Fragment

Wdrażanie AI w organizacji. Analiza prawna, ocena ryzyka i metodyka zapewnienia zgodności + wzory do pobrania

Część I. Fundamenty organizacyjne i model zarządzania ryzykiem wdrożenia

Rozdział I. Pełna zgodność AI w 16 krokach

1. Uwagi wstępne

Integracja AI z organizacją nie przypomina już zwykłego zakupu kolejnego systemu IT. Tradycyjne oprogramowanie działało według ustalonych reguł, przewidywalnych ścieżek i zamkniętych instrukcji. Sztuczna inteligencja, zwłaszcza generatywna, agentowa i oparta na rozbudowanych integracjach z bazami wiedzy, działa inaczej. Odpowiada probabilistycznie, uczy się na kontekście, korzysta z danych z wielu warstw organizacji i bardzo szybko przekracza granice, które w prezentacji projektowej wydawały się oczywiste. W rezultacie dotychczasowe modele odpowiedzialności, nadzoru i rozliczalności zaczynają pękać dokładnie tam, gdzie zarząd liczył na przyspieszenie, oszczędność i szybkość decyzyjną.

Najtrudniejszy odcinek leży między językiem prawa a logiką systemu. Z jednej strony, występują normy takie jak przejrzystość, bezpieczeństwo, proporcjonalność, niedyskryminacja i znacząca kontrola człowieka. Z drugiej strony, występują prompty systemowe, uprawnienia, bazy danych wektorowych, RAG, logi, bramki API, semantyczne zapory, kontrola dostępu oraz decyzje o tym, czy model może wykonać akcję autonomicznie. Właśnie tutaj powstaje luka Lex-Machina. To nie jest poetycka metafora. To praktyczny problem wdrożeniowy. Jeżeli twoja organizacja nie zbuduje pomostu prawno-technologicznego Lex-Machina, bardzo szybko okaże się, że system formalnie wygląda na zgodny, ale operacyjnie podejmuje działania, których nikt nie umie dobrze obronić przed klientem, regulatorem, sądem albo własnym zarządem.

Właśnie dlatego ta książka nie została pomyślana jako komentarz do przepisów. Nie ma cię zachwycić teorią. **Ma cię przeprowadzić przez wdrożenie.** Jej celem jest przekształcenie obawy przed sankcją, incydentem, pozwem, utratą danych albo kompromitacją projektu w uporządkowany, powtarzalny i dowodowy proces zarządzania. Ten proces ma dwa cele naraz. Po pierwsze, ma odsiewać projekty, których nie da się obronić. Po drugie, ma przyspieszać te wdrożenia, które rzeczywiście wnoszą wartość biznesową i dają

się utrzymać w ryzach prawa, bezpieczeństwa i odpowiedzialności. Tylko taki układ daje sens pojęciom AI Governance & Compliance.

Punktem ciężkości tej książki jest **metodyka 16 kroków oraz protokół ARIA – AI Risk & Impact Assessment**. To nie jest kolejny formularz do odhaczenia. To mechanizm pracy z dwoma silnikami jednocześnie. Silnik Optymalizacji szuka Zielonych Świateł, czyli miejsc, w których AI daje realny efekt, skraca cykl, usuwa wąskie gardła, odciąża ludzi i poprawia jakość decyzji. Silnik Ryzyka szuka Czerwonych Flag, czyli sytuacji, w których koszt błędu, szkody dla człowieka, naruszenia praw podstawowych, utraty tajemnicy przedsiębiorstwa albo zaburzenia ciągłości działania jest zbyt wysoki. Ta książka zakłada, że oba silniki pracują równolegle. Nie ma zgody na to, żeby wysokie ROI przykrywało projekt, którego nie da się uczciwie obronić.

W tle działa jeszcze jeden fakt, którego nie warto lekceważyć. **Reżim odpowiedzialności wokół AI twardnieje.** Unijny AI Act wprost nakłada obowiązek zapewnienia odpowiedniego poziomu kompetencji AI po stronie personelu i innych osób działających w imieniu dostawców i podmiotów stosujących[1]. Zrewidowana dyrektywa 2024/2853 o odpowiedzialności za produkty wadliwe traktuje oprogramowanie jako produkt, przewiduje transpozycję do prawa krajowego do 9.12.2026 r. i wzmacnia ryzyko roszczeń za szkody spowodowane wadliwym oprogramowaniem[2]. Orzecznictwo TSUE w sprawie SCHUFA pokazuje z kolei, że człowiek przyklejony do procesu jak pieczątka nie rozwiązuje problemu zautomatyzowanego decydowania. Potrzebna jest realna, a nie dekoracyjna kontrola człowieka.

W konsekwencji w tej książce nie będziesz pracować na hasłach, ale na dowodach[3]. Na przypisaniu właścicieli. Na zmapowanych przepływach pracy. Na opisanych zasobach. Na wpisanych do raportu Zielonych Światłach i Czerwonych Flagach. Na udokumentowanym nadzorze człowieka. Na Teczce Obronnej, którą budujesz nie po kryzysie, lecz zanim kryzys się pojawi. To jest różnica między organizacją, która korzysta z AI, a organizacją, która rzeczywiście panuje nad AI.

---

1 Era powierzchownej adaptacji technologii uległa zakończeniu. Prawo europejskie delegalizuje ignorancję operacyjną, wymuszając na organizacjach proaktywne zarządzanie wiedzą. Organizacje nie mogą już polegać wyłącznie na intuicji użytkowników; muszą wdrożyć systemowe środki zapewniające wystarczający poziom kompetencji w dziedzinie AI (tzw. *AI literacy*) swojego personelu, dostosowane do poziomu ryzyka i stopnia ingerencji systemu w dane wejściowe. Zob. AI Act, art. 4.

2 Erozja tradycyjnej linii obrony wdrożeniowców IT staje się faktem. Zrewidowana dyrektywa unijna 2024/2853 wprost i kategorycznie kwalifikuje każde oprogramowanie – w tym złożone, probabilistyczne modele AI – jako „produkt". Wymagająca transpozycji do 9.12.2026 r. regulacja drastycznie obniża próg wejścia w spór sądowy dla poszkodowanych, nakładając domniemanie wadliwości w sytuacjach wysokiej złożoności technicznej (*black-box*), co wymusza na organizacjach budowę żelaznej architektury dowodowej przed wystąpieniem incydentu. Zob. dyrektywa 2024/2853 w sprawie odpowiedzialności za produkty wadliwe.

