NR STANDARDU	TYP I NAZWA STANDARDU	NUMER/DATA
NTL-L-001	**Lista kontrolna - Sztuczna Inteligencja - wstępna weryfikacja**	**2/2024-05-19**

1. Niniejszy standard ma na celu weryfikację, czy korzystanie z danego sytemu AI przez podmiot, który będzie wdrażał biznesowo rozwiązanie, spełniać będzie wymagania wynikające z RODO. Aby w tym zakresie zapewnić zgodność, należy uwzględnić następujące konteksty / etapy:
    a. budowanie modelu (systemu AI), w ramach którego może dochodzić do przetwarzania danych osobowych (np. pozyskiwanie, łączenie danych, wykorzystanie w ramach trenowania algorytmu, testowania i walidacji);
    b. wdrożenie i wykorzystanie modelu AI (do czego mogą być wykorzystywane dane osobowe jako input lub output, np. w ramach generatywnej AI);
    c. dalsze uczenie i poprawianie działania modelu wykorzystywanego w ramach systemu AI (m.in. na bazie informacji uzyskanych na etapie wdrożenia modelu AI).

2. Chociaż dla podmiotu wdrażającego rozwiązania AI (systemy AI) głównym obszarem zainteresowania jest wykorzystanie systemu AI, to jednak część ryzyk, które musi uwzględnić może wynikać z błędów na etapie budowania modelu w ramach systemu AI (np. ryzyka związane z błędami na etapie definiowania zakresu danych treningowych).

3. **Słownik pojęć**

Na podstawie:

- https://www.europarl.europa.eu/doceo/document/TA-9-2024-03-13_PL.html#sdocta2 oraz https://eur-lex.europa.eu/legal-content/PL/TXT/HTML/?uri=CELEX:52021PC0206 (Wniosek ROZPORZĄDZENIE PARLAMENTU EUROPEJSKIEGO I RADY USTANAWIAJĄCE ZHARMONIZOWANE PRZEPISY DOTYCZĄCE SZTUCZNEJ INTELIGENCJI (AKT W SPRAWIE SZTUCZNEJ INTELIGENCJI) I ZMIENIAJĄCE NIEKTÓRE AKTY USTAWODAWCZE UNII
- Rozporządzenie Parlamentu Europejskiego i Rady (UE) 2016/679 https://eur-lex.europa.eu/legal-content/PL/TXT/?uri=celex%3A32016R0679
- ENISA Multilayer Framework for Good Cybersecurity Practices for AI https://www.enisa.europa.eu/publications/multilayer-framework-for-good-cybersecurity-practices-for-ai
- NIST AI 100-1 AI RISK MANAGEMENT FRAMEWORK https://www.nist.gov/itl/ai-risk-management-framework

1) **System AI** - system maszynowy, zaprojektowany do działania z różnym poziomem autonomii, który po wdrożeniu może wykazywać zdolność adaptacji i który – do wyraźnych lub dorozumianych celów – wnioskuje, jak na podstawie danych

wejściowych generować wyniki, takie jak predykcje, treści, zalecenia lub decyzje, które mogą wpływać na środowisko fizyczne lub wirtualne.

