiedel, N., Warkentin, T., Dafoe, A., Faust, A., Farabet, C., & Legg, S. (2023). *Levels of AGI for operationalizing progress on the path to AGI*. arXiv. <https://doi.org/10.48550/arXiv.2311.02462>
- Mühlhoff, R. (2025). The ethics of AI: Power, critique, responsibility. Bristol University Press. <https://doi.org/10.51952/9781529249262>
- Munch, L. A., Mainz, J. T., & Bjerring, J. C. (2023). The value of responsibility gaps in algorithmic decision-making. *Ethics and Information Technology*, 25(1), art. 21. <https://doi.org/10.1007/s10676-023-09699-6>
- Nagel, T. (1979). Moral luck. W: *Mortal questions* (s. 24–38). Cambridge University Press.
- Naudé, W., & Dimitri, N. (2020). The race for an artificial general intelligence. *AI & Society*, 35(2), 367–377.
- Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: verbal reports on mental processes. *Psychological Review*, 84(3), 231–259.
- Nyholm, S. (2023). *This is AI ethics*. Oxford University Press.

- Olson, E. T. (1997). *The human animal: personal identity without psychology*. Oxford University Press.
- Ord, T. (2020). *The precipice: existential risk and the future of humanity*. Bloomsbury.
- Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Aspell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). *Training language models to follow instructions with human feedback*. arXiv. <https://doi.org/10.48550/arXiv.2203.02155>
- Palumbo, G., Carneiro, D., & Alves, V. (2025). Objective metrics for ethical AI: a systematic literature review. *International Journal of Data Science and Analytics*, 20(2), 247–267. <https://doi.org/10.1007/s41060-024-00541-w>
- Parfit, D. (1984). *Reasons and persons*. Clarendon Press.
- Passmore, J. (1980). *Man's responsibility for nature: ecological problems and Western traditions* (wyd. 2). Duckworth. (oryg. wyd. 1974).
- Pearl, J. (2009). *Causality: models, reasoning, and inference* (wyd. 2). Cambridge University Press. <https://doi.org/10.1017/CBO9780511803161>
- Pereboom, D. (2001). *Living without free will*. Cambridge University Press.
- Picard, R. W. (1995). *Affective computing* (MIT Media Laboratory Technical Report No. 321). Massachusetts Institute of Technology.
- Popper, K. R. (1959). *The logic of scientific discovery*. Hutchinson.
- Raman, R., Kowalski, R., Achuthan, K., Iyer, A., & Nedungadi, P. (2025). Navigating artificial general intelligence development: societal, technological, ethical, and brain-inspired pathways. *Scientific Reports*, 15, art. 8443. <https://doi.org/10.1038/s41598-025-92190-7>
- Rawls, J. (1971). *A theory of justice*. Harvard University Press.
- Robbins, S. (2024). The many meanings of meaningful human control. *AI and Ethics*, 4, 1377–1388. <https://doi.org/10.1007/s43681-023-00320-6>
- Ross, W. D. (2002). *The right and the good*. Oxford University Press.
- Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206–215. <https://doi.org/10.1038/s42256-019-0048-x>

- Runge, R. F. (2025). The ontological and moral status of whole brain emulations in neo-Aristotelian naturalism. *AI and Ethics*, 5(4), 4211–4222. <https://doi.org/10.1007/s43681-025-00711-x>
- Russell, S., & Norvig, P. (2021). *Artificial intelligence: a modern approach* (wyd. 4). Pearson.
- Sandberg, A. (2014). Ethics of brain emulations. *Journal of Experimental & Theoretical Artificial Intelligence*, 26(3), 439–457. <https://doi.org/10.1080/0952813X.2014.895114>
- Sandberg, A., & Bostrom, N. (2008). *Whole brain emulation: a roadmap*. Future of Humanity Institute, University of Oxford.
- Santoni de Sio, F., & Mecacci, G. (2021). Four responsibility gaps with artificial intelligence: why they matter and how to address them. *Philosophy & Technology*, 34, 1057–1084. <https://doi.org/10.1007/s13347-021-00450-x>
- Santoni de Sio, F., Mecacci, G., Calvert, S., Heikoop, D., Hagenzieker, M., & van Arem, B. (2023). Realising meaningful human control over automated driving systems: a multidisciplinary approach. *Minds and Machines*, 33, 587–611. <https://doi.org/10.1007/s11023-022-09608-8>
- Santoni de Sio, F., & van den Hoven, J. (2018). Meaningful human control over autonomous systems. *Ethics and Information Technology*, 20(1), 1–14.
- Scanlon, T. M. (1998). *What we owe to each other*. Harvard University Press.
- Scanlon, T. M. (2008). *Moral dimensions: permissibility, meaning, blame*. Belknap Press of Harvard University Press.
- Schechtman, M. (1996). *The constitution of selves*. Cornell University Press.
- Shevlin, H. (2021). How could we know when a robot was a moral patient? *Cambridge Quarterly of Healthcare Ethics*, 30(3), 459–471. <https://doi.org/10.1017/S0963180120001012>
- Siebert, L. C., Lupetti, M. L., Aizenberg, E., Beckers, N., Zgonnikov, A., Veluwenkamp, H., Abbink, D., Giaccardi, E., Houben, G.-J., Jonker, C. M., van den Hoven, J., Forster, D., & Lagendijk, R. L. (2022). Meaningful human control: actionable properties for AI system development. *AI and Ethics*, 3, 241–255. <https://doi.org/10.1007/s43681-022-00167-3>