3 Wdrażanie systemów probabilistycznych w przestrzeni gospodarczej nie może opierać się na zaufaniu, lecz na twardym materiale dowodowym, niezbędnym w procesie sądowym. Zgodnie z fundamentalną zasadą ciężaru dowodu (*onus probandi*) obowiązek wykazania, iż wada nie leżała po stronie architektury systemu, spoczywa na podmiocie, który wywodzi z tego faktu skutki prawne. Co więcej, lekkomyślne wdrażanie systemów przez jednostki organizacyjne niesie ryzyko odpowiedzialności subsydiarnej ich członków za zobowiązania powstałe wskutek działania niekontrolowanych agentów AI. Zob. art. 6, art. 33[1] § 2 KC.

Ważne

Tę książkę czytaj jak instrukcję wdrożeniową. Pracuj wyłącznie na rzeczywistych procesach swojej organizacji. Miej obok otwarty załączony Raport AI Governance & Compliance oraz Procedurę wdrożenia AI Governance & Compliance. Traktuj je jak arkusze audytowe prowadzone równolegle z lekturą. Uzupełniaj je na bieżąco. Nie po rozdziale. Po akapicie, po warsztacie, po ustaleniu właściciela, po wykryciu ryzyka, po decyzji o zatrzymaniu albo dopuszczeniu projektu. Właśnie w ten sposób budujesz twardą architekturę dowodową.

Kolejność ma znaczenie. Najpierw ustalasz, gdzie AI rzeczywiście występuje albo ma zostać uruchomiona. Potem rozpisujesz przepływ pracy krok po kroku. Następnie identyfikujesz zasoby, dane, integracje i właścicieli. Dopiero wtedy uruchamiasz ARIA, czyli ocenę możliwości, zgodności, wagi zagrożeń i prawdopodobieństwa naruszeń. Środki organizacyjne i techniczne projektujesz dopiero po zrozumieniu rzeczywistego procesu. To wydaje się oczywiste, ale właśnie tu wiele organizacji traci pieniądze. Najpierw kupują narzędzie. Potem robią pilotaż. Następnie tworzą politykę. A dopiero na końcu odkrywają, że nikt nie potrafi powiedzieć, kto odpowiada za wynik modelu, gdzie trafiają dane wejściowe, z jakich źródeł korzysta RAG i czy istnieje wyłącznik awaryjny.

Pracuj więc jak konsultant we własnej organizacji. **Zbieraj fakty, nie deklaracje.** Jeżeli dział HR mówi, że „używa AI tylko pomocniczo", sprawdź, co dokładnie trafia do promptu, czy kandydat został właściwie poinformowany, czy wynik modelu wpływa na kolejność rozpatrywania CV i czy człowiek naprawdę może zakwestionować rekomendację. Jeżeli dział sprzedaży twierdzi, że „to tylko asystent", sprawdź, czy chatbot nie wysyła klientowi cen, rabatów albo obietnic bez właściwej autoryzacji[4]. Jeżeli w organizacji działa M365 Copilot lub podobny asystent, sprawdź uprawnienia do folderów, zakres indeksowania plików, politykę retencji i to, czy poufne dane płacowe albo dokumenty z procesów *due diligence* nie stają się – na skutek błędnych uprawnień – treścią łatwo dostępną dla modelu do streszczenia.

Jeżeli widzisz darmowe konta, prywatne prompty, lokalne automatyzacje, rozszerzenia przeglądarki albo narzędzia agentowe uruchamiane poza oficjalnym kanałem IT, wpisz to do rejestru jako Shadow AI. Nie jako ciekawostkę. Nie jako „obejście, bo firma jeszcze nic nie wdrożyła". **Shadow AI to wyraźny sygnał ostrzegawczy, ponieważ zwykle oznacza brak kontroli nad przepływem danych**, brak pewności co do sposobu trenowania modelu, brak reguł retencji, brak ścieżki autoryzacji i brak odpowiedzialnego właściciela[5]. W logice tej książki wykrycie takiego zjawiska powinno zasilać tabelę Czerwo-

[4] Lekceważenie autonomii agentów konwersacyjnych stanowi najprostszą drogę do katastrofy prawnej. Kiedy dział sprzedaży utrzymuje, że nowo wdrożony chatbot to „tylko pomocny asystent", należy bezwzględnie zweryfikować warstwę walidacji wyjścia (*output handling*). Precedensowa sprawa *Moffatt v. Air Canada* udowadnia, że sądy traktują chatboty zintegrowane z witrynami komercyjnymi nie jako wyizolowane, nieodpowiedzialne podmioty, lecz jako cyfrowe ramię przedsiębiorstwa. Za wprowadzające w błąd oświadczenia (*negligent misrepresentations*) wygenerowane wskutek halucynacji modelu organizacja ponosi bezpośrednią i pełną odpowiedzialność odszkodowawczą i kontraktową. Zob. *Moffatt v. Air Canada*, 2024 BCCRT 149.

[5] Tolerowanie praktyk Shadow AI w strukturach firmowych to zaproszenie do katastrofy własności intelektualnej i ochrony danych. Niekontrolowane wprowadzanie dokumentów do publicznych API oraz korzystanie z wektorów wiedzy o nieustalonym pochodzeniu naraża firmę na „zakażenie" łańcucha modeli (*supply chain poisoning*). Historyczna ugoda opiewająca na 1,5 miliarda dolarów w precedensowej sprawie zbiorowej

nych Flag w Podsumowaniu dla Kierownictwa i uruchamiać decyzję o izolacji, autoryzacji albo zatrzymaniu projektu.

Nie czytaj tej książki liniowo, ale zadaniowo. Po każdym większym fragmencie zatrzymaj się i odpowiedz sobie na 5 prostych pytań. Co w mojej organizacji faktycznie się dzieje? Kto za to odpowiada? Jakie dane i zasoby są wpięte w proces? Jakie Zielone Światło daje ten proces? Jaka Czerwona Flaga może go zatrzymać? To nie jest ćwiczenie stylistyczne. To jest metoda odzyskiwania kontroli nad projektem, zanim kontrolę przejmie przyzwyczajenie użytkowników, dostawca chmurowy albo pozornie „inteligentne" obejście wdrożone przez zespół pod presją terminu.

Korzystaj z tej książki zespołowo, ale bez rozmywania odpowiedzialności. Zarząd podejmuje ostateczne decyzje na podstawie formalnego Podsumowania dla Kierownictwa. AI Officer koordynuje całość i działa jako architekt zgodności, który tłumaczy abstrakcyjne wymogi prawne na inżynieryjne ograniczenia, logikę procesu i bramki decyzyjne. Organizacja nie powinna chować się za rozmytym „komitetem" albo „grupą roboczą", która spotyka się raz w miesiącu i niczego nie umie zatrzymać. Potrzebny jest operacyjny Zespół wdrożeniowy AI z prawem do eskalacji i z mandatem do aktywacji wyłącznika awaryjnego[6]. Właściciel Procesu odpowiada za sens biznesowy, sposób pracy ludzi, wpływ wdrożenia na prawa podstawowe i ocenę FRIA lub DPIA tam, gdzie jest to potrzebne. Właściciel Zasobu odpowiada za architekturę techniczną, cyberbezpieczeństwo, integracje API, weryfikację SBOM, logowanie i odporność środowiska[7].

