SPOŁECZNY ZESPÓŁ EKSPERTÓW PRZY PUODO

Bilet w jedną stronę

Jak dowieść nieodwracalności anonimizacji w projektach AI i data science?

MARIOLA WIĘCKOWSKA

Agenda

Część 1

Pułapka pseudonimizacji w modelach AI. Dlaczego tradycyjne metody (haszowanie, maskowanie) to za mało w świetle wytycznych EROD 01/2025.

Część 2

Dane syntetyczne jako tarcza chroniąca prywatność. Konstrukcja matematycznych kopii danych rzeczywistych zgodna z Privacy by Design.

Część 3 I 4

Warsztat rozliczalności przed UODO. Metryki PETs, testy re-identyfikacji (single out/wyodrębnienie, linkability/możliwość łączenia, inference/wnioskowanie) oraz checklista IOD.

CZĘŚĆ PIERWSZA

Pułapka pseudonimizacji

Dlaczego tradycyjne podejście IT do modeli sztucznej inteligencji stanowi krytyczne ryzyko prawne i jak chronić administratora

Pułapka pseudonimizacji

Mylenie **bezpieczeństwa danych** z ich **prawną anonimizacją** to najczęstszy błąd architektów IT.

- **Dane spseudonimizowane to nadal dane osobowe** - zgodnie z *Wytycznymi EROD Pseudonimizacja*, usunięcie bezpośrednich identyfikatorów bez całkowitej nieodwracalności nie wyłącza RODO.
- **Haszowanie to nie anonimizacja** - funkcje skrótu bez silnej soli w postaci szumu/klucza są trywialne do odwrócenia metodami brute-force przy użyciu współczesnej mocy obliczeniowej.

Wytyczne nr 01/2025 EROD w sprawie pseudonimizacji. Przyjęto w dniu 16. stycznia 2025 r.
https://www.edpb.europa.eu/our-work-tools/documents/public-consultations/2025/guidelines-012025-pseudonymisation_pl

AI to otwarta księga?

Zapamiętywanie (memorisation)

Modele (zwłaszcza LLM) mają tendencję do zapamiętywania rzadkich wartości ze zbioru treningowego bezpośrednio w wagach sieci neuronowej.

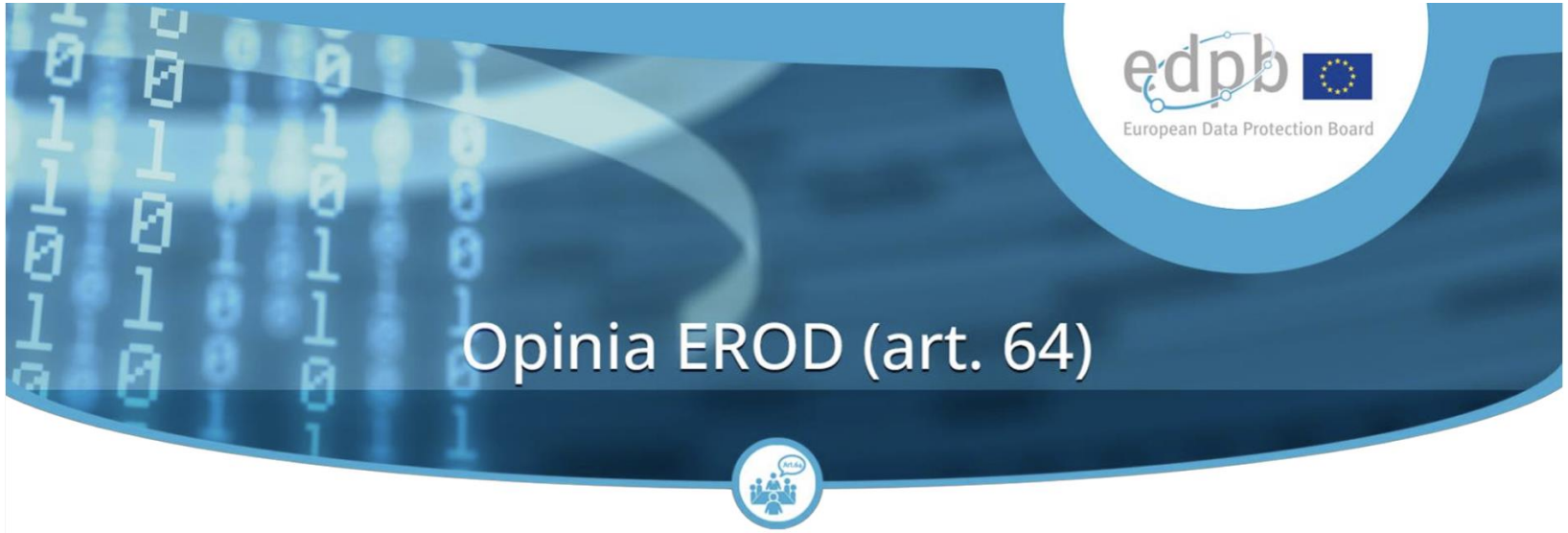
Wnioskowanie (inference)

Poprzez korelację statystyczną model potrafi odtworzyć wrażliwe cechy pacjenta lub klienta, mimo że nie zostały one jawnie podane w zapytaniu.

Wyciek przez interfejs

Odpowiednie manipulowanie promptami (prompt injection) pozwala na obejście filtrów i zmuszenie AI do ujawnienia danych treningowych.

Model jako zbiór danych



Przełomowa ocena statusu prawnego wytrenowanych modeli sztucznej inteligencji:

Zgodnie z Opinią 28/2024 w sprawie niektórych aspektów ochrony danych związanych z przetwarzaniem danych osobowych w kontekście modeli AI (przyjętej 17 grudnia 2024 r.) wytrenowany model AI **nie staje się automatycznie anonimowy**.

- Jeśli prawdopodobieństwo odzyskania danych z parametrów modelu jest wyższe niż znikome – **cały model podlega RODO**.
- Samo uczenie maszynowe nie jest uznawane za technologiczną barierę anonimizacyjną.