2) **Dostawca** - osoba fizyczna lub prawna, organ publiczny, agencja lub inny podmiot, który opracowuje system AI lub model AI ogólnego przeznaczenia lub zleca jego opracowanie i który wprowadza go do obrotu lub oddaje system AI do użytku pod własną nazwą handlową lub własnym znakiem towarowym, odpłatnie lub nieodpłatnie.
3) **Podmiot stosujący AI** - osoba fizyczna lub prawna, organ publiczny, agencja lub inny podmiot, która korzysta z systemu AI i sprawuje nad nim kontrolę, z wyjątkiem sytuacji, gdy system AI jest wykorzystywany w ramach osobistej działalności pozazawodowej.
4) **Dane treningowe** - dane wykorzystywane do trenowania systemu AI poprzez dopasowanie jego parametrów podlegających uczeniu.
5) **Dane walidacyjne** - dane służące do oceny trenowanego systemu AI oraz do dostosowywania jego parametrów niepodlegających uczeniu oraz procesu uczenia, między innymi w celu zapobiegania niedostatecznemu wytrenowaniu lub przetrenowaniu.
6) **Dane testowe** - dane wykorzystywane do przeprowadzenia niezależnej oceny systemu AI w celu potwierdzenia oczekiwanej skuteczności działania tego systemu przed wprowadzeniem go do obrotu lub oddaniem go do użytku.
7) **Dane wejściowe** - dane dostarczone do systemu AI lub bezpośrednio przez niego pozyskiwane, na podstawie których system ten generuje wynik działania.
8) **Szczególne kategorie danych osobowych** - kategorie danych osobowych, o których mowa w art. 9 ust. 1 rozporządzenia (UE) 2016/679, art. 10 dyrektywy (UE) 2016/680 i art. 10 ust. 1 rozporządzenia (UE) 2018/1725.
9) **Dane osobowe** - dane osobowe zdefiniowane w art. 4 pkt 1 rozporządzenia (UE) 2016/679.
10) **Dane nieosobowe** - dane inne niż dane osobowe zdefiniowane w art. 4 pkt 1 rozporządzenia (UE) 2016/679.
11) **Dane syntetyczne** - dane wygenerowane sztucznie, nierzeczywiste.
12) **Ryzyko** - połączenie prawdopodobieństwa wystąpienia szkody oraz stopnia jej powagi.
13) **Instrukcja obsługi** - informacje podane przez dostawcę w celu poinformowania operatora sztucznej inteligencji w szczególności o przeznaczeniu i właściwym użytkowaniu systemu sztucznej inteligencji, a także informacje o wszelkich środkach ostrożności, jakie należy podjąć, w tym informacje o szczególnym kontekście geograficznym, behawioralnym lub funkcjonalnym, w którym ma być wykorzystywany system sztucznej inteligencji wysokiego ryzyka.
14) **System rozpoznawania emocji** - system sztucznej inteligencji służący do rozpoznawania lub odgadywania emocji, myśli, stanów umysłu lub zamiarów poszczególnych osób lub grup osób na podstawie ich danych biometrycznych i danych opartych na biometrii.

4. Lista kontrolna

1) Czy zamierzone działanie systemu AI wymaga użycia danych osobowych (np. system rozpoznawania twarzy) na jakimkolwiek etapie korzystania z systemu (input, output, etc.)?
2) Czy zamierzone działanie systemu AI wymaga użycia szczególnych kategorii danych osobowych na jakimkolwiek etapie korzystania z systemu?
3) Czy system AI dopuszcza (nawet jeżeli nie jest to zamierzone działanie) użycie danych osobowych (np. system rozpoznawania twarzy) na jakimkolwiek etapie korzystania z systemu (input, output, etc.)?
4) Czy system AI dopuszcza (nawet jeżeli nie jest to zamierzone działanie) użycie szczególnych kategorii danych osobowych (np. system rozpoznawania twarzy) na jakimkolwiek etapie korzystania z systemu (input, output, etc.)?
5) Czy system AI ma wbudowane zabezpieczenie zapobiegające użyciu danych osobowych w ramach danych typu "input" lub "output"?
6) Czy budując model AI, w tym w ramach trenowania algorytmu, wykorzystano dane osobowe? Jeżeli tak, to czy spełniono wymagania wynikające z RODO m.in. w zakresie podstawy przetwarzania danych (legalność np. w kontekście zmiany celu przetwarzania) i przejrzystości (obowiązki informacyjne)?
7) Czy w ramach budowania modelu lub wdrożenia systemu AI wykorzystano jedną z technologii PET[1] (np. trenowanie algorytmu w oparciu o dane syntetyczne lub tzw. federated learning)?
8) Czy system daje możliwość poprawiania działania algorytmu (np. fine-tuning, RAG, etc.)?
9) Czy dostawca dostarcza wraz z usługą dokumentację techniczną, w tym instrukcję obsługi?
10) Czy dostawca na bieżącą monitoruje zgodność działania systemu AI z przekazaną dokumentacją?
11) Czy dostawca przeprowadził / przeprowadza ocenę ryzyka, tak aby uwzględnić ryzyka dla praw lub wolności osób fizycznych na etapie budowania modelu w ramach systemu AI (w szczególności wykorzystania danych treningowych)?
12) Czy przeprowadzona ocena ryzyka uwzględnia możliwość wystąpienia uprzedzeń (bias), które mogą prowadzić np. do dyskryminacji?
13) Czy w ramach struktury modelu AI (trening, walidacja, testowanie) wykorzystano odpowiednio opracowane dane, zastosowane do celu budowy modelu tak, aby zapobiec m.in. dyskryminacji?
14) Czy wdrożono zabezpieczenia, aby zapobiec etykietowaniu danych, które będzie prowadzić do negatywnych skutków dla osób fizycznych?
15) Czy definiując zbiory danych szkoleniowych, walidacyjnych i testowych zastosowano odpowiednie cechy, właściwości lub funkcje (w tym odpowiednio zróżnicowane dane), tak aby zapewnić prawidłowe działanie systemu AI?

