FUNDACJA PANOPTYKON

# ALGORYTMY UCZENIA MASZYNOWEGO W KOMUNIKACJI Z KONSUMENTAMI

## JAK I PO CO O NICH MÓWIĆ? WNIOSKI Z BADANIA

## WPROWADZENIE

Systemy wykorzystujące zaawansowane algorytmy i **uczenie maszynowe** (zwane potocznie „sztuczną inteligencją", a w żargonie branżowym po prostu **ML**)[1] są już w powszechnym użyciu. Mają z nimi kontakt nie tylko obsługujący je profesjonaliści, ale także – choć często nieświadomie – odbiorcy usług internetowych, użytkownicy tzw. inteligentnych urządzeń i obywatele w relacji z organami władzy.

Świadomość tego, czym są i jak działają algorytmy uczenia maszynowego, jest wśród nieprofesjonalistów bardzo niska. Konsumenci oczekują potocznie rozumianej sztucznej inteligencji w produktach typu *smart* (nawet jeśli to tylko proste urządzenie elektroniczne z kilkoma czujnikami) i w „inteligentnych" asystentach (nawet jeśli po drugiej stronie znajduje się jedynie chatbot posługujący się kilkoma skryptami). Nie spodziewają się, że naprawdę wyrafinowana technologia, nierzadko wykorzystująca ich dane osobowe, kryje się za spersonalizowanymi rekomendacjami, wynikami wyszukiwania na stronie czy reklamą behawioralną. Na skutek tego w zetknięciu z rozwiązaniami uczenia maszynowego konsumenci pozostają nieświadomi i pasywni – nie są w stanie zgłaszać problemów ani przekazywać konstruktywnej informacji zwrotnej.

Rzetelną rozmowę o celach, którym służą takie rozwiązania w biznesie, utrudnia fakt, że pojęcie sztucznej inteligencji zostało popkulturowo silnie skojarzone z działającym autonomicznie, groźnym dla człowieka systemem. Dodatkowa trudność bierze się z obiektywnego skomplikowania tej dziedziny i wysokiego progu wejścia do merytorycznej rozmowy. Poza profesjonalnym żargonem brakuje słów do opisu tego, jak działają algorytmy uczenia maszynowego. Nie istnieją wspólne wyobrażenia, do których mogą się odwołać firmy, by wytłumaczyć, jak działa ich technologia (nade wszystko: jakie cele realizuje i jakie ryzyka rzeczywiście może stwarzać). W efekcie, kiedy media donoszą o aferach związanych z działaniem sztucznej inteligencji, konsumenci reagują nieracjonalnie, nierzadko panicznie.

Pomieszanie pojęć w debacie publicznej oraz brak wspólnego języka do opisu zjawisk, z którymi coraz powszechniej spotykamy się w usługach, nie służy komunikacji na linii firma–konsument. Bez zdziwienia obserwujemy, że zdecydowana większość przedsiębiorstw korzystających z rozwiązań uczenia maszynowego nie komunikuje tego faktu klientom. Jesteśmy jednak przekonani, że w interesie i konsumentów, i samych firm leży zmiana tego stanu rzeczy.

---

[1] ML, machine learning (z ang. uczenie maszynowe) to podkategoria sztucznej inteligencji, w której maszyna – bez konieczności dokładnego jej programowania – uczy się poprzez identyfikowanie wzorców lub trendów w danych wejściowych oraz dostosowywanie się do nich. Uczenie maszynowe polega na automatycznym przeprogramowywaniu się systemu na podstawie danych.

W tym briefie wyjaśniamy podstawy naszego przekonania. Odwołujemy się do **prac regulacyjnych** (prowadzonych pod egidą Unii Europejskiej)[2] oraz własnych doświadczeń i **wniosków z dialogu,** w który zaangażowało się sześć firm oferujących usługi bazujące na rozwiązaniach uczenia maszynowego (szczegóły i podziękowania w aneksie **„O badaniu"**).

## SPIS RZECZY

W pierwszej części briefu prezentujemy wnioski z dialogu z firmami i nasz pomysł na warstwowe podejście do przejrzystości, w którym każda warstwa komunikacji (tj. każdy komunikacyjny „produkt") realizuje nieco inny cel i odpowiada na inną potrzebę odbiorców. Uwzględniamy tutaj oczekiwania konsumentów, niezależnych ekspertów i organizacji społecznych (takich jak Fundacja Panoptykon) oraz uzasadnione – z uwagi na względy bezpieczeństwa lub ochronę własności intelektualnej – obawy biznesu dotyczące granic przejrzystości.

W dalszej części briefu podsumowujemy wynik rozmów z przedstawicielami firm oraz analizy ich oficjalnych stanowisk i istniejących praktyk, a mianowicie:

- jakie są przekonania biznesu na temat tego, czy i ew. kiedy warto komunikować rozwiązania uczenia maszynowego;
- jakie dobre praktyki komunikacyjne już się ukształtowały i czemu one służą;
- co stanowi najważniejszy czynnik utrudniający komunikację pomiędzy firmą a konsumentem.

W ostatniej części briefu proponujemy dalsze kroki, to jest:

- głębsze zbadanie oczekiwań konsumentów,
- wypracowanie dobrych praktyk komunikacyjnych w duchu unijnego podejścia do regulacji sztucznej inteligencji[3].

W aneksie opisujemy sam proces badawczy:

- profil firm, z którymi rozmawialiśmy;
- konkretne wątki, które się pojawiły w naszym dialogu;
- materiały informacyjne i przykłady algorytmów ML, na których pracowaliśmy.

[2] 21 kwietnia 2021 r. Komisja Europejska przedstawiła kompleksową propozycję uregulowania sztucznej inteligencji (m.in. w usługach konsumenckich): Artificial Intelligence Act, https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/excellence-trust-artificial-intelligence_pl. W dalszej części briefu posługujemy się określeniem „proponowana regulacja AI". Więcej o założeniach reformy piszemy na stronie Fundacji Panoptykon: https://panoptykon.org/wiadomosc/unia-szykuje-przepisy-dotyczace-ai-5-problemow.

[3] Artykuł 69 proponowanej regulacji AI zachęca do tworzenia kodeksów dobrych praktyk również twórców systemów, które nie zostały zakwalifikowane jako systemy wysokiego ryzyka: „The Commission and the Member States shall encourage and facilitate the drawing up of codes of conduct intended to foster the voluntary application to AI systems other than high-risk AI systems of the requirements set out in Title III, Chapter 2 on the basis of technical specifications and solutions that are appropriate means of ensuring compliance with such requirements in light of the intended purpose of the system", https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence-artificial-intelligence.

# I. POSTULATY PANOPTYKONU

## Co zrozumieliśmy w dialogu z firmami?

Do dialogu z firmami, które świadczą usługi konsumenckie wykorzystujące ML, przystąpiliśmy z mniej lub bardziej uświadomionymi hipotezami. Ponieważ przejrzystość jest dla nas wartością samą w sobie, wyszliśmy z założenia, że będzie ona równie cenna dla odbiorców usług, w których wykorzystywane są rozwiązania ML. Przyjęliśmy, że wśród konsumentów są ludzie, którzy po prostu chcą wiedzieć, jak to wszystko działa. W naszym świecie obowiązuje też zasada ograniczonego zaufania („nie przyjmuj deklaracji, których nie możesz samodzielnie zweryfikować, najlepiej w oparciu o dane"). Jednocześnie byliśmy i jesteśmy przekonani, że o większości rozwiązań uczenia maszynowego można opowiedzieć przystępnym językiem. Nawet bardziej wyrafinowane systemy uczące się bazują na szeregu decyzji, które – w procesie projektowania i testowania – podejmują ludzie[4].

W związku z tymi założeniami oczekiwaliśmy od firm gotowości do rozmowy o szczegółach nieledwie technicznych, takich jak:

- klasa i generacja wykorzystywanego algorytmu;
- funkcja strat lub inne parametry pokazujące, które błędy popełniane przez algorytm są tolerowane, a które minimalizowane;
- przyjęte metryki sprawiedliwości;
- struktura danych, na których algorytm był trenowany.

W trakcie dialogu zrozumieliśmy, że ujawnianie tak szczegółowych informacji przez firmy w istocie może zagrozić ich uzasadnionym interesom (np. zostać wykorzystane przez konkurencję albo nieuczciwych partnerów biznesowych, tzw. złych aktorów, rozgrywających system na własną korzyść). Jednocześnie komunikaty o charakterze specjalistycznym – z perspektywy odbiorcy, który nie tworzy systemów ML i nie dysponuje wiedzą ekspercką – nie mają dużej wartości. W tym kontekście **błędem byłoby narzucanie konsumentom** (np. w formie wyskakujących pop-upów) **jakichkolwiek technicznych wyjaśnień**. Takie działanie mogłoby zakłócić proces *user experience*, zniechęcić użytkownika do korzystania z usługi wykorzystującej rozwiązanie uczenia maszynowego lub go po prostu przestraszyć.

Cenniejsza – również dla nas, zaawansowanych użytkowników – jest **otwarta rozmowa o celach i założeniach systemu opartego na uczeniu maszynowym**, a w szczególności rzetelne przedstawienie problemu, który firma przy pomocy tej technologii próbuje rozwiązać.

Taki dialog nie wymaga wnikania w szczegóły techniczne ani ujawniania informacji z perspektywy firmy wrażliwych. Jako niezależni eksperci jesteśmy w stanie wyrobić sobie pogląd na temat celów systemu, wbudowanych weń wartości i założeń, już na podstawie rzetelnie opowiedzianej logiki optymalizacji (bez dostępu do technicznych parametrów). Wystarczy informacja o tym:

[4] Szczegółowo wyjaśniamy to przekonanie w tekście pt. „Black-Boxed Politics: Opacity is a Choice in AI Systems", https://medium.com/@szymielewicz/black-boxed-politics-cebcod5a54ad.

- na jaki efekt obliczone jest działanie algorytmów ML (cel optymalizacji opowiedziany w języku biznesowym, nie technicznym);
- po czym firma poznaje, że wykorzystywany przez nią system działa dobrze (podstawowe metryki);
- jakiego typu błędy stara się wyeliminować, a jakie akceptuje – i dlaczego (wartości wpisane w system)[5].

Rozmową na powyższe tematy będą zainteresowani przede wszystkim zaawansowani użytkownicy systemu, nierzadko eksperci zawodowo zajmujący się problematyką ML. Chętnie sięgną oni do **głębszych** (choć nadal publicznie dostępnych) **opracowań na temat założeń systemu**. Nie ma powodu, by tego typu informacje prezentować w marketingowym, mocno uproszczonym języku, celując w możliwości i oczekiwania przeciętnego konsumenta. Wystarczy, że będzie to język zrozumiały dla osób spoza branży ML.