Ważne

Tu obowiązuje jedna twarda zasada. Dla każdego procesu i dla każdego zasobu wpisujesz do raportu imię i nazwisko decydenta. Nie „dział HR". Nie „IT". Nie „biznes". Organ nadzorczy, audytor, sąd i własny zarząd nie potrzebują nazwy jednostki organizacyjnej. Potrzebują odpowiedzi na py-

Bartz v. Anthropic dowodzi, że budowanie lub dostrajanie modeli na podstawie treści pozyskanych z naruszeniem licencji autorskich (np. z tzw. *shadow libraries*) jest traktowane przez sądy jako nieodwracalne piractwo, które z góry wyklucza jakiekolwiek powołanie się na doktrynę dozwolonego użytku (*fair use*). Zob. *D. Hansen*, The Bartz v. Anthropic Settlement: Understanding America's Largest Copyright Settlement, Kluwer Copyright Blog, 10 November 2025, https://legalblogs.wolterskluwer.com/copyright-blog/the-bartz-v-anthropic-settlement-understanding-americas-largest-copyright-settlement/ (dostęp: 17.3.2026 r.).

6 Rozmywanie odpowiedzialności personalnej w ciałach kolegialnych stanowi drastyczny błąd w świetle standardów ładu korporacyjnego. Skuteczny system kierowania technologicznym rozwojem nie może funkcjonować jako wirtualny komitet; musi być operacyjnym organem wspierającym zarząd. Zgodnie z pryncypiami normy ISO 37000 oraz wytycznymi ISO/IEC 38507 właściwy Governance bazuje wyłącznie na systemie międzyludzkim opartym na imiennie przypisanych uprawnieniach nadzorczych. Tworząc Zespół Wdrożeniowy AI, należy mu przekazać bezwzględne, twarde umocowanie w postaci prawa do eskalacji, ewaluacji nowych wektorów ryzyka oraz uruchomienia wyłącznika awaryjnego w momencie uchybienia normom proporcjonalności i poszanowania prawa. Zob. ISO 37000:2021, Governance of organizations – Guidance oraz ISO/IEC 38507:2022, Information technology – Governance of IT – Governance implications of the use of artificial intelligence by organizations.

7 Odpowiedzialność techniczna musi być personalnie przypisana do określonego decydenta, zgodnie z wymogami ustawowymi dotyczącymi odporności infrastruktury. Właściciel Zasobu odpowiada za egzekwowanie cyberbezpieczeństwa, definiowanego jako absolutna odporność systemów informacyjnych na działania uderzające w poufność, integralność (np. poprzez *data poisoning*) i dostępność przetwarzanych danych. Na szczeblu strategicznym kierownik podmiotu kluczowego musi zapewnić, by integracje API, telemetria oraz warstwa RAG działały w oparciu o pełną widoczność zagrożeń. Zob. CyberbezpU, w tym w szczególności definicje legalne i obowiązki kierownika podmiotu kluczowego.

tania, kto podejmował decyzje, kto miał mandat, kto widział ryzyko i kto akceptował wynik. Projekt bez imiennego właściciela bardzo szybko staje się projektem wspólnym, czyli niczyim. A projekt niczyj to naturalne środowisko dla Shadow AI, *automation bias* i niekontrolowanego dryfu odpowiedzialności.

Ta książka ma ci też pomóc odróżnić wdrożenie sensowne od wdrożenia modnego. Nie każde użycie AI powinno zostać doprowadzone do produkcji. Z mojego doświadczenia wiele organizacji za długo utrzymuje projekty, które dobrze wyglądają w slajdach, a słabo działają w pracy. Jeżeli przez kilka miesięcy system generuje więcej ręcznej poprawy niż wartości, zwiększa ryzyko prawne, spowalnia operatorów albo wytwarza chaos w odpowiedzialności, zamknięcie projektu jest oznaką dojrzałości, a nie porażki. Silnik Optymalizacji nie ma służyć uzasadnianiu wdrożeń za wszelką cenę. Ma służyć uczciwej decyzji.

Twoja operacyjna lista zadań:

1. Otwórz fizycznie Raport AI Governance & Compliance oraz Procedurę wdrożenia i prowadź je równolegle z lekturą.

2. Wybierz jeden realny proces biznesowy i na nim wykonuj kolejne kroki, bez skrótów i bez przeskakiwania do zabezpieczeń.

3. Wpisz imiennie Właściciela Procesu i Właściciela Zasobu do odpowiednich rubryk raportu.

4. Przygotowuj od początku Podsumowanie dla Kierownictwa, w którym kompilujesz Zielone Światła i Czerwone Flagi.

5. Dokumentuj każdy stan faktyczny, decyzję projektową, ograniczenie architektoniczne i każde zabezpieczenie, bo z tego powstaje Teczka Obronna[8].

6. Traktuj Shadow AI jak krytyczne ryzyko operacyjne, a nie poboczny incydent.

7. Nie projektuj warstwy kontrolnej, dopóki nie zmapujesz łańcucha pochodzenia danych, nie rozpiszesz przepływów pracy i nie ocenisz swojej roli jako Podmiotu stosującego albo Dostawcy.

2. Cztery fazy, 16 kroków

Skuteczne zarządzanie sztuczną inteligencją wymaga systemu, który pracuje razem z organizacją. Modele są aktualizowane. Dostawcy zmieniają funkcje, zakres usług i regulaminy. Dane wejściowe dryfują. Użytkownicy budują własne obejścia. Integracje API rosną

[8] Wspieraj kulturę dokumentowania każdego etapu wdrożenia, w tym łańcuchów pochodzenia danych, architektonicznych zapór (*guardrails*) oraz decyzji ról biznesowych, jako że stanowi to kamień węgielny Teczki Obronnej. W przypadku postępowań przed potężnymi organami regulacyjnymi w USA czy UE brak namacalnych logów legalności treningu nie kończy się na karze finansowej. Rozpowszechnienie się orzecznictwa sankcyjnego w postaci konfiskaty algorytmicznej (*algorithmic disgorgement*) sprawia, że regulator dysponuje instrumentem przymusowego fizycznego wykasowania całych generacji modeli neuronowych wytrenowanych z domieszką danych nieuprawnionych, nieodwracalnie niweczące wielomilionowe inwestycje oparte na brudnym, niezaudytowanym środowisku startowym. Zob. *K. Doyle*, Algorithmic Disgorgement: Destruction of AI Models as FTC Remedy, U. Richmond J.L. & Tech. 2023.

szybciej niż dokumentacja. Z tego powodu **cała metodologia tej książki została oparta na cyklu zaplanuj – wykonaj – sprawdź – popraw, czyli *Plan-Do-Check-Act***. Ten układ nie jest ozdobą. Daje organizacji rytm pracy, który pozwala stale odświeżać ocenę ryzyka, utrzymywać gotowość kontrolną i nie gubić związku między prawem, architekturą i procesem biznesowym.