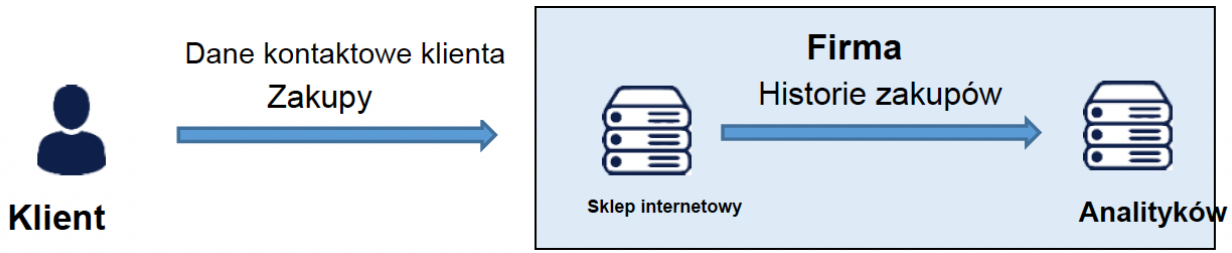
https://www.edpb.europa.eu/system/files/2025-05/edpb_opinion_202428_ai-models_pl_0.pdf

Checklista audytowa IOD

KLUCZOWE WYMOGI WDROŻENIOWE DLA AI:

- **Zakaz uczenia (no-training)** - brak zgody na trenowanie modeli dostawców na danych organizacji.
- **Retencja u źródła** - fizyczne niszczenie pierwotnych zbiorów danych po treningu.
- **Testy podatności -** audyt modelu pod kątem podatności na odzyskanie i wyciek danych.
- **Audytowalność** - zapewnienie prawa do dostępu do raportów, np. z testów wycieku danych u procesora.

Zmniejszenie ryzyka uzasadniające dalsze przetwarzanie



Kontekst i cel przetwarzania	Firma prowadzi duży sklep internetowy z różnymi produktami. Dane o zakupach klientów są przechowywane i prezentowane na kontach klientów. Spółka zamierza pozyskać dane z bazy danych w celu znalezienia korelacji między zakupionymi produktami lub usługami.
Problem do rozwiązania	Ze względu na szerokie spektrum produktów i usług oferowanych przez Sklep internetowy, rejestry zakupów mogą pozwolić na wyciągnięcie istotnych wniosków dotyczących osób, których dane dotyczą, a także mogą pozwolić na ocenę osobistych aspektów związanych z sytuacją ekonomiczną, zdrowiem, osobistymi preferencjami, zainteresowaniami lub zachowaniem osób, których dane dotyczą. Aby dane osobowe mogły zostać uznane za zgodne z celem, dla którego dane osobowe zostały pierwotnie zebrane, i aby uniknąć profilowania klientów zgodnie z kryteriami określonymi w art. 4 ust. 4 RODO, dane muszą być przetwarzane w taki sposób, aby analitycy nie mogli ich już przypisać do konkretnych osób, których dane dotyczą.
Oryginalne dane	– Profil użytkownika – historia zakupów
Domena pseudonimizacji	Zespół Analityków
Dane pseudonimizowane	Historia zakupów z usuniętymi wszystkimi zindywidualizowanymi wpisami (np. odzież z napisem wybranym przez klienta)
Dodatkowe informacje	Oryginalne konto klienta.
Procedura pseudonimizacji	Spółka wyodrębnia historię zakupów z pominięciem wszystkich zindywidualizowanych wpisów i bezpośrednio identyfikujących atrybuty, a analizę przypisuje do Jednostki Organizacyjnej Analityków bez dostępu do dalszych danych o klientach.
Przetwarzanie danych pseudonimizowanych	Analitycy przeprowadzają żądaną analizę i podsumowują wyniki w formie zagregowanej. Po upływie tego okresu Jednostka Organizacyjna usuwa wszystkie posiadane przez nią dane osobowe.
Efekt	Jest mało prawdopodobne, aby przetwarzanie dokonywane w ten sposób miało wpływ na osoby, których dane dotyczą. Administrator może wykorzystać ten efekt pseudonimizacji do oceny zgodności celów zgodnie z art. 6 ust. 4 RODO. Biorąc również pod uwagę inne czynniki wymienione w art. 6 ust. 4 RODO oraz w zależności od specyfiki konkretnego przypadku, ocena może prowadzić do wniosku, że cel analizy można uznać za zgodny z celem, dla którego dane osobowe zostały pierwotnie zebrane.

CZĘŚĆ DRUGA

Dane syntetyczne

Zamiast nieskutecznie anonimizować wrażliwe dane rzeczywiste – stwórzmy ich matematyczne repliki

Dane syntetyczne

Według Europejskiego Inspektora Ochrony Danych (EIOD) "***Dane syntetyczne to sztuczne dane**, które są generowane na podstawie oryginalnych danych i modelu, który jest szkolony w celu odtworzenia cech i struktury oryginalnych danych (...). Proces generowania, zwany również syntezą, może być wykonywany przy użyciu różnych technik, takich jak drzewa decyzyjne lub algorytmy głębokiego uczenia. Dane syntetyczne można sklasyfikować ze względu na rodzaj danych oryginalnych:*

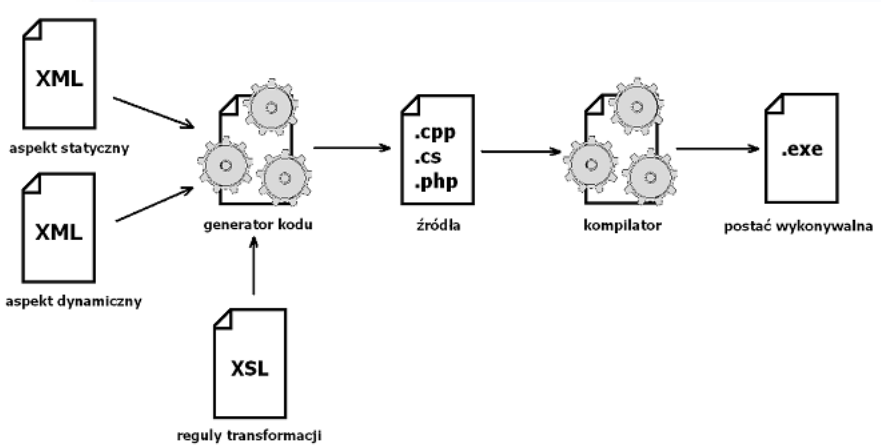
- *pierwszy typ wykorzystuje rzeczywiste zbiory danych*
- *drugi wykorzystuje wiedzę zgromadzoną przez analityków*
- *a trzeci typ jest kombinacją tych dwóch."*

Zasadniczo **dane syntetyczne to dane generowane komputerowo**, które pochodzą z istniejących danych rzeczywistych lub z algorytmów i modeli, które w pełni lub częściowo replikują cechy, wzorce i właściwości danych rzeczywistych.