Z perspektywy zwykłego odbiorcy ważne jest, by na możliwie wczesnym etapie jego komunikacji z firmą oferującą usługi wykorzystujące ML pojawił się **komunikat: „wchodzisz w kontakt z systemem napędzanym »sztuczną inteligencją«"**. Taka informacja, podana zrozumiałym językiem i w przystępnej formie, może pełnić ważną funkcję edukacyjną, a nawet ostrzegawczą.

Podobny pogląd prezentuje Komisja Europejska: zgodnie z proponowaną regulacją AI systemy sztucznej inteligencji, które z założenia mają wchodzić w interakcję z człowiekiem, powinny być projektowane tak, by osoba fizyczna wiedziała, że ma do czynienia ze „sztuczną inteligencją"[6]. W tym miejscu, naszym zdaniem, otwiera się ciekawe **pole do eksperymentów z etykietowaniem usług wykorzystujących ML** (z zakresu *user experience*, ale też poszukiwania odpowiedniego języka do komunikacji z konsumentami)[7].

Cel ostrzegawczy ma znaczenie w przypadku usług, w których interes konsumenta nie (zawsze) jest tożsamy z interesem firmy wykorzystującej ML. Z taką rozbieżnością interesów możemy mieć do czynienia między innymi:

- w sektorze usług finansowych (jeśli bank optymalizuje działanie algorytmów ML tak, by minimalizować swoje ryzyko finansowe, a nie tak, by zminimalizować ryzyko dla klienta – np. nadmiernego zadłużenia lub missellingu[8]);
- w sektorze e-commerce (jeśli platforma zakupowa dopuszcza reklamy czy oferty targetowane w oparciu o dane wrażliwe, np. zdrowie, które mogą wykorzystywać słabości konsumentów);

[5] W części „Skowronki lepszych praktyk" cytujemy publicznie dostępne, choć zaawansowane i przeznaczone raczej dla ekspertów, materiały firm Netflix i YouTube, które odpowiadają na te właśnie pytania.

[6] Artykuł 52 ust. 1 proponowanej regulacji AI: „Providers shall ensure that AI systems intended to interact with natural persons are designed and developed in such a way that natural persons are informed that they are interacting with an AI system, unless it is obvious from the circumstances and the context of use".

[7] O tym wyzwaniu szerzej piszemy w części „Dalsze kroki: co proponujemy?".

[8] Misseling to nieuczciwa sprzedaż. Na zjawisko składają się zachowania, procedury sprzedaży i działania marketingowe, mające na celu wprowadzenie konsumentów w błąd.

- w serwisach oferujących treści wideo i na platformach społecznościowych (jeśli algorytm jest zoptymalizowany na efekt ciągłego scrollowania lub angażowania użytkowników za wszelką cenę, co może przyczyniać się do powstania lub zwiększenia problemów psychicznych).

W sytuacji rozbieżności interesów ryzyko, że działanie algorytmu bazującego na uczeniu maszynowym wywoła negatywny skutek dla konsumenta, wydaje się realne. Z tego względu organizacje społeczne, takie jak Fundacja Panoptykon, oczekują, że dla algorytmów ML związanych z wysokim ryzykiem będzie prowadzona niezależna ocena skutków, uwzględniająca perspektywę praw człowieka (**Human Rights Impact Assessment**)[9].

Firmy, z którymi rozmawialiśmy, deklarowały, że cele, na jakie są zoptymalizowane ich algorytmy ML, nie stoją w sprzeczności z interesami odbiorców, a ewentualne ryzyko dla praw (i innych uzasadnionych interesów) konsumentów wydaje się raczej niskie. Nie polemizując z tymi zapewnieniami[10], stoimy na stanowisku, **że przeprowadzona z perspektywy konsumenta ocena ryzyka jest dobrą praktyką** zarówno w przypadku algorytmów niskiego ryzyka, jak i w przypadku usług, w które nie jest wpisana sprzeczność interesów. Nawet jeśli algorytm ML został zoptymalizowany z myślą o obopólnych korzyściach, w praktyce mogą się pojawić efekty niezamierzone i zwyczajne błędy. **Publicznie dostępna ocena ryzyka (lub odpowiednio zredagowany wyciąg z takiego dokumentu)** stanowi materiał, w którym firma może wykazać, jak przeciwdziała pojawieniu się tych właśnie niezamierzonych efektów i błędów – a tym samym jest doskonałą metodą na budowanie zaufania do wartości i usług danej marki.

Dla pełnej jasności: nie oczekujemy, że osoby postronne będą dopuszczone do danych i parametrów technicznych, które potwierdzą tezy publicznie głoszone przez firmy. Żeby faktycznie budować zaufanie do usług opartych o ML, ktoś jednak musi być w stanie te tezy i zapewnienia zweryfikować (i nie może to być audytor wybrany i opłacony przez samą firmę). W duchu proponowanej regulacji AI tę funkcję mogą pełnić akredytowane organizacje (podmioty komercyjne lub niekomercyjne, również niezależne instytuty badawcze) z odpowiednią wiedzą ekspercką, które będą zobowiązane do zachowania poufności[11]. Wreszcie: nad całym sektorem usług wykorzystujących ML powinien czuwać niezależny organ, który w razie sygnału od zaniepokojonych konsumentów lub organizacji społecznych będzie w stanie zapukać i zajrzeć na zaplecze.

Dialog z firmami utwierdził nas w przekonaniu, że **informacja zwrotna od konsumenta lub testera** (w tej roli może wystąpić organizacja społeczna) do firmy wdrażającej lub rozwijającej system wykorzystujący ML może być bardzo cenna. Może pomóc w wychwyceniu błędów i niepożądanych efektów, a także dostarczyć wiarygodne – bo pochodzące bezpośrednio od konsumenta – dane wejściowe dla algorytmów ML, np. o zainteresowaniach lub oczekiwaniach konkretnej osoby. Jednak aby informacja zwrotna mogła spełnić swoją funkcję (np. przełożyć się na większą trafność sugerowanych podpowiedzi), musi być precyzyjna, podana we właściwym momencie i poważnie potraktowana po drugiej stronie (tj. nie tylko odebrana, ale też przeanalizowana i uwzględniona

[9] W tym kierunku zmierzają prace regulacyjne prowadzone na poziomie Rady Europy, https://www.coe.int/en/web/artificial-intelligence/cahai.

[10] W ramach tego dialogu chcieliśmy jedynie poznać praktyki i przekonania firm. Nie dążyliśmy do weryfikacji żadnych tez.

[11] Por. art. 33 proponowanej regulacji AI.

w procesie ewaluacji rozwiązań uczenia maszynowego). Takiej komunikacji nie udrożni tradycyjny formularz: potrzebujemy nowych, intuicyjnych interfejsów dla konsumentów, którzy chcą (na własnych warunkach, ale w ramach możliwości stworzonych przez usługodawcę) wejść w kontakt ze sztuczną inteligencją i wpływać na kształtowanie własnego doświadczenia.

## Przejrzystość ma warstwy

Jak zatem mógłby wyglądać **złoty standard przejrzystości i wyjaśnialności w świecie konsumenckich usług bazujących na rozwiązaniach uczenia maszynowego**? Standard, który godzi uzasadniony interes firm (ochronę własności intelektualnej i konieczność zabezpieczenia systemu przed nieuczciwymi aktorami) z potrzebą, jaką sami zgłaszamy: budowania zaufania w oparciu o fakty, a nie deklaracje?

Nasz pomysł na „warstwową przejrzystość" usług opartych na rozwiązaniach uczenia maszynowego zarysowujemy poniżej. Bazuje on na przekonaniu, że **poziom szczegółowości tego, co jest ujawniane, musi być uzależniony zarówno od kategorii odbiorcy** (m.in. od tego, czy dysponuje on wiedzą ekspercką oraz czy będzie zobowiązany do zachowania poufności), **jak i od sytuacji, w której ten odbiorca się znajduje** (w szczególności: czy ma podstawy, by sądzić, że system działa na jego niekorzyść, albo inne powody do niepokoju).

Każda z przedstawionych w tabeli warstw przejrzystości realizuje inny cel i odpowiada na inną potrzebę innego odbiorcy.

### ❖ Warstwa 1. Etykieta

Ta warstwa obejmuje informacje, które są ciekawe i zrozumiałe dla przeciętnego odbiorcy. Nie zawiera żadnych szczegółów technicznych. Musi być dostępna w prostym języku. Warto, by była uzupełniona elementami graficznymi.

| **Komunikat:** „Wchodzisz w kontakt z systemem wykorzystującym tzw. sztuczną inteligencję". | **Strona edukacyjna:** „Chcesz wiedzieć więcej?" |
|---|---|
| 1. Zwięzła, widoczna informacja o tym, że:<br>▪ korzystając z tej usługi, konsument wchodzi w kontakt z systemem uczącym się, który będzie m.in. przewidywał jego zachowania, przydzielał go do grupy osób o podobnych zainteresowaniach, dobierał wyświetlane treści pod kątem jego ujawnionych preferencji;<br>▪ system został zaprojektowany do określonego celu lub zadania, jeśli zatem konsument ma inną intencję (stawia przed systemem inne zadanie), system może zadziałać w sposób nieoptymalny.<br>2. Odesłanie do strony „Chcesz wiedzieć więcej?" (patrz: kolumna po prawej stronie).<br>3. Informacja o tym, czy i w jaki sposób można | Narracja wciągająca konsumenta w świat rozwiązań opartych na uczeniu maszynowym, napisana obrazowym i prostym językiem, która wyjaśnia poniższe zagadnienia.<br>▪ Na jaki problem firma próbuje odpowiedzieć, wykorzystując uczenie maszynowe? W jakich usługach je stosuje?<br>▪ Jaką wartość dodaną dla konsumentów wytwarza zastosowanie ML w tych usługach? Po czym przeciętny konsument może poznać, że system działa dobrze? |

skontaktować się z człowiekiem (*human in the loop*).

- Czy z działaniem systemu są związane ryzyka dla konsumentów? W szczególności: jakie grupy ludzi i dlaczego mogą być negatywnie dotknięte działaniem systemu? Co firma robi, żeby temu zapobiegać?
- Dla każdego typu usługi opartej o ML osobno:
  - Jakie są modelowe przypadki użycia ML (warunki, w których system powinien działać poprawnie)?
  - Do jakich celów nie należy tego systemu wykorzystywać (w jakich warunkach nie będzie działał poprawnie)?
- Kto tworzy rozwiązania uczenia maszynowego (jeśli firma korzysta z zewnętrznych dostawców)? Gdzie można się zgłaszać po więcej informacji?
- W przypadku systemów wysokiego ryzyka – link do wpisu dotyczącego konkretnego systemu w dostępnej publicznie unijnej bazie (o której mówi art. 60 proponowanej regulacji AI).

## ❖ Warstwa 2. Dialog z zaniepokojonym lub zaciekawionym konsumentem

Ta warstwa zawiera informacje przeznaczone dla konkretnej osoby, która aktywnie zgłasza zainteresowanie (chce wiedzieć więcej).