- Smith, A. M. (2005). Responsibility for attitudes: activity and passivity in mental life. *Ethics*, 115(2), 236–271.
- Speith, T. (2024). Unlocking the potential of machine ethics with explainability: extended abstract. W: *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024)* (s. 2477–2479). International Foundation for Autonomous Agents and Multiagent Systems.
- Strawson, G. (1994). The impossibility of moral responsibility. *Philosophical Studies*, 75(1–2), 5–24. <https://doi.org/10.1007/BF00989879>
- Strawson, P. F. (2003). *Individuals: an essay in descriptive metaphysics*. Routledge. (oryg. wyd. 1959).
- Strawson, P. F. (2008). Freedom and resentment. W: *Freedom and resentment and other essays* (s. 1–28). Routledge.
- Strawson, P. F. (2019). Wolność a uraza (J. Jarocki, tłum.). *Przegląd Filozoficzny – Nowa Seria*, 28(4), 35–57. <https://doi.org/10.24425/pfns.2019.130910>
- Taddeo, M., & Floridi, L. (2021). How AI can be a force for good. W: L. Floridi (red.), *Ethics, governance, and policies in artificial intelligence* (s. 91–96). Springer. https://doi.org/10.1007/978-3-030-81907-1_7
- Tariq, S., Iftikhar, A., Chaudhary, P., & Khurshid, K. (2023). Is the «technological singularity scenario» possible: can AI parallel and surpass all human mental capabilities? *World Futures*, 79(2), 200–266. <https://doi.org/10.1080/02604027.2022.2050879>
- Thorstad, D. (2025). Against the singularity hypothesis. *Ethics*, 135(4), 1627–1651.
- Torrance, S. (2008). Ethics and consciousness in artificial agents. *AI & Society*, 22(4), 495–521.
- Tuomela, R. (2007). *The philosophy of sociality: the shared point of view*. Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780195313390.001.0001>
- Turkle, S. (2011). *Alone together: why we expect more from technology and less from each other*. Basic Books.
- Vallor, S., & Vierkant, T. (2024). Find the gap: AI, responsible agency and vulnerability. *Minds and Machines*, 34(3), art. 20. <https://doi.org/10.1007/s11023-024-09674-0>
- van Inwagen, P. (1983). *An essay on free will*. Oxford University Press.