Praktyczne podejście do danych syntetycznych?

Według EIOD są to dane **wygenerowane sztucznie**, odtwarzające statystyczną strukturę i zależności oryginału.

- **Brak relacji 1:1** - brak bezpośredniego powiązania rekordu syntetycznego z konkretną tożsamością osoby fizycznej.
- **Przydatność dla AI** - zachowują trendy statystyczne, korelacje i użyteczność badawczą rzeczywistej bazy danych.
- **Rekomendacja IOD** - narzędzie eliminujące konieczność masowego przetwarzania rzeczywistych danych osobowych (art. 5 ust. 1 lit. c RODO).



Uwaga:
Proces samego uczenia generatora to faza, która wymaga ścisłego monitorowania przez iOD i podstawy prawnej (art. 6 RODO)

Korzyści biznesowe i prawne danych syntetycznych

- **Zapewnienie prywatności i zgodność z przepisami (ok. 35-40%) -** bezpieczne wykorzystanie danych bez ryzyka wycieków i naruszeń RODO. Umożliwia to zgodne z prawem udostępnianie danych zespołom zewnętrznym.
- **Przyspieszenie prac badawczo-rozwojowych i testów (ok. 25-30%)** - generowanie zbiorów na żądanie skraca czas oczekiwania w procesie testów jakości QA (Quality Assurance) i testów oprogramowania. Eliminuje to biurokratyczne procedury pozyskiwania i maskowania danych produkcyjnych.
- **Uzupełnienie braków danych i skalowalność (ok. 15-20%)** - zwiększenie rozmiaru zestawów treningowych oraz tworzenie scenariuszy dla trudnych lub rzadkich przypadków, takich jak np. nowe formy oszustw w bankowości.
- **Redukcja kosztów infrastruktury (ok. 10-15%)** - mniejsze wydatki na licencjonowanie danych, magazynowanie wrażliwych informacji oraz audyty bezpieczeństwa.

*Statystyczny podział głównych zalet operacyjnych przy wdrożeniu generatorów danych syntetycznych (SDG) wg analityków IT.

Rola IOD w Privacy by Design

Wybór algorytmu

IOD powinien uczestniczyć i wspierać proces oceny przydatności algorytmu syntezy, np. drzewa decyzyjne, pod kątem ochrony danych.

Selekcja wejściowa

Wymuszenie filtrowania quasi-identyfikatorów i unikalnych rekordów przed zasileniem nimi generatora syntezy.

Matematyczny szum

Zapewnienie wdrożenia technologii differential privacy w celu uniemożliwienia odtworzenia tożsamości z modelu generatywnego.

Ulepszenia procesu anonimizacji danych przy użyciu danych syntetycznych

W celu zmaksymalizowania korzyści z generowania danych syntetycznych i innych metod anonimizacji danych, organizacje powinny również wdrożyć dodatkowe strategie.

- **Oceń swoje dane i aplikacje -** oceń typy danych przechowywanych, zbieranych i przetwarzanych w aplikacjach i systemach. Zidentyfikuj zestawy danych i ustal priorytety, które zestawy danych wymagają anonimizacji lub de-identyfikacji.
- **Opracuj politykę zarządzania danymi** - polityka zarządzania danymi powinna być zgodna zarówno z RODO, jak i wewnętrznymi standardami. Regularnie aktualizuj dokumentację zgodnie z pojawiającymi się trendami co pozwoli na minimalizację ryzyka naruszenia danych.
- **Utrzymuj środowisko nieprodukcyjne** - przygotuj oddzielne, bezpieczne środowisko do tworzenia, utrzymywania i kontrolowania zanonimizowanych danych testowych. Utrzymanie tego środowiska oddzielnie od systemów produkcyjnych zapobiega przypadkowym wyciekom danych i zapewnia bezpieczną przestrzeń do testowania.
- **Ciągła kontrola danych syntetycznych** - stosuj dedykowane protokoły testowe, aby zapewnić zgodność syntetycznych danych z prawem, przy jednoczesnym zachowaniu właściwości statystycznych oryginalnego zestawu danych (może będzie trzeba użyć technik zwiększających prywatność).
- **Organizuj szkolenia personelu** - zainwestuj w szkolenia, aby nauczyć swój zespół najlepszych praktyk anonimizacji danych i danych syntetycznych. Zadbaj o to, aby IT rozumiało kluczowe wymogi regulacyjne i podstawy bezpiecznego przetwarzania danych.

Ograniczenia technik anonimizacji danych

Każda stosowana technika wiąże się z własnymi wyzwaniami i ograniczeniami, które należy zrozumieć, aby osiągnąć zgodność:

- **Degradacja jakości danych -** anonimizacja może usunąć ważne elementy danych, korelacje i atrybuty. Nadmierna anonimizacja danych może pozbawić istotnych szczegółów potrzebnych do celu analizy. Badania medyczne, analizy finansowe i szkolenia w zakresie uczenia maszynowego generują w tym obszarze wysokie ryzyko, np. anonimizacja transakcji finansowych może usunąć kluczowy kontekst, taki jak dokładne lokalizacje lub znaczniki czasu.
- **Wymagania dotyczące zasobów i złożoność** - wdrożenie anonimizacji danych wymaga zasobów obliczeniowych i wiedzy technicznej w zespole IT. Należy starannie wybrać odpowiednie techniki: maskowanie danych, pseudonimizację, generowanie syntetycznych danych dla konkretnych przypadków użycia i typów danych. Każda metoda wiąże się z własnym zestawem wymagań technicznych i rozważań.
- **Konsekwencje kosztowe** - anonimizacja może prowadzić do długoterminowych oszczędności, jednak początkowa konfiguracja i bieżąca konserwacja mogą być kosztowne. Trzeba będzie zainwestować w infrastrukturę, oprogramowanie i szkolenie pracowników oraz będzie trzeba regularnie uaktualniać algorytmy, aby sprostać zmieniającym się zagrożeniom i wymogom regulacyjnym.
- **Ryzyko ponownej identyfikacji** - większość metod anonimizacji danych niesie ze sobą ryzyko potencjalnej ponownej identyfikacji. Zaawansowane techniki lub dodatkowe źródła danych mogą umożliwić atakującym powiązanie zanonimizowanych informacji z osobami, np. zanonimizowane dokumentacje medyczne mogą być skorelowane z publicznymi danych demograficznych co może potencjalnie ujawnić tożsamość pacjentów.
- **Problemy ze skalowalnością** - utrzymanie skutecznej anonimizacji w dużych, dynamicznych zestawach danych jest trudne. Wraz ze wzrostem i zmianą wolumenu danych wzrasta złożoność anonimizacji, np. anonimizacja strumieni danych z urządzeń IoT w czasie rzeczywistym wymaga solidnych i skalowalnych rozwiązań, aby zapewnić ciągłą ochronę prywatności.