Żeby zasypać lukę Lex-Machina, projekt podzieliłem na 4 fazy i 16 kroków. Każda faza ma własny cel operacyjny i własne odzwierciedlenie w załączonym Raporcie AI Governance & Compliance. Faza Identyfikacji wiąże się przede wszystkim z częścią 5A i 5B raportu, gdzie inwentaryzujesz zasoby AI i infrastrukturę wspierającą. Faza Oceny pracuje przede wszystkim na części 3 i 4 raportu, gdzie uruchamiasz ARIA dla operacji wewnętrznych i dla operacji dostawcy. Wnioski z tej fazy trafiają następnie do części 2, czyli do Podsumowania dla Kierownictwa z Zielonymi Światłami i Czerwonymi Flagami. Fazy Wdrażania i Stosowania żyją z kolei w części 6 raportu, która służy do dokumentowania obowiązków ogólnych, szkoleń, zasad AUP, działań korygujących i reakcji na incydenty.

Dobrze to widać w prostym mapowaniu przedstawionym w tabeli 1.

Tabela 1. Fazy wdrożenia systemu AI a odpowiadające im dowody audytowe

Faza	Co robisz	Gdzie zostawiasz ślad dowodowy
Identyfikuj	Ustalasz procesy, *workflow*, zasoby, role i wstępne klasy ryzyka	Część 5A i 5B raportu oraz rubryki właścicieli
Oceń	Uruchamiasz ARIA, liczysz Zielone Światła i Czerwone Flagi	Część 3, 4 i 2 raportu
Wdróż	Przekładasz ustalenia na środki organizacyjne i techniczne	Część 6 raportu oraz Teczka Obronna
Stosuj	Mierzysz, szkolisz, audytujesz, bronisz i poprawiasz	Część 6 raportu, logi, standardy, Teczka Obronna

Źródło: Opracowanie własne.

W praktyce każda faza odpowiada na inne pytanie. Gdzie AI działa? Czy warto ją tu utrzymać? Jak ją zabezpieczyć? Jak udowodnić, że nadal działa pod kontrolą?

Ważne

Nie mieszaj faz. Organizacje najczęściej przegrywają nie dlatego, że brakuje im chęci, lecz dlatego, że zaczynają od końca. Najpierw piszą polityki. Potem zamawiają szkolenia. Następnie kupują *middleware*. A dopiero później zadają pytanie, czy w ogóle rozumieją proces, który chcą zabezpieczać. Taka kolejność daje iluzję profesjonalizmu, ale nie daje kontroli. Najpierw rozpoznaj rzeczywistość operacyjną. Potem ją oceń. Następnie projektuj osłonę. Na końcu egzekwuj i utrzymuj. To jest sekwencja robocza. Nie skracaj jej.

2.1. Faza 1: Identyfikuj

Ta faza służy odzyskaniu widoczności. Interesuje cię stan faktyczny, a nie deklarowany. Chodzi o **pełne zmapowanie miejsc, w których AI już pracuje, ma zacząć pracować**

albo działa bez formalnej zgody organizacji. W wielu firmach największe ryzyko nie kryje się w zatwierdzonych projektach, lecz w nieoficjalnym wykorzystaniu publicznych chatbotów, automatyzacji *no-code*, prywatnych narzędzi do transkrypcji, własnych agentów podłączonych do skrzynki mailowej albo w pospiesznym fine-tuningu modelu wykonanego „tylko na potrzeby testu". Ta faza kończy się dopiero wtedy, gdy potrafisz narysować pełną mapę procesu, danych, zasobów i odpowiedzialności.

2.1.1. Krok 1: Zidentyfikuj procesy biznesowe

Zacznij od szerokich obszarów działania organizacji. Rekrutacja. Obsługa klienta. Księgowość. Sprzedaż. Marketing. Analiza umów. Cyberbezpieczeństwo. Operacje zakupowe. Nie pytaj tylko, „czy używamy AI". Pytaj, **gdzie AI wpływa na czas, jakość, koszt, zgodność albo decyzję**.

Przykład

Jeżeli dział prawny używa modelu do streszczania umów, a następnie te streszczenia wpływają na negocjacje, to nie jest ciekawostka. To element procesu biznesowego. Jeżeli centrum obsługi klienta korzysta z modelu do sugerowania odpowiedzi, a konsultant zwykle je zatwierdza bez zmian, to AI realnie kształtuje komunikację z klientem. Wpisz te procesy do raportu od razu, zamiast liczyć na to, że „wszyscy wiedzą, gdzie AI działa".

2.1.2. Krok 2: Zidentyfikuj przepływy pracy

Teraz schodzisz z poziomu procesu na poziom sekwencji działań. Kto uruchamia model? Jakie dane wkłada do systemu? Z jakiego źródła pochodzą te dane? Gdzie trafia wynik? Kto z niego korzysta? Czy wynik uruchamia kolejne kroki automatycznie? Czy człowiek ma realną możliwość zatrzymania albo zmiany rekomendacji? W tym miejscu trzeba też bardzo **ostro odróżnić rolę Podmiotu stosującego od roli Dostawcy**. To nie jest kwestia słownika. To kwestia reżimu prawnego. Jeżeli organizacja bierze gotowe narzędzie SaaS i używa go zgodnie z przeznaczeniem, pozostaje zwykle Podmiotem stosującym. Jeżeli jednak zaczyna istotnie modyfikować system wysokiego ryzyka, w tym przez dalsze trenowanie zmieniające jego przeznaczenie, ryzykuje wejście w rolę Dostawcy z całym ciężarem dokumentacji technicznej, oceny zgodności i certyfikacji. Ten moment trzeba wyłapać wcześniej, nie po wdrożeniu.