[1] PET - Privacy-Enhancing Technologies; więcej informacji pod adresem: https://arxiv.org/html/2404.03509v1

16) Czy w ramach budowania modelu AI istniała ścisła kontrola nad dostępem do danych, w tym danych treningowych?
17) Czy w ramach budowania modelu AI zdefiniowano uprawnienia dostępowe według zasady "need to know", opierając się na precyzyjnym podziale ról i odpowiedzialności?
18) Czy w ramach budowania modelu AI uwzględniono ryzyko, że systemy AI mogą generować nieuczciwe wyniki dla poszczególnych osób, spowodowane przez niewystarczające zróżnicowanie danych szkoleniowych? Czy wprowadzono odpowiednie zabezpieczenia, aby temu zapobiec?
19) Czy zidentyfikowano wszystkie aktywa wspierające przetwarzanie informacji przez system AI?
20) Czy zapewniono odpowiedni poziom bezpieczeństwa wszelkich aktywów wspierających dotyczących systemu AI, w tym oprogramowania i sieci neuronowych?
21) Czy wdrożono zabezpieczenia mające na celu minimalizację ryzyka zbyt szerokiego zakresu danych wykorzystywanych do budowania modelu AI?
22) Czy dostawca zapewnia, aby system sztucznej inteligencji projektowano i opracowywano w taki sposób, aby zawierał on funkcję umożliwiającą automatyczne rejestrowanie zdarzeń („rejestry zdarzeń") podczas działania systemu?
23) Czy algorytm rozwijany jest / adaptuje się w sposób przewidywalny? Jeśli nie, to:
    - jakie ryzyka może rodzić nieprzewidywalny rozwój systemu AI?
    - jakie środki zaplanowano, aby próbować przewidywać / planować rozwój systemu AI?
24) Czy wdrożono adekwatne zabezpieczenia przed sytuacją, w której algorytm zwraca niedokładne lub nieuczciwe wynik dotyczące osób fizycznych?

    Scenariusz, w ramach którego algorytm uczący zwraca zbyt dużą uwagę na specyficzne cechy zbiorów danych uczących. W rezultacie osoby, które nie są podobne do osób w zbiorach danych szkoleniowych, uzyskują niedokładne i nieuczciwe wyniki.
25) Czy na bieżąco testowane jest bezpieczeństwo oprogramowania służącego do budowania i wykorzystywania algorytmu?
26) Czy / jakie dane podmiotu stosującego system AI są wykorzystywane do poprawy działania algorytmu (w tym dane treningowe, prompty, odpowiedzi zwracane przez system AI)? Czy istnieje ryzyko dostępu do tych danych przez inny podłączony system IT?
27) Czy dostawca dysponuje odpowiednio wykwalifikowanym personelem, który zapewnia rozwój systemu AI, monitorowanie jego wykorzystania w sposób etyczny i zgodny z wymaganiami prawnymi?
28) Czy dostawca waliduje i testuje system AI pod kątem zapewnienia prawidłowości działania w ramach zdefiniowanych parametrów przed udostępnieniem go na rynku?

29) Czy dostawca zapewnia, aby system sztucznej inteligencji projektowano i opracowywano w taki sposób aby osoby fizyczne mogły sprawować nad nim skuteczny nadzór, współmiernie do związanego z tym systemem ryzyka?
30) Czy dostawca gwarantuje wsparcie – na każde żądanie – co do wyjaśnienia działania samego algorytmu i tego w jaki sposób i dlaczego został wygenerowany ten a nie inny wynik (została podjęta ta a nie inna decyzja)? (dotyczy to także innych praw np. żądania kopii danych)?
31) Czy dostawca gwarantuje wsparcie – na każde żądanie – w przypadku potrzeby wykazania zgodności działania systemu z regulacjami prawnymi (w tym RODO)?
32) Czy dostawca wdrożył odpowiednie środki techniczne i organizacyjne, aby informować / ostrzegać podmioty stosujące AI w przypadku wystąpienia ryzyka naruszenia zasad ochrony danych, w tym nieuprawnioną / niezamierzoną re-identyfikację (np. w ramach generowanego głosu, generowanych treści)?