Ryzyko re-identyfikacji

Ostrzeżenie ICO

Dane syntetyczne **nie stają się automatycznie danymi anonimowymi**.

Im bardziej dane syntetyczne przypominają rzeczywiste, tym wyższe prawdopodobieństwo wystąpienia tzw. **przeuczenia (overfittingu)** generatora.

Zagrożenie polega na tym, że generator może zapamiętać i wiernie odtworzyć skrajne wartości odstające (unikalne rekordy), co prowadzi do ujawnienia tożsamości konkretnych osób.

Przykłady wykorzystania danych anonimowych

Obszar	Opis	Przykłady
Zdrowie	Anonimizacja danych pacjentów pozwala dostawcom usług opieki zdrowotnej i badaczom badać trendy zdrowotne i wyniki leczenia bez ujawniania tożsamości pacjentów. Wspiera badania medyczne i zdrowie publiczne, jednocześnie spełniając standardy prywatności.	**Badania medyczne** - szpitale i kliniki anonimizują dane pacjentów onkologicznych w celu testowania różnych protokołów leczenia. **Badania kliniczne** - firmy farmaceutyczne usuwają dane identyfikacyjne, aby zapewnić zgodność z przepisami podczas testowania bezpieczeństwa i skuteczności nowych leków.
Finanse	Banki i instytucje finansowe stosują anonimizację w celu ochrony poufnych informacji i podejmowania decyzji na podstawie danych, jednocześnie dbając o prywatność klientów.	**Wykrywanie oszustw** - instytucje finansowe anonimizują i badają dane dotyczące transakcji w celu identyfikacji i analizy wzorców oszustw. **Zarządzanie ryzykiem -** banki i firmy ubezpieczeniowe wymieniają się anonimowymi danymi, aby ocenić ryzyko kredytowe i opracować modele udzielania pożyczek oraz zawierania ubezpieczeń.
Telekomunikacja	Firmy telekomunikacyjne anonimizują dane klientów w celu optymalizacji wydajności sieci, opracowywania strategii marketingowych i analizowania wzorców użytkowania.	**Optymalizacja sieci** - dostawcy usług telekomunikacyjnych anonimizują dane dotyczące użytkowania, aby identyfikować luki w zasięgu i optymalizować wydajność sieci. **Analiza klienta** - anonimizowanie rejestrów połączeń i wykorzystania danych pozwala firmom telekomunikacyjnym dowiedzieć się więcej o zachowaniach i preferencjach klientów, nie naruszając przy tym przepisów o ochronie prywatności.

Ryzyka użycia danych syntetycznych

- Dane syntetyczne niekoniecznie muszą odpowiadać danym anonimowym - ryzyko **ponownej identyfikacji** pozostanie
- Dane syntetyczne replikują dane ze świata rzeczywistego - zwiększenie ryzyka ponownej identyfikacji, nie **zniknie ryzyko wnioskowania i połączenia danych z konkretną osobą**

  Uwaga ICO. *"Należy skupić się na stopniu, w jakim ludzie są identyfikowalni w danych syntetycznych oraz jakie informacje o nich zostaną ujawnione, jeśli identyfikacja się powiedzie. Wykazano, że niektóre metody generowania danych syntetycznych są podatne na ataki polegające na **odwróceniu modelu**, **atakach wnioskowania** o członkostwie i ryzyku **ujawnienia atrybutów**. Mogą one zwiększać ryzyko wywnioskowania tożsamości danej osoby.... „*
- **Stosowanie innych PET**, takich jak prywatność różnicowa, lub pomijanie wartości odstających (quasi identyfikatorów) może służyć **zmniejszeniu ryzyka** ponownej identyfikacji danych osobowych, ale nie może go całkowicie wyeliminować
- **Faza generowania danych syntetycznych** może wiązać się z przetwarzaniem danych osobowych, w szczególności przy gromadzeniu i analizie rzeczywistych zbiorów danych, co wiąże się z koniecznością przestrzegania RODO i związanych z nim obowiązków.
- Udzielenie **pełnej informacji zgodnie z art. 13 RODO** osobom, których dane dotyczą, których dane są gromadzone, a następnie wykorzystywane do celów szkoleniowych w zakresie sztucznej inteligencji, a także **określenie podstawy prawnej przetwarzania zgodnie z art. 6 RODO**

CZĘŚĆ TRZECIA

Warsztat rozliczalności

Dowód techniczny i dokumentacja przed Prezesem UODO, czyli jak udowodnić nieodwracalność anonimizacji z pomocą metryk PETs i testów

Weryfikacja technologii PETs

Technologia PET	Kluczowy mechanizm	Podatność / Ryzyko prawne	Użyteczność w AI
k-Anonimowość	Maskowanie atrybutów w grupach o liczności minimum k.	Atak homogeniczności i wiedzy pomocniczej.	Niska (niewystarczająca dla LLM).
l-Różnorodność	Zapewnienie minimum l różnych cech wrażliwych w grupie.	Podatność na ataki semantyczne i asymetrię rozkładu.	Średnia (wymaga t-bliskości).
Prywatność różnicowa	Dodawanie szumu matematycznego do wag modelu.	Koszt użyteczności danych (utrata precyzji).	**Wysoka (doby standard dla LLM)**.

Technologie zwiększające prywatność (PET) - szeroki wachlarz **technologii, rozwiązań sprzętowych lub programowych**, których celem jest pełne wykorzystanie danych bez **narażania ich prywatności i bezpieczeństwa.**
https://www.enisa.europa.eu/publications/pets-maturity-tool

Poprzez minimalizację wykorzystania danych osobowych, maksymalizację bezpieczeństwa danych i wzmocnienie praw osób PET chroni prywatność osób w usługach lub aplikacjach dostępnych w Internecie bez utraty ich funkcjonalności.

Prywatność różnicowa (DP - Differential Privacy)

$$P[M(D) \in S] \leq e^{\varepsilon} \cdot P[M(D') \in S] + \delta$$

*Matematyczne równanie (ε, δ)-Differential Privacy określające rozkład prawdopodobieństwa.