- Vijayaraghavan, A., & Badea, C. (2025). Minimum levels of interpretability for artificial moral agents. *AI and Ethics*, 5, 2071–2087. <https://doi.org/10.1007/s43681-024-00536-0>
- Waelen, A. (2022). Procedural fairness and pluralism in AI ethics. *AI and Ethics*, 2(1), art. 9.
- Wallace, R. J. (1998). *Responsibility and the moral sentiments*. Harvard University Press.
- Wallace, R. J. (2019). *The moral nexus*. Princeton University Press.
- Wallach, W., & Allen, C. (2009). *Moral machines: teaching robots right from wrong*. Oxford University Press.
- Wang, H., & Blok, V. (2025). Why putting artificial intelligence ethics into practice is not enough: towards a multi-level framework. *Big Data & Society*, 12(2), 1–14.
- Watson, G. (1996). Two faces of responsibility. *Philosophical Topics*, 24(2), 227–248.
- Whittlestone, J., Nystrup, R., Alexandrova, A., & Cave, S. (2019). The role and limits of principles in AI ethics: towards a focus on tensions. W: *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (s. 195–200). ACM.
- Williams, B. (1981). Moral luck. W: *Moral luck: philosophical papers 1973–1980* (s. 20–39). Cambridge University Press.
- Winfield, A. F. T. (2024). AI and common sense: ambitions and frictions. W: M. Coeckelbergh & E. Di’Domenico (red.), *AI and the future of contemporary societies* (s. 66–81). Routledge.
- Winfield, A. F. T., Blum, C., & Liu, W. (2014). Towards an ethical robot: internal models, consequences and ethical action selection. W: M. Mistry, A. Leonardis, M. Witkowski, & C. Melhuish (red.), *Advances in autonomous robotics systems* (LNCS 8717, s. 85–96). Springer.
- Winfield, A. F. T., van Maris, A., Salvini, P., & Jirotko, M. (2022). An ethical black box for social robots: a draft open standard. W: *ICRES 2022: Seventh International Conference on Robot Ethics and Standards*. <https://doi.org/10.13180/icres.2022.18-19.07.019>
- Woleński, J. (1998). Scepticism and logic. *History of Philosophy & Logical Analysis*, 1(1), 187–194.
- Wolf, S. (1990). *Freedom within reason*. Oxford University Press.

- Woodward, J. (2003). *Making things happen: a theory of causal explanation*. Oxford University Press. <https://doi.org/10.1093/0195155270.001.0001>
- Wróblewski, P., & Fortuna, P. (2023). Ethical AI in human–machine interactions: a psychological perspective. *Annals of Psychology / Roczniki Psychologiczne*, 26(3), 221–240.
- Zafar, M. (2025). Normativity and AI moral agency. *AI and Ethics*, 5, 2605–2622. <https://doi.org/10.1007/s43681-024-00566-8>

Akty prawne i dokumenty instytucjonalne

- Council of Europe. (2024). *Council of Europe framework convention on artificial intelligence and human rights, democracy and the rule of law* (CETS No. 225). Wilno, 5 września 2024. <https://www.coe.int/en/web/conventions/full-list?module=treaty-detail&treatynum=225>
- Digital Diagnostics. (b.d.). *Indications for use – US*. Pobrano 3 września 2026, z <https://www.digitaldiagnostics.com/products/eye-disease/indications-for-use-us/>
- High-Level Expert Group on Artificial Intelligence. (2019). *Ethics guidelines for trustworthy AI*. European Commission.
- International Committee of the Red Cross. (2021, 12 maja). *ICRC position on autonomous weapon systems*. <https://www.icrc.org/en/document/icrc-position-autonomous-weapon-systems>
- International Committee of the Red Cross. (2025, 12 maja). *Preserving human control over the use of force: A call to regulate lethal autonomous weapon systems under international law* [Stanowisko]. <https://www.icrc.org/en/statement/preserving-human-control-over-use-force-call-regulate-lethal-autonomous-weapon-systems>
- Parlament Europejski i Rada Unii Europejskiej. (2024a). *Rozporządzenie Parlamentu Europejskiego i Rady (UE) 2024/1689 z dnia 13 czerwca 2024 r. w sprawie ustanowienia zharmonizowanych przepisów dotyczących sztucznej inteligencji oraz zmiany rozporządzeń (WE) nr 300/2008, (UE) nr 167/2013, (UE) nr 168/2013, (UE) 2018/858, (UE) 2018/1139 i (UE) 2019/2144 oraz dyrektyw 2014/90/UE, (UE) 2016/797 i (UE) 2020/1828 (akt w sprawie sztucznej inteligencji)*. *Dziennik Urzędowy Unii Europejskiej*, L, 2024/1689. <http://data.europa.eu/eli/reg/2024/1689/oj>