33) Czy system AI zapewnia następujące zasady wyrażone przez **ENISA** oraz **NIST**:
    a) **Odpowiedzialność**. Zapewnienie odpowiedzialności za system AI, w tym wyjaśnienia i uzasadnienia. Ludzie i organizacje powinni być w stanie odpowiedzieć i ponosić odpowiedzialność za wyniki działania systemów AI, szczególnie za negatywne skutki wynikające z ryzyka.
    b) **Dokładność**. Poprawność wyników w porównaniu z rzeczywistością. Procesy zarządzania ryzykiem powinny uwzględniać potencjalne ryzyka, które mogą wynikać z nieważności / błędnego wnioskowania systemu AI o związku przyczynowo-skutkowym.
    c) **Możliwość wyjaśnienia**. Dostarczanie opisu wniosku / decyzji w sposób zrozumiały dla człowieka. Ryzyka związane ze zrozumiałością mogą wynikać z wielu przyczyn, w tym na przykład z braku wierności lub spójności w metodologiach wyjaśniania, lub jeśli ludzie błędnie wnioskują o działaniu modelu, albo też model nie działa zgodnie z oczekiwaniami.
    d) **Uczciwość**. Neutralność dowodów, niezależność od osobistych preferencji, emocji, czy innych ograniczeń wprowadzonych przez kontekst, równość (płci i szans). Uczciwość to pojęcie odrębne, ale powiązane z uprzedzeniami. Zgodnie z ISO/IEC TR 24027:2021, uprzedzenia mogą wpływać na uczciwość. Uprzedzenia mogą być społeczne lub statystyczne, mogą być odzwierciedlane lub powstawać w różnych komponentach systemu i mogą być wprowadzane lub rozpowszechniane na różnych etapach cyklu życia rozwoju i wdrażania systemu AI.
    e) **Prywatność**. Bezpieczne zarządzanie (proces, analiza, przechowywanie, transfer, komunikacja) danymi osobowymi i modelami szkoleniowymi oraz zdolność do działania bez ujawniania informacji (danych, modelu). Identyfikacja wpływu ryzyk związanych z problemami związanymi z prywatnością jest kontekstowa i różni się w zależności od kultury i indywidualnych osób.

f) **Niezawodność**. Zdolność do utrzymania minimalnego poziomu wydajności i konsekwentnego generowania tych samych wyników w granicach dopuszczalnych błędów statystycznych. Może to dać wgląd w ryzyko związane z dekontekstualizacją.

g) **Odporność**. Zdolność do zminimalizowania wpływu, przywrócenia bezpiecznych warunków pracy i wyjścia z ataku przeciwnika.

h) **Solidność**. Zdolność systemu AI do utrzymania wcześniej uzgodnionego minimalnego poziomu wydajności w każdych okolicznościach.Przyczynia się to do analizy wrażliwości w procesie oceny ryzyka.

i) **Bezpieczeństwo (safety)**. Zapobieganie niezamierzonemu lub szkodliwemu zachowaniu systemu AI dla ludzi lub społeczeństwa. Bezpieczeństwo jest silnie skorelowane z ryzykiem.

j) **Bezpieczeństwo (security)**. Zdolność do zapobiegania odchyleniom od bezpiecznych warunków działania w przypadku wystąpienia niepożądanych zdarzeń. Zdolność do odpierania ataków. Zapewnia poufność, integralność, autentyczność, niezaprzeczalność, dostępność danych, procesów, usług i modeli.

k) **Przejrzystość**. Zdolność do wspierania ogólnego zrozumienia systemów AI, uświadamiania zainteresowanym stronom ich interakcji z systemami AI i umożliwienia osobom "dotkniętym" przez te system zrozumienia wyniku. Umożliwia również osobom, na które system AI ma negatywny wpływ, zakwestionowanie jego wyników w oparciu o proste i zrozumiałe informacje na temat czynników i logiki, która posłużyła za podstawę prognozy, zalecenia lub decyzji.

34) Czy przy ocenie ryzyka uwzględniono typowe ryzyko dotyczące cyberbezpieczeństwa w ramach algorytmu w tym systemie poznawczym (uwzględniając OWASP Top Ten):
https://owasp.org/www-project-machine-learning-security-top-10/

a) ML01:2023 Atak polegający na manipulacji danymi wejściowymi.

b) ML02:2023 Atak polegający na zatruwaniu danych.

c) ML03:2023 Atak polegający na odwróceniu modelu.

d) ML04:2023 Atak polegający na wnioskowaniu o członkostwie.

e) ML05:2023 Kradzież modelu.

f) ML06:2023 Ataki na łańcuch dostaw AI.

g) ML07:2023 Atak na naukę transferu.

h) ML08:2023 Pochylenie modelu.

i) ML09:2023 Atak na integralność danych wyjściowych.

j) ML10:2023 Zatrucie modelowe.

Metryka dokumentu

NUMER/DATA	OPIS ZMIAN / AUTORZY
1/2024-04-12	Pierwsze wydanie standardu. M.Gumularz, T.Izydorczyk, B.Śliwiński
2/2024-05-12	Drugie wydanie. Dodatno nowe pytania do listy. M.Gumular