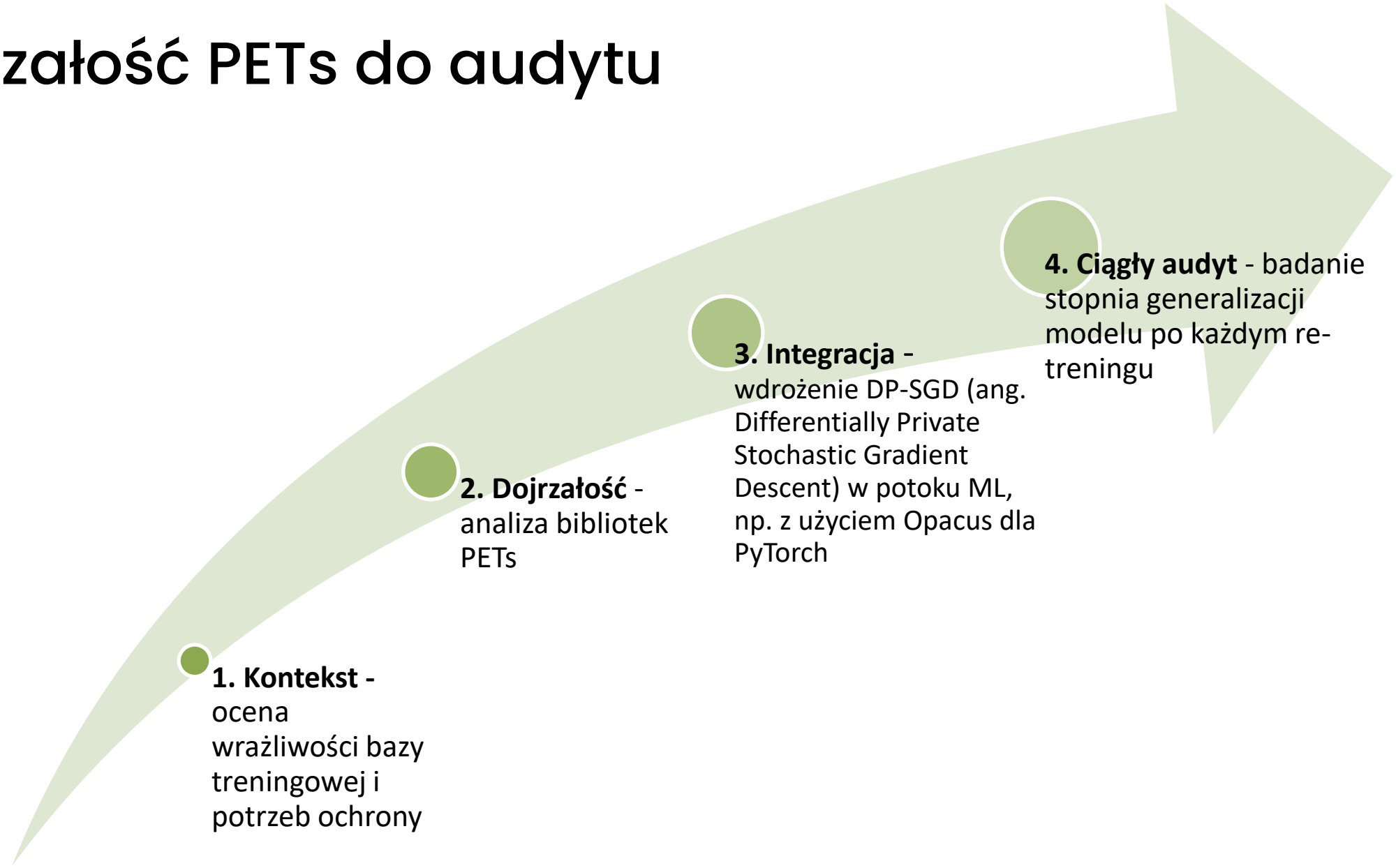
Standard matematyczny

Gwarantuje, że dołączenie lub usunięcie pojedynczego rekordu (osoby) z bazy treningowej **nie wpływa znacząco** na zachowanie modelu.

- **Budżet epsilon (ε) zwany budżetem prywatności** - im mniejszy ε, tym silniejsza ochrona prywatności, ale wyższa utrata precyzji.

- **Parametr delta (δ)** - prawdopodobieństwo całkowitego wycieku (musi dążyć do zera).

Dojrzałość PETs do audytu



Techniki i rodzaje anonimizacji danych

Metoda	Oryginalne dane	Po zmianie
Maskowanie danych	Jan Kimble	Jan Kowalski lub Klient943
Pseudonimizacja danych	1234-5678-9876-5432	1111-2222-3333-4444
Uogólnienie danych	Wiek: 27	Wiek: 25-30
Zaburzenie danych - dodaje losowy szum do danych	Wynagrodzenie: $ 50,000	Wynagrodzenie: $ 49,550
Podmiana danych	01/15/1985	03/22/1990
Dane syntetyczne	$123.45	$126.78

Klient	PESEL	Data ur.	Email	Miasto	Kod poczt.	Nr karty kred.	Typ	Wartość
Jan Nowak	69072206253	22-07-1969	jan.nowak@email.com	Warszawa	PL00-193	6287 8107 3365 9842	Indywidualny	100 344
Dane w spoczynku			**Tokenizacja**					
Hdu Jnkow	67122835031	04-12-1972	0m6.h4jk7@lhu1d.xyk	Kjauiwek	PL60-034	3490 3343 3884 3902	Indywidualny	97 234
Dane w użyciu – Instytucja finansowa			**Detokenizacja**					
Jan Nowak	69072206253	22-07-1969	jan.nowak@email.com	Warszawa	PL00-193	6287 8107 3365 9842	Indywidualny	100 344
Dane w użyciu – Dział obsługi klienta			**Maskowanie**					
Jan Nowak	*******	6253	22-07-	****	0m6.h4jk7@lhu1d.xyk	Warszawa	PL00-193	**** **** **** 9842	Indywidualny	*** ***
Dane w użyciu – Uczenia maszynowe			**Anonimizacja**					
Hdu Jnkow	67122835031	50 do 59 lat	0m6.h4jk7@lhu1d.xyk	******	PL00-	***	3490 3343 3884 3902	Indywidualny	>80 000