2.1.3. Krok 3: Zidentyfikuj zasoby

W praktyce chodzi o zbudowanie pełnego rejestru tego, co zasila albo otacza AI. Modele SaaS. Integracje API. Modele lokalne i *open-weight*. Bazy danych wektorowych. Repozytoria wiedzy. Skrypty automatyzujące. Środowiska testowe. Legacy IT. Współdzielone foldery. Magazyny danych. Karty modeli. Zestawienia SBOM. To właśnie tutaj wychodzi na jaw, czy twój chatbot odpowiada wyłącznie na podstawie zatwierdzonej bazy wiedzy, czy może czyta wszystko, do czego ktoś kiedyś nadał zbyt szerokie uprawnienia. To tutaj widzisz, czy asystent AI streszcza poufne pliki płacowe, bo błędnie skonfigurowano indeksowanie, albo czy agent wykonuje działania na zasobach, których nikt nie objął realnym nadzorem. Dobrze zrobiona inwentaryzacja to początek **obrony przed *prompt***

***injection*, zatruciem danych, wyciekiem tajemnicy przedsiębiorstwa i konfiskatą algorytmiczną** wynikającą z braku dowodu legalnego pochodzenia danych[9].

2.1.4. Krok 4: Zidentyfikuj kategorie ryzyka i właścicieli procesów i zasobów

Na tym etapie ustalasz dwie rzeczy naraz. Po pierwsze, **do jakiej klasy ryzyka z grubsza zmierza wdrożenie**. Minimalne, ograniczone, wysokie, a może niedopuszczalne. Po drugie, **kto za nie odpowiada**. Właściciel Procesu odpowiada za sens wdrożenia, zmianę sposobu pracy i skutki dla ludzi. Właściciel Zasobu odpowiada za warstwę techniczną, bezpieczeństwo, integracje i odporność środowiska. W wielu organizacjach właśnie tu zaczyna się zdrowie projektu, bo kończy się wygodne stwierdzenie, że „to wspólna inicjatywa". Nie. Tu kończy się wspólność, a zaczyna **odpowiedzialność**. Dobra praktyka polega na tym, że po tym kroku potrafisz wskazać nie tylko system, ale człowieka odpowiedzialnego za jego cel biznesowy i człowieka odpowiedzialnego za jego techniczną warstwę działania.

Po fazie identyfikacji nie powinieneś już działać na intuicji. Masz wiedzieć, czy rekruterzy wklejają CV do publicznego narzędzia, czy dział zakupów używa generatywnego asystenta do redagowania negocjacji, czy dział cyberbezpieczeństwa korzysta z modelu do klasyfikacji alertów, **czy organizacja jest tylko Podmiotem stosującym, czy już ociera się o status Dostawcy**. Jeżeli tego nie wiesz, każdy kolejny krok będzie budowany na piasku.

2.2. Faza 2: Oceń

Tutaj uruchamiasz protokół ARIA i wchodzisz w serce całej metodyki. Ta faza działa na Podwójnym Silniku. Silnik Optymalizacji pyta, czy wdrożenie daje realny efekt biznesowy i czy organizacja ma zdolność operacyjną, by z tego efektu korzystać. Silnik Ryzyka pyta, jak poważna może być szkoda i jak łatwo może dojść do naruszenia. **Te silniki nie pracują osobno. Właśnie ich zderzenie daje uczciwy obraz wdrożenia.** To tutaj otwierasz część 3 raportu dla operacji wewnętrznych albo część 4 dla operacji dostawcy i prowadzisz ocenę w taki sposób, aby wnioski od razu zasilały Podsumowanie dla Kierownictwa. Zielone Światło bez oceny ryzyka nic nie znaczy. Czerwona Flaga bez odniesienia do wartości biznesowej też jest niepełna. Dopiero zestawienie obu daje dobrą decyzję.

---

[9] Rygorystyczna inwentaryzacja zasobów i ścisła kontrola uprawnień (RBAC) to jedyna skuteczna tarcza przed wektorami ataków takimi jak *prompt injection* czy zanieczyszczenie łańcucha treningowego (*data poisoning*). Co więcej, to kluczowy mechanizm obronny przed interwencją najsurowszych organów regulacyjnych. Brak dowodów legalnego pochodzenia danych treningowych naraża organizację na zjawisko „konfiskaty algorytmicznej" (*algorithmic disgorgement*) – radykalnej sankcji stosowanej m.in. przez Federalną Komisję Handlu (FTC), polegającej na nakazie całkowitego fizycznego zniszczenia modeli sztucznej inteligencji, które wyuczono na bazach zgromadzonych z naruszeniem rygorów prywatności i dostępu, co miało miejsce w głośnych postępowaniach wobec podmiotów Big Tech. Zob. FTC, FTC Sends Refunds to Ring Customers Stemming from 2023 Settlement over Charges the Company Failed to Block Employees and Hackers from Accessing Consumer Videos, 23 April 2024, https://www.ftc.gov/news-events/news/press-releases/2024/04/ftc-sends-refunds-ring-customers-stemming-2023-settlement-over-charges-company-failed-block (dostęp: 16.3.2026 r.).

2.2.1. Krok 5: Oceń możliwości

Nie pytaj, czy AI „robi wrażenie", ale czy usuwa konkretne wąskie gardło. Czy skraca czas cyklu? Czy odciąża ludzi od pracy powtarzalnej? Czy poprawia jakość pierwszej wersji dokumentu? Czy zwiększa przepustowość obsługi klienta? Czy przyspiesza analizę reklamacji? Czy daje przewidywalny zwrot z inwestycji? Tutaj powstają Zielone Światła. **Ale nie opisuj ich ogólnie.** Zapisz w raporcie stan obecny, stan docelowy, miernik czasu, miernik jakości, koszt ręcznej obsługi, koszt poprawy błędów i gotowość zespołu do pracy z narzędziem. Jeżeli organizacja nie ma przeszkolonego personelu, sensownych danych wejściowych albo stabilnej infrastruktury, to nawet dobre narzędzie nie daje jeszcze Zielonego Światła. Daje co najwyżej obietnicę.

2.2.2. Krok 6: Oceń zgodność

Teraz przekładasz przepisy na warunki wdrożenia. Sprawdzasz, czy nie wchodzisz w praktyki zakazane, czy występują obowiązki przejrzystości, czy pojawia się problem danych osobowych, tajemnicy przedsiębiorstwa, ograniczeń sektorowych, wymogów kontraktowych albo transferów transgranicznych. Tu również sprawdzasz, czy umowa z dostawcą nie zostawia mu zbyt szerokiej swobody w dalszym wykorzystywaniu danych, czy istnieją zasady retencji i usuwania danych, czy prawo do usunięcia danych jest technicznie wykonalne w architekturze AI i czy można obronić zakres dostępu do logów. **To nie jest etap na ogólną deklarację, że „dział prawny już spojrzał".** Tu trzeba przetłumaczyć przepis na konkretne warunki uruchomienia procesu[10].