Wychodzimy z założenia, że za takim zainteresowaniem nie stoi pusta ciekawość, ale **zaniepokojenie** automatyczną decyzją podjętą przez system („chcę wiedzieć, dlaczego system tak zadziałał") albo **gotowość do przekazania informacji zwrotnej** („chcę, żeby system działał inaczej").

**Spersonalizowane wyjaśnienie:** „Dlaczego system wygenerował taką, a nie inną decyzję w mojej sprawie?"

W ramach spersonalizowanego wyjaśnienia (podawanego konkretnej osobie na jej wyraźne żądanie) konsument powinien móc się dowiedzieć:

- jaka była logika automatycznie podjętej decyzji (np. jaka została zastosowana reguła; w ludzkim – a nie matematycznym – języku);
- jakie dane (nie tylko osobowe) zostały wzięte pod uwagę w tym konkretnym przypadku i skąd pochodziły, a w szczególności: jaka cecha zadecydowała o takim, a nie innym traktowaniu konsumenta (jeśli da się ją wyodrębnić);

Możliwość **przekazania informacji zwrotnej** „Chcę, żeby system traktował mnie inaczej!"

W idealnym scenariuszu konsument, który chce przekazać informację zwrotną, nie musi sam jej konstruować. Jest w tym celu przekierowywany do intuicyjnego formularza lub w pełni **interaktywnego interfejsu**, w którym może:

- skorygować surowe dane wejściowe (np. usunąć pozycję z historii zakupów czy historii oglądania);
- zmodyfikować swój profil (np. przypisać się do innej kategorii wg deklarowanych lub zaobserwowanych zainteresowań);

- do jakiej kategorii konsument został zaklasyfikowany w wyniku automatycznej oceny lub jaka decyzja została wobec niego podjęta;
- czy w proces podejmowania decyzji był albo nadal może być zaangażowany człowiek (jeśli tak – w jakiej roli; z kim w tym celu należy się skontaktować; jakie będą kolejne kroki etc.).

Spersonalizowane wyjaśnienie **nie musi być generowane natychmiastowo** (*on the spot*). Natomiast wydaje się nam ważne to, by konsument mógł zażądać wyjaśnienia w łatwy, intuicyjny sposób – **w momencie, w którym natrafia na problem** (np. poprzez kliknięcie w ikonkę ze znakiem zapytania, która odsyła do strony lub formularza z wyjaśnieniem).

Ten wysoki i niewątpliwie wymagający dodatkowych nakładów (finansowych, organizacyjnych) standard rekomendujemy dla systemów, które podejmują **istotne (w tym prawnie wiążące) decyzje**, dotykające konkretnego człowieka.

- wybrać (inne) kategorie rekomendowanych treści (np. poprzez użycie tagów tematycznych);
- wybrać metodę sortowania rekomendowanych treści (np. chronologicznie albo wg trafności);
- wybrać inny cel działania algorytmu, jeśli system przewiduje taką możliwość (np. „chcę spędzać mniej czasu przed ekranem" albo „chcę mieć dostęp do różnych, losowo dobieranych treści, a nie tylko sprofilowanej oferty").

Ten wymagający opracowania odpowiedniego interfejsu standard zalecamy jako **dobrą praktykę** dla systemów wytwarzających rekomendacje, podpowiedzi i spersonalizowane oferty.

Warto w tym miejscu odnotować, że w proponowanej przez Komisję Europejską regulacji dla platform internetowych (**Digital Services Act**) dominujące platformy (*very large online platforms*) mają mieć prawny obowiązek zagwarantowania swoim użytkownikom opcji **opt out z personalizacji.** Projekt regulacji wspiera także takie rozwiązania, które pozwalają użytkownikom samodzielnie definiować kluczowe parametry dla systemów rekomendacyjnych.

## ❖ Warstwa 3. Szczegóły dla zaawansowanych

Ta warstwa obejmuje informacje techniczne i elementy oceny ryzyka, które są zrozumiałe dla zaawansowanych odbiorców i mogą ich (np. organizacje strażnicze i niezależnych badaczy) zainteresować – ale tylko w takim w zakresie, w jakim ich ujawnienie nie naraża własności intelektualnej ani bezpieczeństwa systemu. Nadal poruszamy się w sferze **publicznie dostępnych** informacji.

Ta warstwa przejrzystości może przybrać różne formy: publicznie dostępnej **analizy**, zawierającej wyjaśnienie społecznie istotnych kwestii (np. wybranych przez twórców systemu definicji i metryk sprawiedliwości)[12], **wyciągu z** przeprowadzonej **oceny ryzyka** (w proponowanej regulacji AI art. 43 – *conformity assessment*) czy napisanej bardziej specjalistycznym językiem **strony internetowej „dla zaawansowanych"**.

Z perspektywy zaawansowanych odbiorców cenne będą informacje o:

- klasie i generacji algorytmu lub algorytmów wykorzystywanych w systemie;

[12] W kulturze anglosaskiej taką funkcję pełni tzw. *paper*, przygotowywany zwykle przez grupę badaczy, prezentowany na konferencjach naukowych i publikowany na odpowiednich portalach. Odnotowaliśmy, że również polscy badacze (m.in. związani z Allegro) publikują takie opracowania w języku angielskim.

- ogólnej logice oraz szczegółowych celach, na jakie system został zoptymalizowany;
- rodzajach danych wejściowych, jakie algorytm wykorzystuje do podejmowania jednostkowych rozstrzygnięć;
- spodziewanych danych wyjściowych (np. typach rozstrzygnięć lub rekomendacji; kategoriach, do których przypisywane są osoby oceniane przez system);
- przyjętych założeniach nt. kategoryzacji osób, które mają związek z celami albo logiką działania systemu (jeśli celem działania algorytmu jest przypisywanie do określonej kategorii albo określone kategorie są wykorzystywane w procesie targetowania lub personalizacji);
- źródłach, z których pochodzą dane wykorzystywane do trenowania algorytmów, oraz kategoriach wykorzystanych danych (bez potrzeby ujawniania jakichkolwiek danych osobowych);
- przyjętych metodach pracy z danymi, które mają zagwarantować ich jakość i reprezentatywność;
- poziomie trafności, odporności (stabilności) i bezpieczeństwa, jakiego oczekują operatorzy systemu w oparciu o przeprowadzone testy;
- kluczowych decyzjach, jakie podjęli projektanci systemu, w tym krokach, jakie mają zagwarantować poprawne działanie systemu (oczekiwany poziom trafności, odporności i bezpieczeństwa) i zminimalizować ryzyko tendencyjnych rozstrzygnięć (*fairness*);
- szacowanym czasie życia systemu i zaplanowanych metodach jego ewaluacji (szczególnie tych z udziałem samych odbiorców);
- środkach organizacyjnych zapewniających nadzór człowieka nad funkcjonowaniem systemu (jeśli zostały wdrożone).

## ❖ Warstwa 4. Wgląd w dane dla uprawnionych organów i akredytowanych audytorów

Ta warstwa **nie jest dostępna publicznie**. Wymaga wiedzy eksperckiej i zachowania poufności. Z perspektywy konsumentów i organizacji strażniczych jest ważne, żeby ktoś kompetentny i niezależny miał do tych informacji dostęp i mógł zweryfikować deklaracje firm. W proponowanej regulacji AI są to *national competent authorities* i *notified bodies*.

Z perspektywy ochrony interesów konsumentów akredytowani audytorzy powinni móc zweryfikować:

- cele optymalizacji wyrażone w języku formuł matematycznych wraz z pełną listą parametrów wykorzystywanych do optymalizacji i przypisanymi im wagami;
- metryki sprawiedliwości wyrażone w języku formuł matematycznych;
- dane potwierdzające to, czy osoby znajdujące się w podobnej sytuacji (lub mające podobny profil) zostały potraktowane identycznie (np. czy w ich przypadku została podjęta taka sama decyzja albo czy zostały zaklasyfikowane do tej samej kategorii);
- parametry techniczne (takie jak funkcja straty) pozwalające odtworzyć i zweryfikować, w jaki sposób system został zoptymalizowany (w szczególności: jakiego typu błędy są tolerowane, a jakie zwalczane);
- wszelkie znane lub możliwe do przewidzenia okoliczności, które mogą mieć wpływ na zakładany poziom trafności, odporności i bezpieczeństwa systemu;
- wszelkie znane lub możliwe do przewidzenia okoliczności, które mogą prowadzić do niezamierzonych rezultatów, takich jak krzywda konkretnego konsumenta albo zwiększenie tendencyjności całego systemu;
- środki techniczne wdrożone w celu minimalizowania ryzyka błędów czy tendencyjności w systemie;
- procedury służące zapewnieniu poprawnego funkcjonowania systemu (np. cykliczne testy i ewaluacje, aktualizacje oprogramowania);
- faktycznie wykorzystane dane treningowe (poprzez wgląd do bazy, również do danych osobowych).

# II. ANALIZA WYWIADÓW, OFICJALNYCH STANOWISK ORAZ ISTNIEJĄCYCH PRAKTYK

## Przekonania firm

Z naszego dialogu z firmami oraz z analizy ich oficjalnych stanowisk wyłaniają się następujące przekonania, które wyznaczają granice gotowości do komunikowania konsumentom rozwiązań uczenia maszynowego[13].

### ❖ Przejrzystość w relacji z konsumentami jest wartością, szczególnie kiedy w grę wchodzą dane osobowe (a więc również RODO).

Przejrzystość, jako **wartość w relacjach z konsumentami i ważny element publicznego wizerunku firmy**, ma ugruntowaną pozycję. Żadna z firm z tym nie dyskutuje. Ważną rolę w promowaniu tej wartości odegrało RODO, które wymusiło transparentność w zakresie tego, jakie dane na temat konsumentów są przetwarzane oraz kiedy i na podstawie jakich informacji podejmowane są, wiążące prawnie lub niosące dla konsumentów istotne skutki, automatyczne decyzje. Ten standard jest przez firmy akceptowany. Inną sprawą jest jego praktyczne wdrożenie w odniesieniu do konkretnych usług (por. kolejne punkty).

Przejrzystość jako **jedno z narzędzi budowania zaufania** może mieć **szczególne znaczenie** w usługach, w których interes konsumenta nie zawsze jest zbieżny z ekonomicznym interesem firmy. Z taką sytuacją możemy mieć do czynienia np. w przypadku ubezpieczeń i innych produktów finansowych. Zdaniem naszych rozmówców z tego sektora ta rozbieżność interesów ma jednak charakter pozorny (krótkoterminowy), bo w dłuższej perspektywie wszystkim opłaca się autentyczne dopasowanie usługi czy produktu do faktycznych możliwości klienta (tj. unikanie missellingu).