Zastosowanie PET w zależności od celu

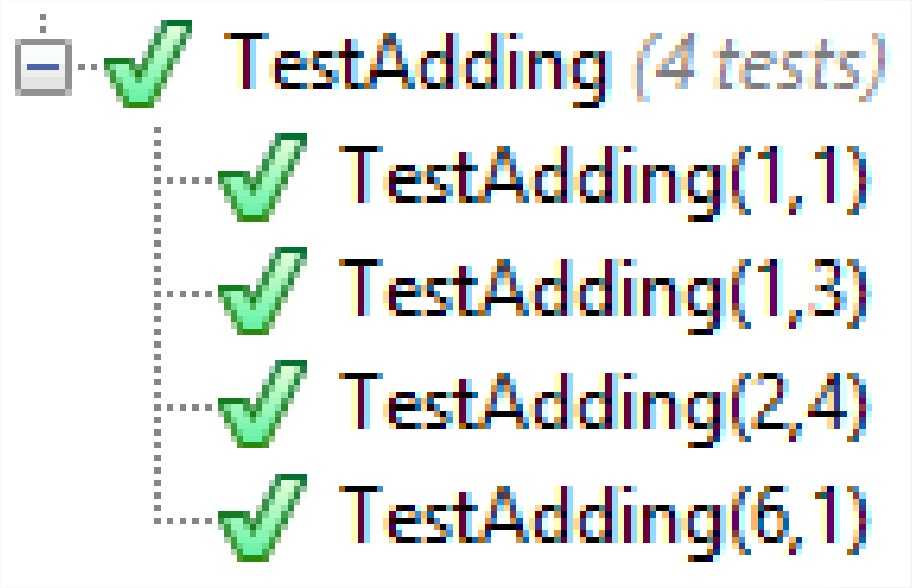
Jak udokumentować proces anonimizacji z perspektywy RODO?

Udokumentuj:

- **art. 5**
- **art. 24**
- **art. 25**
- **art. 30**
- **art. 35**

Dokumentacja - Zasada rozliczalności (art. 5 ust. 2 RODO)

1. Informacje dotyczące **oceny skutków dla ochrony danych**, w tym wszelkie oceny i decyzje, w których stwierdzono, że ocena skutków dla ochrony danych nie jest konieczna.
2. Wszelkie konsultacje, porady lub informacje zwrotne udzielone przez **Inspektora Ochrony Danych.**
3. Informacje na temat **środków technicznych i organizacyjnych** zastosowanych podczas projektowania modelu sztucznej inteligencji w celu zmniejszenia prawdopodobieństwa identyfikacji.
4. Środki techniczne i organizacyjne podjęte na **wszystkich etapach cyklu życia modelu**, które przyczyniły się do braku danych osobowych w modelu lub zweryfikowały ten brak.
5. Dokumentację demonstrującą teoretyczną **odporność modelu sztucznej inteligencji na techniki ponownej identyfikacji**, a także mechanizmy kontroli mające na celu ograniczenie lub ocenę powodzenia i wpływu ataków, np. wnioskowania, w tym m.in. wskaźniki prawdopodobieństwa ponownej identyfikacji w oparciu o aktualny stan wiedzy, sprawozdania na temat tego, **w jaki sposób model został przetestowany (przez kogo, kiedy, jak i w jakim zakresie) oraz wyniki testów**.
6. Dokumentację stosowaną przez wdrażających model, w szczególności dokumentację dotyczącą środków podjętych w celu **zmniejszenia prawdopodobieństwa identyfikacji** oraz dotyczącą ewentualnego ryzyka szczątkowego.



CZĘŚĆ CZWARTA

Testy re-identyfikacji

Jakie testy penetracyjne i ataki symulacyjne należy przeprowadzić, by zgromadzić dowody na przeprowadzenie odpowiednich testów

Kryteria oceny skuteczności anonimizacji danych

K1

- Czy nadal możliwe jest **wyodrębnienie** konkretnej osoby fizycznej?

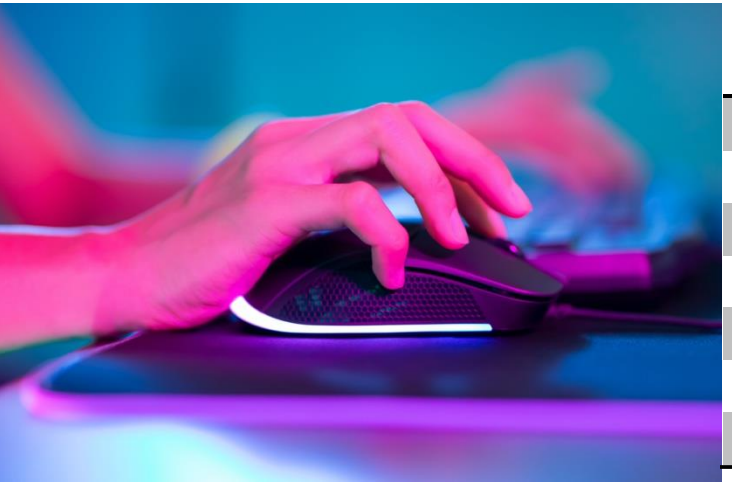
K2

- Czy nadal możliwe jest **powiązanie** zapisów dotyczących konkretnej osoby fizycznej?

K3

- Czy można **wywnioskować** informacje w odniesieniu do konkretnej osoby fizycznej?

Jaką technikę anonimizacji danych wybrać?



	Czy nadal istnieje ryzyko wyodrębnienia?	Czy nadal istnieje ryzyko możliwości tworzenia powiązań?	Czy nadal istnieje ryzyko wnioskowania?
Pseudonimizacja	Tak	Tak	Tak
Dodawanie zakłóceń	Tak	Być może nie	Być może nie
Zastąpienie	Tak	Tak	Być może nie
Agregacja lub k-anonimizacja	Nie	Tak	Tak
L-dywersyfikacja	Nie	Tak	Być może nie
Prywatność różnicowa	Być może nie	Być może nie	Być może nie
Skracanie/Tokenizacja	Tak	Tak	Być może nie

Testy wyodrębnienia

Badanie izolacji rekordu

Cel testu:

Sprawdzenie, czy po procesie syntezy lub treningu możliwe jest zidentyfikowanie i wyodrębnienie unikalnej osoby fizycznej.

- **Metoda ataku** - wykorzystanie unikalnych korelacji cech, np. rzadka kombinacja kodu pocztowego, wieku i specyficznego zawodu.

- **Kryterium sukcesu** - prawdopodobieństwo poprawnego przypisania pojedynczego, rzadkiego profilu w zbiorze syntetycznym musi być statystycznie zbliżone do czystego losowania.