2.2.3. Krok 7: Oceń wagę zagrożeń

To jest pytanie o skalę szkody, nie o prawdopodobieństwo. Co się stanie, jeżeli model się pomyli? Kto poniesie koszt? Czy szkoda uderzy w człowieka, jego prawa podstawowe, bezpieczeństwo fizyczne, reputację, dostęp do usługi, zatrudnienie albo zdolność organizacji do kontynuowania działania[11]? Jeżeli system działa w obszarze zatrudnienia, kredytowania, dostępu do świadczeń, edukacji albo w innym obszarze wysokiej stawki, nie

[10] Autoryzacja przetwarzania danych przez rozwiązania SaaS staje się wysoce rygorystyczna. Kontraktowa weryfikacja dostawców musi obejmować twarde wyłączenie prawa do wtórnego wykorzystywania wprowadzanych logów do tzw. „udoskonalania modeli" (*retraining*). Europejska Rada Ochrony Danych w wytycznych uderzających bezpośrednio w rynek LLM (Opinia 28/2024) narzuca skumulowany, ścisły test prawnie uzasadnionego interesu dla fazy deweloperskiej. Ponadto organ nadzorczy wprost wskazuje na ryzyka re-identyfikacji i ataków inwersyjnych (*model inversion*) polegających na ekstrakcji zapisanych danych z wag architektonicznych samego modelu, co bezwzględnie wymaga ustanowienia technicznie wykonalnego prawa do bycia zapomnianym poprzez politykę Zero-Data Retention. Zob. EDPB, Opinion 28/2024 on certain data protection aspects related to the processing of personal data in the context of AI models, 17 December 2024.

[11] Silnik Ryzyka ewaluujący wagę zagrożeń musi priorytetyzować analizę skali szkody dla praw człowieka, pomijając na tym etapie rozważania o statystycznym prawdopodobieństwie błędu. Ekstremalnym przypadkiem urzeczywistnienia tego ryzyka są systemy klasy SyRI. Orzeczenie Sądu Okręgowego w Hadze dobitnie wykazało, że zautomatyzowane platformy analizujące masowe, wrażliwe zbiory danych w celu przewidywania oszustw lub profilowania behawioralnego, funkcjonujące w reżimie całkowitej nieprzejrzystości (*black-box*), łamią gwarancje art. 8 EKPCz. Działania polegające na nieproporcjonalnej inwigilacji algorytmicznej wymuszają na projektantach wprowadzenie kategorycznej zasady rozliczalności oraz wymogu FRIA przed dopuszczeniem systemu do ruchu. Zob. Rechtbank Den Haag, 5 februari 2020, C/09/550982 / HA ZA 18-388, ECLI:NL:RBDHA:2020:865 (SyRI).

wystarczy ogólna ocena etyczna[12]. W materiałach tej książki **FRIA jest traktowana jako twardy warunek progowy dla określonych zastosowań wysokiego ryzyka**. To znaczy, że jeżeli analiza wpływu na prawa podstawowe pokazuje szkodę nieproporcjonalną albo trudną do odwrócenia, projekt nie przechodzi dalej tylko dlatego, że daje dobre ROI. To właśnie tutaj Silnik Ryzyka ma prawo powiedzieć „stop"[13].

2.2.4. Krok 8: Oceń prawdopodobieństwo naruszeń

Dopiero teraz pytasz, jak łatwo może dojść do błędu lub ataku. Analizujesz *prompt injection*, zatrucie danych, kaskady halucynacji, błędy integracyjne, słabą kontrolę dostępu, niewłaściwą izolację środowisk, błędne uprawnienia do baz wiedzy, niską jakość danych wejściowych, *automation bias* po stronie operatora i zbyt dużą autonomię agentów. W praktyce to tutaj dobrze wychodzi **różnica między ryzykiem procesu a ryzykiem zasobu**. Ten sam chatbot może być względnie bezpieczny w generowaniu szkiców wewnętrznych notatek i bardzo niebezpieczny wtedy, gdy wolno mu wysłać gotową wiadomość do klienta albo uruchomić działanie w systemie ERP. Prawdopodobieństwo naruszenia mierzy się więc nie tylko przez podatność techniczną modelu, lecz także przez to, gdzie i jak model został wpięty w działanie organizacji[14].

Po fazie oceny powinieneś mieć już coś więcej niż opis projektu – decyzję zarządczą w formie, którą da się obronić. Co daje Zielone Światło i dlaczego? Co generuje Czerwoną Flagę i dlaczego? Jaki jest próg uruchomienia? Jaki jest próg zatrzymania? Jakie warunki trzeba spełnić, zanim projekt przejdzie dalej? Taki sposób pracy skraca spory wewnętrzne, bo zamienia chaos opinii na mierzalny zapis w raporcie. Właśnie tu rodzi się prawdziwa Szybkość decyzyjna[15].

---

12 Oceny powagi zagrożenia nie wolno sprowadzać do ryzyka wizerunkowego. Musi ona obejmować precyzyjne mapowanie naruszeń praw podstawowych człowieka, w tym bezpieczeństwa fizycznego, dostępu do usług, zatrudnienia oraz prawa do decydowania o swoim życiu osobistym. Konstrukcja zautomatyzowanych systemów decyzyjnych w obszarach wrażliwych musi bezwzględnie szanować przyrodzoną i niezbywalną godność jednostki, która na terytorium RP stanowi nienaruszalne źródło wszelkich wolności i granicę dopuszczalnej automatyzacji procesów. Zob. Konstytucja RP, art. 30, 47.

13 Wdrożenie procedury FRIA dla systemów wysokiego ryzyka wymaga porzucenia deklaratywnej etyki na rzecz rygorystycznej metodologii inżynieryjno-prawnej, operacjonalizującej konwencję Rady Europy. Instrumentem do tego służącym jest metodologia HUDERIA, ugruntowana na szczegółowej, opartej na kontekście analizie ryzyka (COBRA). Kluczowym wektorem wdrożeniowym stają się tu twarde „pytania zerowe" (*Zero Questions*), które pełnią w organizacji funkcję nadrzędnej bramki bezpieczeństwa (*Kill Switch*). Jeżeli w ramach procesu analizy proporcjonalności i wpływu na prawa podstawowe odpowiedzi wykazują trudną do skompensowania nierównowagę lub brak ścieżki obronnej dla osoby fizycznej, projekt zostaje kategorycznie wstrzymany, bez względu na obiecywaną efektywność kosztową (ROI). Zob. Council of Europe, HUDERIA (Human Rights, Democracy and the Rule of Law Impact Assessment) methodology for AI.