Poniższe cytaty ilustrują ten kierunek myślenia:

> *We believe it is important to give our users a helpful sense of what data is used for what purposes and how, and an understanding of how the algorithm works in organising and prioritising content for them. We recognise that users are seeking more transparency and control over their online experience, including the role of algorithms, and we have developed a number of tools for users.*
>
> *We acknowledge the desire of the Commission and others to provide users with transparency over why they are being recommended certain content. We are willing to be a constructive participant in dialogue over practical mechanisms that provide meaningful transparency to users, while avoiding the risks of poorly designed requirements. In particular, we must ensure new transparency requirements do not risk commercially-sensitive information,*

[13] W przeważającej części są to stanowiska wypracowane w związku z proponowaną przez Komisję Europejską regulacją sztucznej inteligencji. Konkretne stanowiska przywołujemy w przypisach do pojawiających się w tej części briefu cytatów.

*violate user privacy or data disclosure laws, nor allow bad actors to game our systems[14].*

Google

*EUTA members would support self-regulatory principles for accountability and transparency to help businesses tread market opportunities and the possibility for unfair bias and discrimination15.*

European Tech Alliance[16], w tym Allegro

*New technology, including AI systems, must be transparent and explainable[17].*

IBM

*Przejrzystość jest potrzebna, by w świecie tajemnicy klienci nie byli jak dzieci. Jest w etycznym biznesie przestrzeń na służenie ludziom i jednocześnie zarabianie pieniędzy – sprzedawanie naprawdę potrzebnych produktów[18].*

ANG

- Firmy są skłonne wyjaśniać konsumentom zasady działania konkretnych funkcjonalności lub usług napędzanych rozwiązaniami uczenia maszynowego.

Podczas analizy istniejących praktyk określonych firm oraz w trakcie rozmów z ich przedstawicielami przekonaliśmy się, że na rynku usług napędzanych rozwiązaniami uczenia maszynowego nie brakuje prób wyjaśnienia (bez wchodzenia w techniczne szczegóły) zasad czy też biznesowej logiki konkretnych funkcjonalności. Najciekawsze przykłady takich wyjaśnień zebraliśmy i omówiliśmy w kolejnym punkcie.

Niemal wszystkie firmy, z którymi rozmawialiśmy lub których materiały analizowaliśmy, podzielają nasze przekonanie na temat tego, że konsument powinien mieć:

1) świadomość kontaktu z systemem bazującym na rozwiązaniach wykorzystujących sztuczną inteligencję;

[14] Stanowisko Google w konsultacjach Digital Services Act: https://blog.google/documents/89/Googles_submission_on_the_Digital_Services_Act_package_1.pdf.

[15] European Tech Alliance High-Level Principles on Artificial Intelligence: http://eutechalliance.eu/wp-content/uploads/2020/02/EUTA-High-Level-Principles-on-AI.pdf.

[16] EUTA zrzesza tylko firmy unijne, które stosują RODO i wiele innych regulacji chroniących konsumenta.

[17] Stanowisko IBM w konsultacjach White Paper on AI: https://www.ibm.com/blogs/policy/wp-content/uploads/2020/06/IBM-Submission-on-the-EC-AI-White-Paper.pdf.

[18] Cytat z wywiadu.

2) podstawowe informacje o tym, co ten system robi i na jaki efekt został zoptymalizowany (m.in. po to, by odbiorca mógł dostosować własne oczekiwania)19;

3) możliwość przekazania informacji zwrotnej (za pośrednictwem łatwego w obsłudze formularza albo interaktywnego interfejsu, np. suwaków), jeśli system robi coś zaskakującego lub niepożądanego.

To przekonanie, poparte przykładami wartościowych komunikatów skierowanych do konsumentów, wybrzmiało również w materiałach, które przekazał nam Facebook[20].

> *As far as our existing efforts on AI explainability and transparency that are geared toward different audiences, you will see that FB has invested in both in-product and out of product features that all seek to address how our AI systems work. Some highlights include:*
>
> - *How the Ranking Algorithm Works;*
> - *Community Standards Enforcement, which leverages AI in finding and removing violating content;*
> - *Most Recent Changes to Controlling Your Feed Features.*

## ❖ Firmy nie są skłonne ujawniać technicznych szczegółów konsumentom.

Firmy żywią przekonanie, że konsumenci nie oczekują przejrzystości w zakresie stosowanych rozwiązań technologicznych i nie chcą zaglądać „pod maskę". Twierdzą, że w dialogu pomiędzy firmą a konsumentem wartość ma uczciwa rozmowa o celach i założeniach biznesowych, a nie wnikanie w szczegóły techniczne. Zdaniem biznesu konsumenci chcą po prostu, by napędzany przez rozwiązania uczenia maszynowego system robił to, czego się od niego oczekuje – w sposób szybki i intuicyjny, bez zaskoczeń i widocznych błędów. W kontakcie z systemem dla konsumentów ważne są: **przewidywalność** i **poczucie bezpieczeństwa**.

Z doświadczenia firm wynika, że tak długo, jak długo ten cel jest realizowany, konsumenci nie zadają pytań o działanie silnika. Większość naszych rozmówców wyrażała wręcz obawę, że skonfrontowanie odbiorcy z aspektami technicznymi (a nawet tylko z pojęciami uczenia maszynowego czy sztucznej inteligencji) mogłoby wzbudzić lęk lub zniecierpliwienie konsumenta (i w efekcie zniechęcić go do korzystania z usługi). Wynika to ze stanu debaty publicznej na temat AI, zmitologizowania tej technologii oraz bardzo wysokiej bariery wejścia do rozmowy o jakichkolwiek szczegółach.

Jednocześnie firmy są przekonane, że zbyt daleko idąca przejrzystość rozwiązań uczenia maszynowego mogłaby posłużyć tzw. złym aktorom do rozgrywania systemów na własną korzyść (np. sprzedawcy, reklamodawcy lub profesjonalni wydawcy mogliby dopasowywać swoje działania pod algorytm, żeby ich treści były lepiej pozycjonowane; przestępcy mogliby obchodzić rozwiązania

---

[19] Wymownie ujęli to w rozmowie przedstawiciele firmy Netflix: **„Is what is happening to me what I think is happening to me?"**.

[20] Ponieważ nie udało nam się umówić na wywiad, nie analizowaliśmy głębiej praktyk komunikacyjnych tej firmy.

antyfraudowe, służące bezpieczeństwu). Z tego powodu większość naszych rozmówców **obawiałaby się**:

- ujawnienia konkretnych parametrów wykorzystywanych w procesie optymalizacji;
- wskazania wszystkich źródeł i rodzajów danych wykorzystywanych do trenowania algorytmu;
- ujawnienia przyjętej przez twórców systemu funkcji straty.

Sposobem na pogodzenie potrzeby budowania zaufania wśród konsumentów z potrzebą ochrony własności intelektualnej (lub ochrony przed złymi aktorami, o których mowa wyżej) może być – zdaniem niektórych firm – wdrażanie **procedur i mechanizmów zapobiegających nadużyciom**, które jednak nie wiążą się z wyższym standardem przejrzystości (np. regularne audyty, nadzór regulatora rynku) oraz publikowanie odpowiednio zredagowanych **wyciągów z ocen ryzyka** (w proponowanej regulacji AI art. 43 – *conformity assessment*).

Poniższe cytaty ilustrują ten kierunek myślenia:

> *As a tool, transparency has potential for both positive and negative impact. It can empower users but amongst those there will also be bad actors who will use information on algorithmic transparency to manipulate the algorithms*[21].
>
> Facebook

> *Fortunately, just as not everyone needs to be an expert mechanic to get a driving licence and trust that a car is safe to drive, nor are explanations always necessary when using AI systems. In considering the level of explainability demanded in a specific instance, it is worth comparing the standards applied to current (non-AI) approaches. For example, an oncologist may struggle to explain the intuition that leads them to believe they fear a patient's cancer has recurred. In contrast, an AI system in the same circumstance may be able to provide biomarker levels and historical scans from 100 similar patients as a reference, even if it remains a struggle to fully grasp how the data are processed to predict an 80% chance of cancer. There is a risk that innovative uses of AI could be inadvertently precluded by demanding that AI systems meet a „gold standard" of explainability that far exceeds that required of established non-AI (including human-based) approaches. A sensible compromise is needed that balances the benefits of using complex AI systems against the practical constraints that different standards of explainability would impose*[22].
>
> Google

> *The most appropriate technical and organisational measures for mitigating AI risks from the outset will be situation-dependent. For low-risk applications, EU*

[21] Stanowisko Facebooka w konsultacjach Digital Services Act: https://about.fb.com/de/wp-content/uploads/sites/10/2020/09/FINAL-FB-Response-to-DSA-Consultations.pdf.

[22] Stanowisko Google w konsultacjach White Paper on AI: https://www.blog.google/documents/77/Googles_submission_to_EC_AI_consultation_1.pdf.

*businesses should have the flexibility to choose measures that will deliver the best outcomes.*

*Transparency could be fed into DPIAs, a key compliance tool for certain forms of AI processing under GDPR[23].*

European Tech Alliance, w tym Allegro

Dobrą praktyką, stanowiącą ukłon w stronę zaawansowanych użytkowników (w tym ekspertów z organizacji społecznych takich jak Panoptykon), jest publikowanie w zwięzłej formie, np. fiszek, zestawienia najważniejszych informacji o modelu uczenia maszynowego.

*Machine learning (ML) model transparency is important across a wide variety of domains that impact peoples' lives, from healthcare to personal finance to employment. The information needed by downstream users will vary, as will the details that developers need in order to decide whether or not a model is appropriate for their use case. This desire for transparency led us to develop a new tool for model transparency, Model Cards[24], which provide a structured framework for reporting on ML model provenance, usage, and ethics-informed evaluation and give a detailed overview of a model's suggested uses and limitations that can benefit developers, regulators, and downstream users alike[25].*

Google

*IBM has proposed the use of FactSheets as a general approach to AI transparency. A FactSheet is a collection of relevant information about an AI model or service that is created during the machine learning lifecycle[26].*

*The goal of the FactSheet project is to foster trust in AI by increasing an increased understanding of how AI was created and deployed and enabling the ability to control how AI is created and deployed. (...) This allows a consumer of the model to determine if it is appropriate for their situation. (...) This can prevent undesirable situations, such as a model training with unapproved datasets, models having biases, or models having unexpected performance variations[27].*

IBM

Wreszcie: większość firm, z którymi rozmawialiśmy, podkreśla swoją gotowość do udostępniania i dyskutowania detali technicznych w formie opracowań skierowanych do ekspertów, publikowanych w kanałach przeznaczonych dla takich użytkowników, jak również pogłębianych na życzenie

[23] http://eutechalliance.eu/wp-content/uploads/2020/02/EUTA-High-Level-Principles-on-AI.pdf

[24] https://arxiv.org/pdf/1810.03993.pdf

[25] https://ai.googleblog.com/2020/07/introducing-model-card-toolkit-for.html

[26] https://www.ibm.com/blogs/policy/wp-content/uploads/2020/06/IBM-Submission-on-the-EC-AI-White-Paper.pdf

[27] https://aifs360.mybluemix.net/introduction

zainteresowanych organizacji społecznych. Przykłady takich opracowań cytujemy i chwalimy w części „Skowronki lepszych praktyk".