Przykład konfiguracji testowej IT:

Dział IT uruchamia skrypt analizujący odległość statystyczną między rzeczywistym zbiorem uczącym z zbiorem syntetycznym. Jeżeli odległość najmniej licznych profili dąży do zera, test wyodrębnienia wykazuje podatność modelu – wymagane jest wzmocnienie k-anonimowości i uogólnienie danych przed syntezą.

Testy powiązania i ataki MIA

Membership Inference Attacks - cyberprzestępca próbuje dowiedzieć się, czy konkretne dane, np. Twoja historia medyczna, zdjęcie zostały użyte do wytrenowania danego modelu sztucznej inteligencji.

Cel testu:

Potwierdzenie odporności modelu na ataki, które sprawdzają, czy dane konkretnego pacjenta/klienta znajdowały się w zbiorze uczącym.

- **Metoda MIA** - atakujący porównuje pewność odpowiedzi modelu dla danych znanych i nieznanych.

- **Kryterium sukcesu** - model zachowuje identyczną pewność odpowiedzi i brak wahań dystrybucyjnych bez względu na obecność rekordu w bazie uczącej.

Cecha	Testy powiązania (linkage)	Ataki MIA
Cel	Odkrycie tożsamości osoby w zanonimizowanej bazie danych.	Ustalenie, czy dany rekord brał udział w uczeniu algorytmu AI.
Źródło informacji	Wiele różnych zbiorów danych (zwykle tabelarycznych).	Interakcja z gotowym, wytrenowanym modelem czarnej skrzynki, np. LLM, sieć neuronowa.
Metoda	Porównywanie wzorców i nakładanie baz danych na siebie.	Analiza pewności siebie modelu, np. funkcji loss, prawdopodobieństwa wyjścia.

Symulacja powiązania (linkability):

Audyt polegający na próbie połączenia zbioru wyjściowego – syntetycznego lub wag modelu z zewnętrzną publiczną bazą danych, np. rejestrem KRS.
Jeżeli ponowne powiązanie danych jest niemożliwe bez dostępu do bazy kluczy pseudonimizacji, to proces spełnia wymagania nieodwracalności.

Testy wywnioskowania - rekonstrukcji danych

Ataki inwersyjne

Cel testu:

Weryfikacja, czy z wyjścia modelu, np. zapytań promptowych można zrekonstruować cechy twarzy, historię medyczną lub dane osobowe.

- **Metoda ataku** - rekurencyjne odpytywanie modelu i analiza statystyczna prawdopodobieństwa generowania określonych tokenów słownych.
- **Kryterium sukcesu** - całkowite zablokowanie możliwości wydobycia pierwotnej tożsamości poprzez filtry wyjściowe i generalizację.

Środki łagodzące IT:

W razie niepowodzenia testów rekonstrukcji, dział IT musi:
1. Wdrożyć **gradient clipping** (obcinanie gradientu) podczas uczenia sieci neuronowej.
2. Wprowadzić **output filters** filtrujące w czasie rzeczywistym wszelkie próby generowania wzorców przypominających m.in. numery PESEL, e-maile czy nazwiska.

Przykładowa ocena IOD

Ocena IOD 12.06.2026 r.						
Pytania weryfikacyjne opracowane na podstawie Opinii EROD 28/2024 oraz powiązanych materiałów dostarczanych przez ENISA i UODO. Pytania te pomogą ocenić IOD stan wdrożenia systemów AI w organizacji z perspektywy przepisów o ochronie danych osobowych oraz rozliczyć się z podejmowanych działań podczas potencjalnej kontroli UODO w zakresie wdrożenia systemów AI w tym konsultacji IOD w projekcie.						
Lp.	**Obszar**	**Pytania**	**Odpowied**	**Szczegół**	**Ocena IOD**	**Uwagi**	**Wskazówki**
1.1	**Podstawa prawna i cel przetwarzania (Zasada zgodności z prawem)**	1.Czy zidentyfikowano oddzielne podstawy prawne dla fazy rozwoju (treningu) i fazy wdrażania modelu AI?					Wyjaśnienie: Przetwarzanie na tych etapach ma różne cele i ryzyka, dlatego wymaga odrębnej analizy.
1.2		2.Jeśli podstawą jest prawnie uzasadniony interes (art. 6 ust. 1 lit. f RODO), czy przeprowadzono test równowagi (PUI)?					Czy wykazano niezbędność przetwarzania do realizacji celu? Czy interesy administratora nie są nadrzędne wobec praw i wolności osób?
1.3		3. Czy cel wdrożenia AI jest zgodny z prawem i nie narusza zakazów zawartych w AI Act?					Przykład: Systemy do profilowania małoletnich w celach reklamowych lub rozpoznawania emocji w edukacji mogą być niedopuszczalne.
2.1	**Szczególne kategorie danych**	4. Czy system AI przetwarza dane wrażliwe (art. 9 RODO)?					Jeśli tak, czy zidentyfikowano jedną z przesłanek z art. 9 ust. 2 RODO (np. wyraźna zgoda lub ważny interes publiczny)?
2.2		5. Czy wdrożono mechanizmy filtrowania danych szczególnych, których gromadzenie nie jest zamierzone?					Zalecenie: EROD przypomina o zakazie przetwarzania SKD, chyba że zachodzi konkretny wyjątek; incydentalne dane powinny być niezwłocznie usuwane.
3.1	**Zasady przetwarzania: dokładność i minimalizacja**	6.Jakie środki podjęto, aby zapewnić dokładność danych generowanych przez AI (uniknięcie halucynacji)?					Czy użytkownik został poinformowany, że odpowiada za weryfikację danych wyjściowych?
3.2		7.Czy rozważono zastosowanie danych syntetycznych zamiast rzeczywistych danych osobowych w fazie testów lub szkolenia?					Dane syntetyczne mogą wzmacniać prywatność i zmniejszać ryzyko naruszeń.