- Parlament Europejski i Rada Unii Europejskiej. (2024b). *Dyrektywa Parlamentu Europejskiego i Rady (UE) 2024/2853 z dnia 23 października 2024 r. w sprawie odpowiedzialności za produkty wadliwe oraz uchylająca dyrektywę Rady 85/374/EWG. Dziennik Urzędowy Unii Europejskiej, L, 2024/2853.* <http://data.europa.eu/eli/dir/2024/2853/oj>
- Parlament Europejski i Rada Unii Europejskiej. (2026). *Rozporządzenie Parlamentu Europejskiego i Rady (UE) 2026/1744 z dnia 8 lipca 2026 r. w sprawie zmiany rozporządzeń (UE) 2024/1689, (UE) 2018/1139 i (UE) 2023/1230 w odniesieniu do uproszczenia wdrażania zharmonizowanych przepisów dotyczących sztucznej inteligencji (akt zbiorczy prawa cyfrowego dotyczący AI). Dziennik Urzędowy Unii Europejskiej, L, 2026/1744.* <http://data.europa.eu/eli/reg/2026/1744/oj>
- Tabassi, E. (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)* (NIST AI 100-1). National Institute of Standards and Technology. <https://doi.org/10.6028/NIST.AI.100-1>
- U.S. Department of Defense. (2023). *Department of Defense directive 3000.09: autonomy in weapon systems.* U.S. Department of Defense. <https://www.esd.whs.mil/Portals/54/Documents/DD/issuances/dodd/300009p.pdf>
- U.S. Food and Drug Administration. (2018a). *De Novo classification request for IDx-DR (DEN180001): Decision summary.* Center for Devices and Radiological Health. https://www.accessdata.fda.gov/cdrh_docs/reviews/DEN180001.pdf
- U.S. Food and Drug Administration. (2018b). *Device classification under Section 513(f)(2)(De Novo): IDx-DR (DEN180001).* Pobrano 3 września 2026, z <https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfpmn/denovo.cfm?id=DEN180001>

Kraków, dnia/date.....

.....
imię, nazwisko / name

OŚWIADCZENIE / DECLARATION

Świadomy/a odpowiedzialności oświadczam, że przedłożona rozprawa doktorska pt.:

Aware of my responsibility, I declare that the submitted dissertation entitled:

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

jest mojego autorstwa.

is my own creation.

Jednocześnie oświadczam, że ww. rozprawa nie narusza praw autorskich w rozumieniu Ustawy z dnia 4 lutego 1994 roku o prawie autorskim i prawach pokrewnych (Dz.U. z 2021 r. poz. 1062 z późn. zmianami) oraz dóbr osobistych chronionych prawem cywilnym.

At the same time, I declare that the aforementioned dissertation does not infringe upon copyrights within the meaning of the Act of 4 February 1994 on Copyright and Related Rights (Journal of Laws of 2021, item 1062, as amended) and personal rights protected by civil law.

Przedłożona praca nie zawiera danych empirycznych ani też informacji, które uzyskałem/am w sposób niedozwolony.

The submitted work does not contain empirical data or information that I have obtained by unauthorised means.

Stwierdzam, iż przedstawiona praca w całości ani też w części nie była wcześniej podstawą żadnej innej urzędowej procedury związanej z zakończeniem kształcenia w szkole doktorskiej.

I certify that the submitted work, in whole or in part, has not previously formed the basis of any other official procedure relating to the completion of training in doctoral school.

.....
Data/date

.....
Podpis doktoranta / PhD student's
signature

Kraków, dnia/date.....

.....

imię, nazwisko / name

OŚWIADCZENIE / DECLARATION

o zgodności wersji elektronicznej z wydrukiem

on the conformity of the electronic version with the hard copy

Ja niżej podpisany/podpisana* oświadczam, że wersja elektroniczna jest zgodna z wersją papierową rozprawy doktorskiej.

I, the undersigned, declare that the electronic version is compatible with the paper version of the dissertation.*

.....

Data/date

.....

*Podpis doktoranta / PhD student's
signature*

*Niewłaściwe skreślić/Delete as appropriate

Kraków, dnia/date.....

.....

imię, nazwisko / name

OŚWIADCZENIE / DECLARATION

Ja niżej podpisany/podpisana* wyrażam zgodę/nie wyrażam zgody* na udostępnianie mojej rozprawy doktorskiej w wersji papierowej i elektronicznej dla celów naukowych i dydaktycznych.

I, the undersigned, consent/do not consent to my dissertation being made available in hard copy and electronic format for research and teaching purposes.*

.....

Data/date

.....

*Podpis doktoranta / PhD student's
signature*

*Niewłaściwe skreślić/Delete as appropriate