## ❖ Firmom brakuje języka do rozmowy o rozwiązaniach uczenia maszynowego z nieekspertami.

Niemal wszystkie firmy, z którymi mieliśmy okazję rozmawiać, potwierdzają gotowość do komunikowania swoich rozwiązań uczenia maszynowego na różnym poziomie, w zależności od potrzeb, interesu prawnego i poziomu zaawansowania odbiorcy. Tę gotowość widać w eksperckich opracowaniach, które regularnie pojawiają się na technologicznych blogach (Allegro[28], Google[29], IBM[30], Netflix[31]) czy w periodykach naukowych. Stale powraca jednak temat bariery, jaką jest brak takiego języka do opisu tych rozwiązań, który byłby zrozumiały dla osób spoza branży *data science*. O tym wyzwaniu więcej piszemy w ostatnim punkcie tej części briefu.

Poniższe cytaty pokazują, jak same firmy definiują problem.

> *Facebook and other AI developers are continually exploring new tools and processes for enhancing the fairness and transparency of our AI systems in line with emerging Responsible AI principles. However, there is still a general lack of consensus on exactly how best to translate these broad responsible AI principles into practice[32].*
>
> Facebook

---

> *AI's greatest value is seeing patterns in complex situations that are beyond human comprehension — thus (by definition) such AI systems will not be fully explainable in a way that a person can grasp. Even if the source code were shared in such a situation (an extreme form of algorithmic transparency which Google does not support) it would not help, as it would still be too complex to fathom even for experts. However, it is a fallacy that AI systems are black boxes. With enough effort and the right tools, it is possible to get some insight into why any AI system behaves in a certain way.*
>
> *The problem is that explainability is costly, either in terms of technical resources or in terms of trade-offs with other goals like model accuracy (if more accurate but harder-to-explain techniques have to be foregone). Tailoring explanations to be meaningful to a range of audiences is also difficult and time intensive. While*

---

[28] https://blog.allegro.tech/

[29] https://research.google/pubs/

[30] https://www.ibm.com/blogs/research/

[31] https://netflixtechblog.com/

[32] Stanowisko Facebooka w konsultacjach White Paper on AI: https://scontent-waw1-1.xx.fbcdn.net/v/t39.8562-6/103231277_1162782850727962_2719421119701851752_n.pdf?_nc_cat=103&ccb=1-3&_nc_sid=ae5e01&_nc_ohc=3hfSdvyecXMAX8pSTlR&_nc_ht=scontent-waw1-1.xx&oh=c10273646643a957294d0bfd583fedbef&oe=60C1C909.

*there has been much progress in tools to support developers, such as Google's recently launched Explainable AI tool for Cloud AI customers, providing explanations at scale remains a challenge because the detail of what is needed varies significantly from sector to sector and across audiences*[33].

Google

## Skowronki lepszych praktyk

Mimo tego, że większość firm, z którymi mieliśmy szansę porozmawiać, nie uważa przejrzystości algorytmów uczenia maszynowego za wartość, którą są gotowe umieścić na sztandarze (w przeciwieństwie do przejrzystości innych procesów, np. przetwarzania danych osobowych) i niezwłocznie przełożyć na język *user experience*, dostrzegamy w ich działaniach komunikacyjnych pierwsze dobre praktyki. Opisujemy je w tej części briefu, pokazując, co nam się w nich szczególnie podoba, a czego nam jeszcze brakuje[34].

Nie krytykujemy, że jest za mało; cieszymy się, że – mimo braku obowiązków prawnych – renomowane firmy budują własny standard przejrzystości i z własnej inicjatywy publikują cenne informacje. Te przykłady pokazują, że nasze pomysły i rekomendacje (sformułowane w pierwszej części briefu) mają mocne zaczepienie w rzeczywistości. W istocie **nie proponujemy „przejrzystościowej" rewolucji**, a jedynie przesunięcie na wyższy poziom schematów komunikacyjnych i procedur, które już funkcjonują na rynku usług wykorzystujących uczenie maszynowe.

### ❖ Netflix

#### Rozwiązanie:

Strona edukacyjna „How Netflix's Recommendations System Works"[35]

#### Nasz komentarz:

Zwięźle i jasno wytłumaczona logika optymalizacji oraz zasad działania systemu (to, jakie dane bierze pod uwagę, co próbuje przewidzieć, jakie rezultaty wytwarza).

*Our recommendations system strives to help you find a show or movie to enjoy with minimal effort. We estimate the likelihood that you will watch a particular title in our catalog based on a number of factors including:*

- *your interactions with our service (such as your viewing history and how you rated other titles),*
- *other members with similar tastes and preferences on our service, and*

[33] https://www.blog.google/documents/77/Googles_submission_to_EC_AI_consultation_1.pdf

[34] Większość z wątpliwości, jakie pozostawiła lektura materiałów omawianych w tej części briefu, udało nam się wyjaśnić w czasie rozmów z przedstawicielami firm. Jednak ze względu na roboczy i poufny charakter naszych rozmów, opisując dobre praktyki, opieramy się wyłącznie na publicznie dostępnych materiałach.

[35] https://help.netflix.com/en/node/100639

- *information about the titles, such as their genre, categories, actors, release year, etc.*

*In addition to knowing what you have watched on Netflix, to best personalize the recommendations we also look at things like:*

- *the time of day you watch,*
- *the devices you are watching Netflix on, and*
- *how long you watch.*

*All of these pieces of data are used as inputs that we process in our algorithms. (…) The recommendations system does not include demographic information (such as age or gender) as part of the decision making process.*

*We take feedback from every visit to the Netflix service and continually re-train our algorithms with those signals to improve the accuracy of their prediction of what you're most likely to watch.*

Mimo zapewnień firmy, że nie zbiera i nie bierze pod uwagę **danych demograficznych**, na stronie dla konsumentów brakuje nam pełnej informacji o tym, jakie **dane behawioralne** mogą być brane pod uwagę przy wytwarzaniu rekomendacji, oraz wyjaśnienia, co oznacza „podobieństwo" między użytkownikami.

- Jak Netflix je ustala?
- Do jakich kategorii w związku z tym przypisuje swoich użytkowników?
- Czy i w jaki sposób (np. poprzez ustawienia prywatności) można te kategorie samodzielnie zweryfikować lub zmienić?

## Rozwiązanie:

Możliwość przekazania informacji zwrotnej (dla konkretnej rekomendacji)

## Nasz komentarz:

Użytkownik w prosty sposób, intuicyjnie („łapka w górę" albo „łapka w dół") może przekazać informację zwrotną dla każdej rekomendacji. Może w dowolnym momencie skorygować tę informację (zmienić zdanie co do wcześniej ocenionych treści). Może też usunąć konkretny tytuł z wyników wyszukiwania, co przełoży się na przyszłe rekomendacje.

## Rozwiązanie:

Strona informacyjna tłumacząca, jak działa przekazywanie informacji zwrotnej[36]

## Nasz komentarz:

Pomocne wyjaśnienie, napisane zwięzłym i prostym językiem. Zachęca do udzielania informacji zwrotnej, na bazie której uczy się algorytm odpowiedzialny za generowanie rekomendacji.

[36] https://help.netflix.com/en/node/9898

Wyjaśnienie nie pozostawia wątpliwości co do tego, że algorytm rekomendujący bierze pod uwagę także inne sygnały (w tym **to, co oglądał dany użytkownik** oraz **jemu podobni**).

Z perspektywy organizacji społecznej, zainteresowanej ochroną danych osobowych, strona nadal jednak pozostawia wątpliwości dotyczące określeń:

- *Your specific profile* (co się na niego składa, czy można go obejrzeć i skorygować);
- *Your viewing history* (co się na nią składa: czy tylko oglądane tytuły, czy również wzorce zachowania osoby oglądającej).[37]

## Rozwiązanie:

Opracowanie „The Netflix Recommender System: Algorithms, Business Value, and Innovation"[38]

## Nasz komentarz:

**Kompleksowe opracowanie na temat zasad działania algorytmów** (jakie dane i inne zmienne biorą pod uwagę, co próbują przewidzieć, jak są trenowane i testowane) **oraz logiki optymalizacji** (zakładane cele, przyjęte metryki, zwalczane błędy, eksperymenty zmierzające do poprawy wyników).

Wybrane fragmenty ilustrują język (wymagający pewnego przygotowania merytorycznego, ale zrozumiały dla czytelnika bez wiedzy branżowej) i poziom szczegółowości (przeważnie satysfakcjonujący).

- Rodzaj wykorzystywanych modeli i algorytmów

  *In general, our different video ranking algorithms use different mathematical and statistical models,different signals and data as input, and require different model trainings designed for the specific purpose each ranker serves. Each of the algorithms in our recommender system relies on statistical and machine-learning techniques. This includes both supervised (classification, regression) and unsupervised approaches (dimensionality reduction through clustering or compression, e.g., through topic models.*

- Dane wejściowe

  *Data that describe what each Netflix member watches, how each member watches (e.g., the device, time of day, day of week, intensity of watching), the place in our product in which each video was discovered, and even the recommendations that were shown but not played in each session.*

[37] Przedstawiciele Netflixa w rozmowie potwierdzili, że **nie tworzą profili behawioralnych** poszczególnych użytkowników: "We only look at the likelihood a profile, within an account, may like a particular title, but we do not build any individual profiles or use any demographic information of end-users". Natomiast w odniesieniu do historii oglądanych treści zapewnili, że "We try and **place control in the hands of the end-users themselves** and making sure they are aware of what viewing history has been collected, how they can download it, and how easy it is to delete any instance or all of it is key."