Przykładowa ocena IOD

4.1	**Realizacja praw osób, których dane dotyczą**	8.Czy opracowano procedurę obsługi żądań sprostowania danych w modelu AI?					Wyzwanie: W modelach LLM izolacja i zmiana konkretnych danych osobowych „zaszytych” w algorytmie jest technicznie trudna.
4.2		9.Czy administrator jest w stanie spełnić prawo do usunięcia danych (prawo do bycia zapomnianym) ze zbiorów treningowych?					
5.1	**Zarządzanie ryzykiem i Transparentność (ochrona danych w fazie projektowania i domyslna ochrona danych)**	10.Czy przeprowadzono ocenę skutków dla ochrony danych (DPIA)?					Czy analiza uwzględnia ryzyka specyficzne dla AI, takie jak uprzedzenia (bias), dyskryminacja czy "dark patterns"?
5.2		11.Czy wdrożono zasady przejrzystości (art. 13/14 RODO)?					Czy osoby wiedzą, że wchodzą w interakcję z systemem AI?
5.3		12.Czy istnieją jasne zasady dotyczące udostępniania danych firmowych do zewnętrznych modeli AI?					Należy ustalić, czy dane wprowadzane do promptów zasilają model dostawcy (ryzyko utraty kontroli).
6.1	**Umowa z dostawcą narzędzia AI**	13. Czy dostawca AI, np. Microsoft, gwarantują wypełnianie zasad chrony danych osobowych i dostarczają dokumentacje to potwierdzającą?					
7.1	**Bezpieczeństwo techniczne**	14. Czy zastosowano technologie wzmacniające prywatność (PETs)?					
7.2		15. Czy system zapewnia poufność, integralność i dostępność danych zgodnie z normami, np. ISO 27001, ISO 42001?					

Rozliczalność

Zasada rozliczalności - weryfikacja IOD 12.06.2026 r.

Według Opinii EROD 28/2024 (wydanej w grudniu 2024 r.), UODO podczas kontroli systemów AI nie będzie skupiał się tylko na samym algorytmie, ale przede wszystkim na całym cyklu życia danych – od momentu zbierania zbiorów treningowych, aż po wyniki generowane przez model.
Poniżej przedstawiam kluczowe obszary, które, zgodnie z opinią EROD, będą przedmiotem weryfikacji przez UODO.

Lp.	Obszar	Pytania	Odpowied:	Zgodno	Uwa	Wskazówki
1.1	**Legalność faz życia AI** (UODO będzie sprawdzać, czy administrator posiada odrębną podstawę prawną dla dwóch różnych procesów)	Faza rozwoju (treningu): Skąd pochodzą dane?				Ważne: EROD wskazuje, że jeśli faza treningu była nielegalna, może to wpłynąc na legalność całego modelu w fazie wdrożenia.
1.2		Faza rozwoju (treningu): Czy pobieranie ich z internetu (web scraping) było legalne?				
1.3		Faza wdrażania (użytkowania): Na jakiej podstawie przetwarzane są dane wprowadzane przez użytkowników (prompty)?				
2.1	**Prawdziwość i rzetelność danych (walka z halucynacjami)** UODO zweryfikuje, jak organizacja radzi sobie z tendencją modeli AI do zmyślonych faktów	Czy system generuje nieprawdziwe dane osobowe na temat konkretnych osób?				
2.2	**Prawdziwość i rzetelność danych (walka z halucynacjami)** UODO zweryfikuje, jak organizacja radzi sobie z tendencją modeli AI do zmyślania faktów	Czy administrator wdrożył mechanizmy pozwalające użytkownikowi na sprostowanie błędnych informacji, co w technologii LLM jest wyzwaniem?				
3.1	**Anonimowość w modelu AI**	Czy z parametrów modelu można pozyskać dane osób ze zbioru treningowego?				To jeden z najtrudniejszych punktów Opinii 28/2024. UODO sprawdzi, czy model AI po wytrenowaniu faktycznie przestał zawierać dane osobowe:
3.2		Jeśli model potrafi wygenerować unikalne dane osoby, na której się uczył, UODO uzna, że model nadal zawiera dane osobowe, a więc podlega pełnemu RODO.				
4.1	**Wykorzystanie prawnie uzasadnionego interesu (art. 6 ust. 1 lit. f)**	Czy interes organizacji, np. stworzenie innowacyjnego narzędzia, nie przeważa nad prywatnością osób, których dane np. zeskrapowano z sieci?				Jeśli organizacja szkoli model na własną rękę, UODO zażąda tzw. testu prawnie uzasadnionego interesu (PIU).
4.2		Czy osoby te mogły racjonalnie oczekiwać, że ich dane zostaną użyte do szkolenia AI?				

Rozliczalność

5.1	**Transparentność i prawa osób** UODO zweryfikuje, czy Twoja polityka prywatności i interfejs systemu AI jasno informują	Czy użytkownik wie, że rozmawia z maszyną?				
5.2		W jaki sposób osoba może wycofać swoje dane ze zbioru treningowego (prawo do bycia zapomnianym)?				
5.3		Jakie są ryzyka związane z przetwarzaniem danych w promptach?				
6.1	**DPIA (Ocena skutków dla ochrony danych)**	Czy DPIA uwzględnia specyficzne ryzyka AI, np. uprzedzenia, błędy poznawcze, halucynacje?				
7.1	**Instrukcja obsługi**	Czy użytkownicy wiedzę, czego nie wolno przekazywać do systemu AI?				
7.2		Czy użytkownicy wiedzą jak mają postępować podczas incydentu bezpieczeńeństwa oraz błędów w dostarczanych przez systemy AI odpowiedziach?				
8.1	**Analiza techniczna**	Czy przeprowadzono testy na wyciek danych z modelu?				
8.2		Czy przeprowadzono testy na poprawność dostarczanych wyników?				
8.3		Czy przeprowadzono testy na skuteczność stosowanych technik anonimizacji danych osobowych?				
9.1	**Umowa z dostawcą narzędzia AI**	Czy był przeprowadzany audyt podmiotu dostarczającego narzędzie AI jeżeli dochodzi do powierzenia danych osobowych?				
9.2		Czy dostawca AI, np. Microsoft, gwarantuje ochronę danych i dostarcza potwierdzające to dokumentację?				

PODSUMOWANIE:
Checklista rozliczalności IOD

1. Paczka dowodowa

Zgromadzenie dowodów: analiza DPIA uwzględniająca Opinię 28/2024 EROD, wyniki testów wyodrębnienie, powiązania, wywnioskowania oraz atestacja prywatności różnicowej (ε, δ).

2. Metodyka i szablony

Wdrożenie i bieżące aktualizowanie formularza jako standardu zarządzania ryzykiem.

3. Edukacja i współpraca

Budowanie dialogu między IOD, Zespołem Data Science i IT. Tłumaczenie wymogów prawnych RODO na konkretne wskaźniki i parametry matematyczne, np. szumu.

Dziękuję za uwagę

Mariola.Wieckowska@LexDigital.pl

https://www.linkedin.com/in/mariolawieckowska/