14 Oceniając podatność organizacyjną i prawdopodobieństwo szkody, należy kategorycznie oddzielić systemy o funkcjonalności doradczej od tych pełniących rolę nadzorczą wobec personelu. W przypadku wykorzystania systemów klasy HR do monitorowania behawioralnego i zwalniania pracowników organizacja nie może skryć się za parawanem zawiłości kodu ani za ochroną własności intelektualnej. Przełomowy wyrok Sądu Okręgowego w Amsterdamie, dotyczący niejawnych zwolnień dyscyplinarnych kierowców platformowych wygenerowanych przez modele klasyfikujące rzekome oszustwa, dobitnie dowodzi, iż zatajanie logiki decyzyjnej pod pretekstem ochrony algorytmu jest prawnie nieproporcjonalne i nie obroni się przed żądaniem pełnej przejrzystości. Zob. Amsterdam District Court/Court of Appeal, *App Drivers & Couriers Union and Worker Info Exchange v. Uber BV and Ola Netherlands BV*.

15 Prawidłowa ewaluacja wektorów zagrożenia technicznego zmusza architektów do myślenia w kategoriach układów socjotechnicznych, a nie jedynie statystycznej podatności samego kodu. Ocenę prawdopodobieństwa

2.3. Faza 3: Wdróż

Tutaj kończy się analiza, a zaczyna architektura. **Faza wdrożenia służy przekuciu ustaleń z ARIA w realne bariery, procedury i ustawienia techniczne.** To jest moment, w którym *compliance* przestaje być dokumentem, a zaczyna być częścią działania systemu. Jeżeli dobrze przeprowadziłeś fazy wcześniejsze, teraz nie budujesz zabezpieczeń „na wszelki wypadek". Budujesz je dokładnie tam, gdzie proces i raport pokazały, że są potrzebne. Dzięki temu nie dusisz lekkich zastosowań nadmiarem formalności, a jednocześnie nie zostawiasz bez osłony obszarów, w których stawka dla człowieka, organizacji albo rynku jest wysoka.

2.3.1. Krok 9: Zoptymalizuj przepływy pracy

Nie dokładaj AI do złego procesu. Najpierw przeprojektuj sam *workflow*. Ustal, gdzie AI ma wspierać człowieka, a gdzie człowiek ma przejąć ster. Zaprojektuj znaczącą kontrolę człowieka w punktach krytycznych. Zdecyduj, czy wynik modelu jest szkicem, rekomendacją, filtrem czy automatycznym wyzwalaczem. **W tym kroku planujesz też tryb ręczny i wyłącznik awaryjny.** Jeżeli dostawca chmurowy ma awarię, model zaczyna halucynować albo agent wykonuje niepożądane akcje, organizacja ma nadal działać. Dobry Kill Switch nie jest ozdobą dla audytu. To warunek utrzymania ciągłości działania i zdrowego rozsądku w relacji człowiek – maszyna.

2.3.2. Krok 10: Wdróż środki organizacyjne

Tu formalizujesz reguły gry. Politykę Dopuszczalnego Użytkowania AI. Wymóg ukończenia szkolenia przed nadaniem dostępu. Zasady pracy z promptami i danymi wrażliwymi. Zasady negocjowania postanowień umownych z dostawcami. Zasady używania własnych baz wiedzy. Zasady retencji Zero-Day dla określonych kategorii danych lub szybkiego usuwania danych, które nie powinny zostać utrwalone w logach czy pamięci pomocniczej. **W praktyce dobrze napisane środki organizacyjne odpowiadają użytkownikowi na bardzo przyziemne pytania.** Czy wolno wkleić całe CV? Czy wolno użyć danych klienta do przygotowania prezentacji? Czy wolno podłączyć zewnętrzne API bez zgody? Czy wolno uruchomić agenta na skrzynce mailowej? Jeśli Polityka Dopuszczalnego Użytkowania AI nie odpowiada na takie pytania, to nie steruje zachowaniem ludzi.

2.3.3. Krok 11: Wdróż środki techniczne

To jest warstwa, w której prawo ma zejść do kodu i konfiguracji. Kontrola dostępu oparta na rolach. Separacja środowisk. Semantyczne zapory i *middleware* filtrujące wejścia i wyj-

usterki należy mapować z chirurgiczną precyzją, analizując nie tylko klasyczne ataki typu *prompt injection* czy intencjonalne zanieczyszczenie łańcucha zasobów (*data poisoning*), lecz także kaskady tzw. konfabulacji (halucynacji modelu) oraz degradację wydajności w architekturach RAG. Zgodnie z najnowszym profilem NIST dedykowanym generatywnej sztucznej inteligencji ocena ta obejmuje błędy integracyjne logiki modelu w systemach chmurowych, rozlanie uprawnień autoryzacyjnych do repozytoriów współdzielonych oraz uprzedzenia algorytmiczne (bias), na które nakłada się zawsze presja decyzyjna wywierana na samego użytkownika. Zob. NIST, Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile, NIST AI 600-1, 2024 (w kontekście zjawiska konfabulacji i podejścia socjotechnicznego).

ścia. Maskowanie danych przed wektoryzacją. Ochrona baz wiedzy. Ograniczenia uprawnień agentów. Niezmienne logi. Telemetria. Weryfikacja SBOM. Sanityzacja danych wejściowych. Wszystko to brzmi technicznie, ale ma bardzo prosty sens biznesowy. Chodzi o to, żeby twoja organizacja mogła wykazać nie tylko, że „miała politykę", lecz także **że zbudowała realne bariery** przeciwko wyciekowi danych, nadużyciu dostępu, zatruciu łańcucha dostaw i błędom autonomicznych działań. Bez tej warstwy nie ma dobrego mostu Lex-Machina. Są tylko deklaracje[16].

2.3.4. Krok 12: Zrealizuj ogólne obowiązki

Na tym etapie wdrażasz przejrzystość, etykietowanie treści generowanych przez AI, kulturę kwestionowania wyniku modelu i środki przeciwdziałające ślepemu ufaniu automatyzacji. Interfejs nie może prowokować *rubber-stamping*, czyli mechanicznego zatwierdzania odpowiedzi maszyny. Operator powinien mieć **realny czas, właściwy kontekst i czytelne źródła, żeby ocenić wynik**[17]. To jest też moment, w którym dokumentujesz szkolenia z kompetencji AI. Obowiązek zapewnienia odpowiedniego poziomu AI literacy wynika wprost z AI Act i dotyczy zarówno dostawców, jak i podmiotów stosujących, z uwzględnieniem kontekstu użycia systemu i osób, których system dotyczy. W praktyce dostęp do narzędzia powinien iść w parze z dowodem, że użytkownik rozumie ograniczenia modelu, zasady użycia i próg eskalacji.