[38] https://dl.acm.org/doi/pdf/10.1145/2843948

*Our member coldstart approach has evolved into a survey given during the sign-up process, during which we ask new members to select videos from an algorithmically populated set that we use as input into all of our algorithms.*

- Cele algorytmów

  *Personalization allows us to significantly increase our chances of success when offering recommendations. One metric that gets at this is the take-rate – the fraction of recommendations offered resulting in a play. The lift in take-rate that we get from recommendations is substantial. But, most important, when produced and used correctly, recommendations lead to meaningful increases in overall engagement with the product (e.g., streaming hours) and lower subscription cancellations rates.*

  *Evidence algorithms decide whether to show that a certain movie won an Oscar or instead show the member that the movie is similar to another video recently watched by that member; they also decide which image out of several versions to use to best support a given recommendation.*

- Przykłady metryk

  *How do we know when an algorithm variant is better or worse than another? Revenue is proportional to the number of members, and three processes directly affect this number: the acquisition rate of new members, member cancellation rates, and the rate at which former members rejoin. If we create a more compelling service by offering better personalized recommendations, we induce members who were on the fence to stay longer, and improve retention. The main measurement target of changes to our recommendation algorithms is improved member retention.*

  *We analyze the resulting data to answer several questions about member behavior from a statistical perspective, including:*

  *– Are members finding the part of the product that was changed relative to the control more useful? For example, are they finding more videos to watch from the videosimilars algorithm than in the control?*

  *– Are members in a test cell streaming more on Netflix than in the control? For example, is the median or other percentile of hours streamed per member for the duration of the test higher in a test cell than in the control?*

  *– Are members in a test cell retaining their Netflix subscription more than members in the control?*

  *There are many other possible metrics that we could use, such as time to first play, sessions without a play, days with a play, number of abandoned plays, and more. Each of these changes, perhaps quite sensitively, with variations in algorithms, but we are unable to judge which changes are for the better.*

- Niepożądane błędy, zaobserwowane problemy

> *Most of our statistical models, as well as the standard mathematical techniques used to generate recommendations, do not take a positive feedback loop into account. It is very likely that better algorithms (…) will remove the potential negative effects of such a feedback loop and result in better recommendations.*
>
> *Children's viewing presents a particular problem in shared profiles, since kid videos tend to be shorter, and because young children have a predilection to view the same movie or episode many times, which is not a behavior typical of adults, and which can lead to very strange biases to the recommendations generated from that data.*

Po lekturze tego tekstu **nadal** jednak **nie jest dla nas jasne**:

- jakie konkretnie dane są zbierane o ludziach (*personalized signals*);
- jakie na tej podstawie są wyciągane wnioski na ich temat (*inferred information*) i czy w związku z tym ludzie są przypisywani do określonych kategorii;
- w jaki sposób i czy w ogóle jest oceniane ryzyko z perspektywy użytkownika (np. uzależnienia od usługi, ekspozycji dzieci na niestosowne dla nich treści, wzmocnienia problemów psychicznych lub emocjonalnych etc.), a nie firmy (niespełnienia zakładanych celów biznesowych).

## ❖ Allegro

### Rozwiązanie:

Strona informacyjna, wyjaśniająca sortowanie po trafności dla sprzedających[39] i dla kupujących[40]

### Nasz komentarz:

**Strona napisana zrozumiałym, zwięzłym językiem**, bez popadania w żargon branżowy. Wyjaśnia najważniejsze wątpliwości, jakie może mieć przeciętny użytkownik systemu (sprzedający lub kupujący na Allegro):

- Cele algorytmu; problem, który Allegro próbuje rozwiązać

> *W sortowaniu po trafności staramy się jak najlepiej dopasować kolejność ofert na liście do tego, czego szuka kupujący.*

[39] https://allegro.pl/dla-sprzedajacych/trafnosc-xGmVjoPwOTo

[40] https://allegro.pl/pomoc/dla-kupujacych/wyszukiwanie-i-obserwowanie/jestem-kupujacym-na-czym-polega-sortowanie-po-trafnosci-lDkngqDYETA

- Zasada działania algorytmu wyjaśniona poprzez przykład

> *Oznacza to, że dla dwóch bardzo podobnych zapytań o produkt, np. „samsung galaxy s3" i „samsung galaxy", możesz zobaczyć inną kolejność ofert na liście. Wpływa na to kilka czynników:*
>
> 1. *zapytanie, jakie kupujący definiuje, gdy szuka danego przedmiotu (np. wyszukuje daną frazę w wyszukiwarce, zaznacza filtry, wybiera kategorie);*
> 2. *dopasowanie tytułu oferty do frazy wyszukiwania;*
> 3. *to, jak zachowywali się na listach ofert inni użytkownicy, np. jakich ofert szukali, jakie oferty oglądali, które oferty kupili.*

- Dane, jakie algorytm bierze pod uwagę

Strona wymienia szereg „danych o ofercie" i „danych o sprzedającym", które kształtują pozycję oferty na liście wyników. Ogólnie definiuje również, co się składa na „Zaangażowanie kupujących" (złożony parametr, jaki algorytm bierze pod uwagę).

- Omówienie problemów, na jakie może trafić klient; wyjaśnienie zasady działania algorytmu na przykładach

Na przykład: „Zmieniłem pierwsze zdjęcie w mojej ofercie i zauważyłem, że zmieniła ona miejsce na liście. Dlaczego?" albo „Moja oferta ma dużą sprzedaż a mimo to jest niżej w sortowaniu po trafności niż inne nowe oferty. Dlaczego?".

Z perspektywy organizacji społecznej, która szuka odpowiedzi na specyficzne pytania (np. na jaki efekt system został zoptymalizowany i czy takie działanie algorytmu może prowadzić do niepożądanych skutków, takich jak dyskryminacja; jakie dane o ludziach system bierze pod uwagę; jak została dla niego przeprowadzona analiza ryzyka), lektura strony **nie wyjaśnia wszystkich wątpliwości**.

- Czy w ogóle, a jeśli tak – to jakie dane użytkownika (według swojej polityki prywatności Allegro gromadzi „informacje o Twoich działaniach na Platformach, m.in. o historii Twoich zakupów, wystawianych Ofertach, wybieranych metodach płatności oraz o treściach komentarzy i wystawianych ocen") mają wpływ na sortowanie ofert?
- Jakie typy błędów (lub przejawy tendencyjności) mogą występować w systemie? W jaki sposób są minimalizowane?
- Jak Allegro przeprowadza analizę ryzyka i ewaluację systemu?

## ❖ YouTube

### Rozwiązanie:

Możliwość przekazania informacji zwrotnej nt. konkretnych rekomendacji[41]

[41] https://support.google.com/youtube/answer/6125535?hl=en&ref_topic=9257501

## Nasz komentarz:

Strona informacyjna dla użytkowników serwisu z **prostymi instrukcjami** na temat tego, jak sami mogą wpłynąć na rekomendowane im treści:

> *There are several ways to influence these recommendations and search results. You can remove specific videos from your watch history and searches from your search history. You can also pause your watch and search history, or start fresh by clearing your watch and search history.*
>
> *Your „Not Interested" feedback may be used to tune your recommendations.*

Z eksperckiego opracowania cytowanego poniżej wynika, że **profil behawioralny użytkownika nie jest brany pod uwagę w procesie personalizowania rekomendacji**. W tym kontekście zastanawiają jednak sformułowania takie jak *may be used to tune your recommendations*, ponieważ sugerują, że „silnik" YouTube'a (wykorzystujący uczenie maszynowe) karmi się **nie tylko** sygnałami wymienionymi na tej stronie. W tym kontekście mamy następujące wątpliwości:

- Co jeszcze jest brane pod uwagę? I czy sam użytkownik może na te (inne, nieosobowe) czynniki w jakikolwiek sposób oddziaływać?
- Czy na rekomendacje wpływają statystyczne wzorce zachowań (np. to, jak angażuje się grupa osób „podobnych" do użytkownika)?

Czy w związku z tym serwis będzie rekomendował w miarę korzystania coraz bardziej radykalne/emocjonujące treści? A jeśli tak, w jaki sposób konkretny użytkownik może przerwać tę eskalację?

## Rozwiązanie:

Opracowanie „Deep Neural Networks for YouTube Recommendations"[42]

## Nasz komentarz:

**Szczegółowe, poparte danymi opracowanie na temat zasad działania algorytmów** (jakie dane i inne zmienne są brane pod uwagę, co próbują przewidzieć, jak są trenowane i testowane) **oraz logiki optymalizacji** (zakładane cele, przyjęte metryki, zwalczane błędy, eksperymenty zmierzające do poprawy wyników).

Wybrane fragmenty ilustrują język (zdecydowanie wymagający wiedzy branżowej) i poziom szczegółowości (więcej niż satysfakcjonujący):

- Rodzaj i ogólne zasady działania algorytmów

> *The system is comprised of two neural networks: one for candidate generation and one for ranking. Presenting a few „best" recommendations in a list requires a fine-level representation to distinguish relative importance among candidates with high recall.*

[42] https://static.googleusercontent.com/media/research.google.com/en//pubs/archive/45530.pdf

*The ranking network accomplishes this task by assigning a score to each video according to a desired objective function using a rich set of features describing the video and user. We pose recommendation as extreme multiclass classification where the prediction problem becomes accurately classifying a specific video among millions of videos(classes) from a corpus V based on a user U and context C.*

*We use a deep neural network with similar architecture as candidate generation to assign an independent score to each video impression using logistic regression.*

- Dane wejściowe (źródła danych, zakres danych i przyjęte kategoryzacje)

  *A user's watch history is represented by a variable-length sequence of sparse video IDs which is mapped to a dense vector representation via the embeddings. Search history is treated similarly to watch history – each query is tokenized into unigrams and bigrams and each to-ken is embedded. Once averaged, the user's tokenized, embedded queries represent a summarized dense search history. Demographic features are important for providing priors so that the recommendations behave reasonably for new users. The user's geographic region and device are embedded and concatenated. Simple binary and continuous features such as the user's gender, logged-in state and age are input directly into the network as real values normalized to [0,1].*

  *During ranking, we have access to many more features describing the video and the user's relationship to the video because only a few hundred videos are being scored rather than the millions scored in candidate generation.*

  *Recommendation systems in particular benefit from specialized features describing past user behavior with items.*

  *We „rollback" a user's history by choosing a random watch and only input actions the user took before the held-out label watch.*

- Cele algorytmów, wykorzystywane metryki

  *Our goal is to predict expected watch time given training examples that are either positive (the video impression was clicked) or negative (the impression was not clicked). Positive examples are annotated with the amount of time the user spent watching the video. To predict expected watch time we use the technique of weighted logistic regression, which was developed for this purpose.*

  *During development, we make extensive use of offline metrics (precision, recall, ranking loss, etc.) to guide iterative improvements to our system.*

  *Logistic regression was modified by weighting training examples with watch time for positive examples and unity for negative examples, allowing us to learn odds that closely model expected watch time.*

- Niepożądane błędy, zaobserwowane problemy

> *A key insight that improved live metrics was to generate a fixed number of training examples per user, effectively weighting our users equally in the loss function. This prevented a small cohort of highly active users from dominating the loss.*
>
> *Great care must be taken to withhold information from the classifier in order to prevent the model from exploiting the structure of the site and overfitting the surrogate problem.*
>
> *Withholding discrimative signals from the classifier was also essential to achieving good results – otherwise the model would overfit the surrogate problem and not transfer well to the homepage.*
>
> *Ranking by click-through rate often promotes deceptive videos that the user does not complete („clickbait") whereas watch time better captures engagement.*
>
> *Using the age of the training example as an input feature removes an inherent bias towards the past and allows the model to represent the time-dependent behavior of popular videos.*

Po lekturze tego tekstu **nadal nie jest dla nas jasne**:

- czy obok tych wyraźnie wymienionych brane są pod uwagę także inne dane osobowe (w tekście pojawia się sformułowanie *such as*, a więc podane przykłady nie są wyczerpujące);
- czy w ramach profilowania na potrzeby działania opisanych algorytmów użytkownicy są przypisywani do stałych kategorii;
- jak (i czy w ogóle) są oceniane ryzyka z perspektywy użytkowników, a nie celów biznesowych firmy.

W technicznym, bardzo precyzyjnym języku (odwołującym się do formuł matematycznych) gubią się ogólniejsze wnioski i założenia, które byłyby cenne dla ekspertów z innych dziedzin (np. prawników).

## Problem na horyzoncie

Dialog z firmami oferującymi usługi oparte na rozwiązaniach uczenia maszynowego potwierdza istnienie problemu, którego byliśmy świadomi już w momencie przystępowania do badania. Dojmująco **brakuje języka do opisu technologii** napędzającej innowacyjne usługi konsumenckie, który byłby i precyzyjny (tj. nie przekłamywałby technologicznej rzeczywistości), i zrozumiały dla ludzi bez wiedzy eksperckiej. Ta bariera powoduje, że nawet firmy, które już widzą korzyść w większej przejrzystości rozwiązań uczenia maszynowego i otwartej komunikacji z konsumentem w tym konkretnym aspekcie (np. w kontekście pozyskiwania informacji zwrotnej dla konkretnych rekomendacji), nie są w stanie zrealizować swoich zamierzeń.

Ten problem ilustrują opisane powyżej „Skowronki lepszych praktyk". W naszej ocenie **dostępne publicznie materiały informacyjne są albo zbyt proste i wybiórcze** (jako kierowane do przeciętnego odbiorcy), **albo zbyt zaawansowane, ponieważ ich zrozumienie wymaga wiedzy branżowej**.

Materiały przyjazne konsumentom są uproszczone do takiego poziomu, który nie pozwala ani na odtworzenie faktycznych zasad działania ML w danej usłudze, ani na ocenę (na własne potrzeby) związanego z nią ryzyka („czy na pewno ta usługa jest dla mnie bezpieczna?"). Materiały dla zaawansowanych użytkowników odpowiadają wprawdzie na szczegółowe pytania, ale tylko te, które ich twórcy uznali za ciekawe. Jest tam miejsce na opis założeń biznesowych, przeprowadzanych eksperymentów (mających na celu ulepszenie usługi) i problemów, z jakimi mierzyli się deweloperzy – natomiast brakuje analizy ryzyk, które miałyby znaczenie z perspektywy konsumenta.

Mając świadomość braku adekwatnego języka, przyznajemy, że zaprojektowanie choćby strony edukacyjnej na temat rozwiązań uczenia maszynowego przeznaczonej dla konsumentów, którzy chcą się dowiedzieć czegoś więcej, to nie lada wyzwanie! **Nie ma ani pojęć** (szczególnie w języku polskim), które opisywałyby tę matematyczno-technologiczną rzeczywistość w sposób zrozumiały dla nieekspertów, **ani uwspólnionych wyobrażeń o tym, czym w istocie są rozwiązania uczenia maszynowego** (nie magia, ale też nie zwyczajna statystyka ani nawet licząca tabela w Excelu). W powijakach jest też kultura otwartej komunikacji pomiędzy firmami i konsumentami, jeśli chodzi o sygnalizowanie problemów i ryzyk (tego, co może „pójść nie tak" i co firma robi, żeby „było dobrze").

Te luki w masowej wyobraźni, języku i kulturze komunikacji domagają się reakcji. Nie jest to jednak zadanie dla jednej firmy (nawet globalnej) ani dla jednej organizacji społecznej (nawet najbardziej zdeterminowanej!). To długofalowe, wielowymiarowe wyzwanie, z którym warto zmierzyć się wspólnie – jako **społeczność złożona z różnorodnych interesariuszy** o uzupełniających się kompetencjach.

W Fundacji Panoptykon podjęliśmy już pierwsze eksperymenty z wykuwaniem nowych pojęć i metafor, które mogłyby umożliwić publiczną, otwartą rozmowę o rozwiązaniach uczenia maszynowego (ich celach, zasadach działania, związanych z nimi ryzykach i ich społecznym oddziaływaniu). W 2020 r. opublikowaliśmy „Sztuczną inteligencję non-fiction", czyli przewodnik po kluczowych dla tej dziedziny pojęciach i zasadach, napisany językiem zrozumiałym dla ludzi spoza branży *data science*[43].

Mimo entuzjastycznego przyjęcia publikacji przez praktyków z innych branż, w tym zajmujących się sztuczną inteligencją prawników, aktywistów i wykładowców akademickich, mamy pełną świadomość, że nadal jest nam bardzo daleko do wypracowania standardów, które udrożnią **komunikację z przeciętnym konsumentem**. Tej przepaści komunikacyjnej sami nie przeskoczymy i dlatego – z pełną odpowiedzialnością za słowo – proponujemy firmom, które świadczą usługi konsumenckie wykorzystujące ML, wspólne działania na tym polu.

[43] https://panoptykon.org/sztuczna-inteligencja-non-fiction

# III. DALSZE KROKI: CO PROPONUJEMY?

## Po pierwsze, zbadajmy (głębiej) oczekiwania konsumentów

Każda szanująca się firma (w tym wszystkie, z którymi mieliśmy okazję porozmawiać) bada oczekiwania i potrzeby swoich odbiorców. To rynkowy standard. Mimo to „sztuczna inteligencja w usługach konsumenckich" nie weszła jeszcze do zestawu tematów, o których ankieterzy rozmawiają wprost z badanymi. Z przeprowadzonych przez nas dyskusji wynika, że do tej pory tylko Google pytał swoich odbiorców (w ramach badań fokusowych i ankiet prowadzonych online) o **oczekiwania dotyczące przejrzystości usług wykorzystujących uczenie maszynowe**. Odpowiedzi potwierdzały, że takich oczekiwań odbiorcy nie mają.

Pozostałe firmy w prowadzonych badaniach nie badały wprost oczekiwań dotyczących przejrzystości działania „silnika" serwisu, natomiast pytały o zadowolenie z tego, jak działa serwis (w tym *user experience*[44] i spersonalizowane rekomendacje[45]), oraz o zaufanie do niego i poczucie bezpieczeństwa[46]. Allegro przeprowadziło badania jakościowe, w których użytkownicy oceniali trafność algorytmu (listy ofert), a dział firmy wspierający sprzedających na bieżąco monitoruje i reaguje na wątpliwości dotyczące rankowania wyników.

Biorąc pod uwagę opisane w poprzednich punktach wyzwania (brak języka; brak wspólnych wyobrażeń; niska społeczna świadomość tego, czym są rozwiązania uczenia maszynowego), zupełnie nas **nie dziwi, że konsumenci nie zgłaszają się do firm z pytaniami o zasady działania algorytmów** i nie sygnalizują potrzeby wyjaśnienia, jak działa ta technologia.

Co więcej, intuicyjnie zgadzamy się z poglądem, który mocno wybrzmiał w naszych rozmowach z przedstawicielami firm: konsument, który nie doświadczył na własnej skórze błędów ani tendencyjności w działaniu systemu napędzanego sztuczną inteligencją, raczej nie będzie miał ochoty wnikać w to, jak system działa. Dla przeciętnego odbiorcy ważniejsze niż przejrzystość i możliwość zajrzenia „pod maskę" jest **poczucie bezpieczeństwa, przewidywalność systemu i intuicyjność interfejsów**. Posiłkując się motoryzacyjną metaforą: oczekiwanie konsumenta jest takie, by samochód po prostu jeździł. Działanie silnika to już zmartwienie producenta.

Ponieważ rozwiązania uczenia maszynowego są obecne w usługach konsumenckich stosunkowo od niedawna (w każdym razie kilka dekad krócej niż samochody), nadal mają status technologii na poły magicznej. Nie bardzo jeszcze wiadomo, **po czym poznać, że produkt lub usługa ma „w środku" sztuczną inteligencję ani czy ta inteligencja rzeczywiście działa** (czy robi to, do czego została przeznaczona; jaką ma skuteczność; czy popełnia błędy; ew. jak poważne są to błędy i w czyje interesy uderzają). W naszym przekonaniu przeciętny konsument, nawet gdyby czuł się zaniepokojony lub skonfundowany efektem działania algorytmu, najpewniej nie wiedziałby nawet, jak sformułować pytanie i jaki konkretnie problem zgłosić. Dlatego jesteśmy przekonani, że milczenie konsumentów na temat rozwiązań uczenia maszynowego nie jest ich wyborem. Może raczej stanowić dowód ich bezradności, niskiej świadomości i poczucia zagubienia.

[44] OLX

[45] Allegro

[46] Netflix

Tę intuicję chętnie byśmy potwierdzili we **wspólnie zaprojektowanych badaniach**, które pozwoliłyby nam określić, co tak naprawdę kryje się za milczeniem konsumentów. Jakie mają przekonania na temat sztucznej inteligencji? Jakie – z ich perspektywy – obietnice wiążą się z jej wykorzystaniem w usługach? Czy chcieliby, żeby ktoś (za nich lub dla nich) sprawdzał, jak ów system działa i czy działa uczciwie? Jak sami definiują „uczciwość" w działaniu ML? Jakich niepożądanych efektów obawiają się najbardziej? Czy chcieliby móc zgłaszać swoje problemy i obawy? A jeśli tak – to w jakiej formie?

Naturalnie, metoda i zakres musiałyby uwzględnić, że nie mamy na celu przeprowadzenia badania istniejących, stabilnych przekonań, ale **analizę potrzeb, które dopiero się kształtują** (nierzadko na bazie irracjonalnych lęków lub nieprawdziwych przeświadczeń na temat działania ML). Warto zbadać stan obecny choćby po to, żeby móc się od niego odbić przy projektowaniu komunikatów i narzędzi wspierających dialog na linii firma–konsument i kształtujących dojrzałe potrzeby odbiorców.

## Po drugie, wypracujmy wspólnie dobre praktyki

Obserwując od dekady rozwój technologii wykorzystywanych w usługach konsumenckich i ewolucję podążających za tym rozwojem europejskich regulacji, jesteśmy przekonani, że lata **2021 i 2022 to czas na budowanie dobrych praktyk w obszarze wykorzystania i komunikowania rozwiązań uczenia maszynowego**.

Komisja Europejska w proponowanej regulacji sztucznej inteligencji zawahała się przed wprowadzeniem dalej idących obowiązków dla firm świadczących usługi uczenia maszynowego ograniczonego i niskiego ryzyka. W miejsce twardych wytycznych (z wyjątkiem podstawowego obowiązku ostrzegania konsumenta, że wchodzi w kontakt ze sztuczną inteligencją, por. art. 52) w art. 69 pojawia się zachęta do samoregulacji, w tym tworzenia **branżowych kodeksów dobrych praktyk**.

Nasze doświadczenie z regulacją reklamy behawioralnej oraz usług świadczonych przez platformy internetowe pokazuje, że europejski prawodawca nie rzuca słów na wiatr. Należy się zatem spodziewać rosnącego zainteresowania Brukseli rynkiem usług napędzanych przez rozwiązania uczenia maszynowego. Prognozujemy, że za wezwaniem do samoregulacji, o ile firmy nie podejmą zaproszenia w poważny sposób, pójdą **twardsze działania regulacyjne**[47].