Po fazie wdrożenia powinieneś mieć coś więcej niż działający system. Powinieneś mieć **system, który działa w sposób ograniczony, opisany i gotowy do obrony**. Człowiek wie, kiedy ufać wynikowi, kiedy go zakwestionować i kiedy uruchomić tryb ręczny. Architektura techniczna wie, co wolno przepuścić, a co trzeba zablokować. Zarząd wie, jaki jest próg akceptacji ryzyka. To właśnie tu zgodność przestaje być kosztem ubocznym projektu, a zaczyna być operacyjnym warunkiem jego skalowania.

---

16 Krok ten stanowi ostateczną architektoniczną implementację wymogów bezpieczeństwa, w której litera prawa materializuje się w rygorystycznym kodzie obronnym. Budowa godnego zaufania środowiska operacyjnego opiera się na wytycznych technicznych organów powołanych do ochrony infrastruktury w UE. Oznacza to obligatoryjne wdrażanie ścisłej kontroli RBAC, odizolowanie poligonów doświadczalnych od środowisk z danymi produkcyjnymi oraz instalację tzw. zapór semantycznych (*semantic middleware routing*). Oprogramowanie pośredniczące, dokonujące dynamicznej sanityzacji strumieni wejścia i wyjścia, neutralizuje podatności modeli na ataki inwersyjne i iniekcję złośliwych instrukcji, blokując je fizycznie przed zintegrowaniem ich w cykl wnioskowania sztucznej inteligencji. Zob. ENISA, Securing Machine Learning Algorithms, 2021 (w tym w szczególności zalecenia w zakresie *semantic middleware routing* oraz mitygacji podatności AI).

17 Przejrzystość decyzyjna jest wymogiem ustawowym, z którego nie zwalnia złożoność technologiczna modelu. Zgodnie z wytycznymi europejskich organów nadzorczych dotyczącymi zautomatyzowanego podejmowania decyzji (ADM) i profilowania osoba dotknięta werdyktem algorytmu dysponuje niezbywalnym prawem do wyjaśnienia (*right to explanation*) logiki stojącej za decyzją systemu. Z perspektywy operacyjnej projektant interfejsu systemu AI jest zobligowany do zapewnienia operatorowi ustrukturyzowanych danych dowodowych (w tym wyjaśnień kontrfaktycznych, bez konieczności całkowitego dekodowania struktury *black-box*), pozwalających człowiekowi na świadome przeanalizowanie wskaźników ryzyka, ich zakwestionowanie w trybie odwoławczym oraz wyeliminowanie mechanicznego zatwierdzania wyników generowanych przez algorytm. Zob. Grupa Robocza Art. 29, Wytyczne w sprawie zautomatyzowanego podejmowania decyzji w indywidualnych przypadkach i profilowania do celów rozporządzenia 2016/679 (WP251rev.01).

2.4. Faza 4: Stosuj

Ostatnia faza zamyka pętlę i od razu otwiera następną. Tutaj system trafia do realnego życia. To moment, w którym użytkownicy zaczynają pracować pod presją czasu, dane się zmieniają, dostawcy aktualizują modele, jakość zaczyna falować, a pięknie napisane zasady są konfrontowane z codziennym nawykiem. Wiele organizacji wpada w tym momencie w pułapkę fałszywego poczucia bezpieczeństwa. Procesy zostały opisane, role rozdzielone, zabezpieczenia (*guardrails*) wdrożone, a szkolenia przeprowadzone. Projekt uznaje się za zakończony. Tymczasem wdrożenie AI tak naprawdę dopiero się zaczyna. Od tej chwili wszystko zależy od tego, czy umiesz mierzyć, audytować, poprawiać i bronić system na bieżąco.

2.4.1. Krok 13: Przygotuj standardy wewnętrzne

Ustal telemetrię, miary jakości, progi błędu, zasady przeglądów, częstotliwość testów, sposób wykrywania dryfu danych i modelu, zakres testów antagonistycznych i Red Teaming AI. Określ też, gdzie potrzebujesz wyjaśnialności i na jakim poziomie. Nie każdy proces wymaga tej samej głębokości XAI. Ale każdy proces wymaga **uczciwej odpowiedzi, jak rozpoznasz, że model przestał działać bezpiecznie**. Standard bez progu reakcji nie steruje działaniem. To tylko opis intencji.

2.4.2. Krok 14: Przyjmij, opublikuj i przeszkol

W tym miejscu dokumenty zyskują moc wiążącą. Zasady mają trafić do ludzi, nie do folderu. Dobrze zrobiony etap wdrożenia kończy się nie tylko publikacją standardu, ale też związaniem dostępu do narzędzia z ukończeniem szkolenia, zrozumieniem zasad AUP i znajomością reguł eskalacji. To także właściwy moment, by **osadzić działanie zapobiegawcze i komunikację wokół incydentów**, skarg oraz podejrzeń naruszeń. W praktyce to tutaj widać, czy twoja organizacja buduje kulturę odpowiedzialnego używania AI, czy tylko archiwum PDF-ów.

2.4.3. Krok 15: Osiągnij rozliczalność

Teraz porządkujesz Teczki Obronne. Gromadzisz logi, wyniki testów, decyzje, akceptacje, macierze RBAC, oceny ARIA, FRIA, dokumenty dostawców, potwierdzenia szkoleń, karty modeli, SBOM i ślady zmian. W tym kroku trzeba **zwrócić szczególną uwagę na realność nadzoru człowieka**. Wyrok TSUE w sprawie SCHUFA pokazuje, że sama obecność człowieka w procesie nie wystarcza, jeżeli jego rola sprowadza się do automatycznego zatwierdzenia wyniku[18]. Dlatego w Teczce Obronnej powinien znaleźć się nie tylko log klik-

[18] Projektowanie nadzoru ludzkiego wymaga wyeliminowania powszechnego w korporacjach zjawiska automatycznego ufania maszynie (*automation bias*). Zgodnie z przełomowym wyrokiem TSUE w sprawie SCHUFA generowanie wartości prawdopodobieństwa (scoringu), która w sposób decydujący determinuje późniejsze ustanowienie, wykonanie lub zakończenie stosunku umownego, wyczerpuje znamiona zautomatyzowanego podejmowania decyzji. Operator zredukowany do funkcji bezrefleksyjnej „pieczątki" nie wyłącza zastosowania rygorów art. 22 RODO, dlatego też Teczka Obronna musi zawierać dowody na istnienie krytycznego tarcia decyzyjnego w interfejsach systemów. Zob. TSUE, *SCHUFA Holding AG*, C-634/21, Legalis (w kontekście zautomatyzowanego podejmowania decyzji i profilowania).