Wyprzedzenie spodziewanych ruchów europejskiego regulatora to tylko jeden z powodów, dla których warto wypracować dobre praktyki (szczególnie komunikacyjne) dla usług wykorzystujących algorytmy uczenia maszynowego. Z perspektywy firm ceniących sobie otwartość i zaufanie w relacji z konsumentami o wiele ważniejszym (choć nadal pragmatycznym) powodem jest **wychowanie świadomych odbiorców, którzy będą w stanie aktywnie zgłaszać błędy i inne niepożądane efekty**, a przez to pomagać w trenowaniu algorytmów ML. Taka grupa kontrolna może się w dłuższej perspektywie okazać bezcennym zasobem przy rozwijaniu coraz bardziej wyrafinowanych, a przez to również obarczonych większym ryzykiem, usług.

[47] Proces, który doprowadził do zaostrzenia kursu regulacyjnego wobec platform internetowych oraz branży adtechowej i pojawienia się w dyskursie dość radykalnego w założeniach pakietu DSA/DMA, opisujemy m.in. w tekście „Demokracja kontra firmy technologiczne. Wielka zmiana na horyzoncie" https://holistic.news/demokracja-kontra-firmy-technologiczne-wielka-zmiana-na-horyzoncie/.

Jesteśmy przekonani, że edukowanie własnych konsumentów w obszarze sztucznej inteligencji to dobra baza pod **zarządzanie przyszłymi kryzysami komunikacyjnymi oraz kryzysami zaufania**. Doświadczenie firm, które znajdowały się w awangardzie usług wykorzystujących uczenie maszynowe (np. w reklamie behawioralnej[48], systemach rekomendacyjnych[49] i scoringu[50]), pokazuje, że pojawienie się takich zarzutów, jak sprzeczna z potrzebami konsumentów logika optymalizacji, ingerowanie w prywatność osób fizycznych czy negatywne skutki społeczne, to tylko kwestia czasu. Im bardziej wyrafinowany produkt i technologia go napędzająca, tym większe prawdopodobieństwo nieporozumienia. Jedną rzeczą jest krytykować zasadnie i wchodzić z firmami w merytoryczny dialog, inną – krytykować z powodu niezrozumienia lub na zapas, w obawie przed nieznaną technologią. Najlepszą **polisą ubezpieczeniową dla firm** obawiających się tego drugiego scenariusza jest właśnie edukacja i otwartość w komunikacji z użytkownikami.

Nie bagatelizujemy zadania stojącego przed firmami, które chciałyby tworzyć dobre praktyki komunikacyjne w nierozpoznanej jeszcze dziedzinie. Rozmiar i powagę wyzwania, jakim będzie wypracowanie odpowiedniego języka i kodów komunikacyjnych, pokazujemy w punkcie „Problem na horyzoncie"). Jesteśmy jednak przekonani, że właśnie tutaj otwiera się **pole do ważnych społecznie eksperymentów** (np. w obszarze projektowania interfejsów i nowego *user experience*) oraz autentycznej innowacji.

Ze swojej strony deklarujemy gotowość do uczestnictwa w takich eksperymentach. Możemy wystąpić zarówno w roli testerów, jak i konsultantów. Interesuje nas szczególnie wspólne wypracowanie pomysłu na:

- etykietę informującą o tym, że konsument wchodzi w kontakt ze sztuczną inteligencją;
- stronę edukacyjną tłumaczącą cele i zasady działania sztucznej inteligencji w konkretnej usłudze konsumenckiej;
- modelowy proces oceny ryzyka dla praw i interesów konsumentów, z którego wnioski byłyby publicznie udostępniane (z myślą o zaawansowanych odbiorcach).

Źródłem inspiracji w projektowaniu takich komunikatów i procesów mogą być standardy wypracowane w obszarach, w których regulacja już wcześniej – mimo podobnych wyzwań (tj. specjalistycznego języka i obiektywnej złożoności produktu) – wymusiła większą przejrzystość. Jako przykład można wskazać:

- system etykiet i ostrzeżeń dla sprzętu elektrycznego[51],
- etykiety dla leków i produktów medycznych[52],

[48] https://algorithmwatch.org/en/automated-discrimination-facebook-google/

[49] https://www.wired.com/story/our-minds-have-been-hijacked-by-our-phones-tristan-harris-wants-to-rescue-them/

[50] https://www.forbes.com/advisor/credit-cards/from-inherent-racial-bias-to-incorrect-data-the-problems-with-current-credit-scoring-models/

[51] https://europa.eu/youreurope/business/product-requirements/labels-markings/index_pl.htm

[52] https://sip.lex.pl/akty-prawne/dzu-dziennik-ustaw/wymagania-dotyczace-oznakowania-opakowan-produktu-leczniczego-i-tresci-17529132

- etykiety informujące o składzie żywności[53] (i przyznawanych certyfikatach)[54],
- standardy oceny ryzyka wypracowane przez ISO[55].

## ANEKS. O BADANIU

W marcu 2021 r. przeprowadziliśmy serię wywiadów z przedstawicielami i przedstawicielkami firm dostarczających usługi wykorzystujące algorytmy uczenia maszynowego (ML). Do rozmowy zaprosiliśmy firmy wykorzystujące ML w następujących sektorach usług:

- na platformach e-commerce, m.in. w celu personalizacji oferty: Allegro i OLX;
- w systemach rekomendacyjnych platform prezentujących treści wideo: Netflix i Google (YouTube) – oraz treści różnego typu: Facebook (z tą firmą nie udało nam się przeprowadzić wywiadu, ale otrzymaliśmy od niej wyjaśnienia mailowo);
- w usługach finansowych, m.in. w celu oszacowania ryzyka i oceny wiarygodności klientów i klientek: IBM, Spółdzielnia ANG.

**W tym miejscu serdecznie dziękujemy naszym rozmówcom za autentyczne zaangażowanie, poświęcony czas i otwartość na dialog!**

W wywiadach pytaliśmy:

- o wykorzystywane rozwiązania uczenia maszynowego

Chcieliśmy się dowiedzieć, jakie algorytmy (typ, klasa) są wykorzystywane przez firmy; jaką wartość dodaną te rozwiązania wnoszą i jak działają (w szczególności: co konkretnie próbują przewidzieć i na jaki efekt są zoptymalizowane).

- o ryzyko i przejrzystość rozwiązań uczenia maszynowego w kontekście przygotowywanej przez Unię Europejską regulacji

Pytaliśmy o to, jak firmy definiują ryzyko związane z wykorzystaniem ML w usługach konsumenckich. Jak rozumieją pojęcia takie jak „rozliczalność" czy „bezpieczeństwo" systemu z perspektywy konsumentów czy „negatywny wpływ na (prawa) człowieka", które pojawiają się w przygotowywanej przez UE regulacji. Chcieliśmy też dowiedzieć się, w jaki sposób firmy przeprowadzają ocenę ryzyka i czy jej wyniki są komunikowane na zewnątrz.

- o informowanie użytkowników i użytkowniczek o wykorzystywanych algorytmach ML

Chcieliśmy się dowiedzieć, czy firmy informują o algorytmach wykorzystywanych w swoich usługach. Jak to robią? Co wiedzą o oczekiwaniach swoich użytkowników i użytkowniczek w tym zakresie (i skąd czerpią tę wiedzę)? Pytaliśmy również o otwartość firm na ujawnianie konkretnych informacji o ich rozwiązaniach uczenia maszynowego, m.in. w odpowiedzi na oczekiwania organizacji społecznych. Od czego jest uzależniona ich gotowość do dzielenia się informacjami? Jakie dostrzegają korzyści lub zagrożenia w tym obszarze?

[53] http://publications.europa.eu/resource/cellar/801c0034-e55a-40ba-9506-d70cc13ffeea.0018.02/DOC_2

[54] https://spolecznosci.fairtrade.org.pl/o-kampanii/certyfikaty-i-oznaczenia-produktow-sprawiedliwego-handlu/

[55] https://www.iso.org.pl/uslugi-zarzadzania/wdrazanie-systemow/zarzadzanie-ryzykiem/iso-31000/

W naszym badaniu wykorzystaliśmy też udostępnione przez firmy materiały (informacyjne, edukacyjne, lobbingowe, naukowe):

- Allegro: strona dla sprzedających „Trafność" [https://allegro.pl/dla-sprzedajacych/trafnosc-xGmVjoPwOTo], strona dla kupujących „Na czym polega sortowanie po trafności?" [https://allegro.pl/pomoc/dla-kupujacych/wyszukiwanie-i-obserwowanie/jestem-kupujacym-na-czym-polega-sortowanie-po-trafnosci-lDkngqDYETA];
- European Tech Alliance: High-Level Principles on Artificial Intelligence [http://eutechalliance.eu/wp-content/uploads/2020/02/EUTA-High-Level-Principles-on-AI.pdf];
- Google: strona dla użytkowników i użytkowniczek nt. działania rekomendacji treści na YouTubie [https://www.youtube.com/intl/ALL_ie/howyoutubeworks/product-features/recommendations/];
- Google: opis sieci neuronowej wykorzystywanej w systemie rekomendacji treści na YouTubie [https://static.googleusercontent.com/media/research.google.com/en/pubs/archive/45530.pdf];
- Google: opis „Model Cards" [https://ai.googleblog.com/2020/07/introducing-model-card-toolkit-for.html];
- Google: stanowisko w konsultacjach Digital Services Act [https://blog.google/documents/89/Googles_submission_on_the_Digital_Services_Act_package_1.pdf];
- Google: stanowisko w konsultacjach White Paper on AI [https://www.blog.google/documents/77/Googles_submission_to_EC_AI_consultation_1.pdf];
- Facebook: stanowisko w konsultacjach Digital Services Act [https://about.fb.com/de/wp-content/uploads/sites/10/2020/09/FINAL-FB-Response-to-DSA-Consultations.pdf];
- Facebook: stanowisko w konsultacjach White Paper on AI [https://scontent-waw1-1.xx.fbcdn.net/v/t39.8562-6/103231277_1162782850727962_2719421119701851752_n.pdf?_nc_cat=103&ccb=1-3&_nc_sid=ae5e01&_nc_ohc=1nR9gYZtRosAX_MhKSI&_nc_ht=scontent-waw1-1.xx&oh=25ab88d45cbbd27d8e7896ac43050912&oe=60B1F709];
- Netflix: strona dla użytkowników i użytkowniczek nt. działania rekomendacji treści na platformie [https://help.netflix.com/en/node/100639];
- Netflix: blog informujący o technicznych aspektach działania platformy [https://netflixtechblog.com/];
- Netflix: opis algorytmów składających się na system rekomendacyjny tej firmy [https://dl.acm.org/doi/pdf/10.1145/2843948];
- IBM: stanowisko w konsultacjach White Paper on AI [https://www.ibm.com/blogs/policy/wp-content/uploads/2020/06/IBM-Submission-on-the-EC-AI-White-Paper.pdf];
- IBM: opis „AI FactSheets" [https://aifs360.mybluemix.net/introduction].