Uczenie maszyn

Czy jest możliwe projektowanie sprawiedliwych algorytmów AI?

Bartosz Rodziewicz, 226105

W ostatnim czasie algorytmy AI zaczynają być wykorzystywane przez coraz więcej firm, podmiotów i do coraz większej ilości zadań. Systemy AI są używane przez duże internetowe korporacje jak Google czy Facebook, by serwować bardziej dopasowane treści do danego odbiorcy, są wykorzystywane przez instytucje bankowe do oceny ryzyka udzielenia kredytu danej osobie, czy w medycynie do wczesnego diagnozowania pacjentów lub oznaczania pacjentów potencjalnego ryzyka. Sztuczna inteligencja zaczyna być stosowana w prawie każdej dziedzinie. Im więcej systemów AI powstało, tym bardziej zaczęto zauważać, że zaczynają one wykazywać bias wobec niektórych osób, przypadków czy zastosowań. Stąd też zaczęto badania nad tzw. sprawiedliwością AI (*ang. AI fairness*).

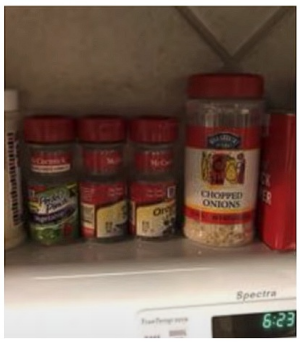
Zgodnie z bardziej powszechnie akceptowaną definicją [1] sprawiedliwa AI to taki system AI, który do podjęcia decyzji nie bierze pod uwagę cech ocenianej jednostki, które nie powinny mieć wpływu na wynik, zwłaszcza cech uważanych za wrażliwe. Przykładem takiego biasu wobec niektórych grup czy jednostek może być system rozpoznawania twarzy, który lepiej radzi sobie z analizą osób o jasnym kolorze skóry w porównaniu do ciemnoskórych osób, czy system oceny ryzyka kredytowego gorzej oceniający osoby o tych samych dochodach, ale mieszkające w "gorszej" dzielnicy miasta [2]. Aktualnie prowadzona dyskusja wobec tego tematu postrzega go jako jednoznacznie zły i konieczny do wyeliminowania [3]. Uważam jednak, że takie postrzeganie tematu jest przejściem ze skrajności w skrajność i postaram się pokazać, że potencjalny bias może nie jest czymś pożądanym, ale może być tolerowany w niektórych przypadkach.

Bias w algorytmach sztucznej inteligencji może pojawić się na każdym etapie tworzenia systemu. Nasz świat nie jest w pełni obiektywny. Dane wykorzystywane do tworzenia systemu mogą zawierać historyczne zaszłości widziane do teraz w naszym świecie, mogą nieodpowiednio reprezentować konkretne grupy osób lub mogą być zebrane na podstawie wycinka populacji, który jest łatwy do analizy, jednak niekoniecznie dobrze przekłada się na całe społeczeństwo (np. studenci). Bias może również pojawić się na etapie tworzenia i ewaluacji modelu, gdy dane lub metryki używane do oceny jakości modelu nie uwzględniają całego społeczeństwa lub w przypadku, gdy jeden model nie może odpowiednio pokryć całego społeczeństwa i konieczne jest wykorzystanie kilku zagregowanych (np. w medycynie wobec osób różniących się od siebie). Dodatkowo bias może pojawić się na etapie analizy działania modelu przez testera, który może przerzucić swoją, potencjalnie obciążoną biasem, wizję tego jakie wyniki powinien algorytm dawać [4].

Temat niesprawiedliwych algorytmów AI pojawił się w ostatnich latach i niedawno pojawiło się wiele prac badawczych poruszających go [5]. Niektóre z nich zaproponowały różnego rodzaju metody przeciwdziałania, które można zastosować na gotowy model, gdy zauważamy, że posiada on bias. Przykładem takich metod są: adversarial debiasing, dynamiczny upsampling danych uczących, czy distributionally robust optimization [2]. Adversarial debiasing polega na takiej modyfikacji algorytmu, aby był on "karany" w procesie uczenia za predykcje, po których można z dużym prawdopodobieństwem wyznaczyć wcześniej zdefiniowane cechy chronione. Korzystając z tych technik możemy próbować stworzyć algorytm, który przynajmniej w naszym rozumieniu jest bardziej sprawiedliwy.

Bias w algorytmach pojawia się z różnych powodów, cześć z nich można nazwać jako błędy przy tworzeniu algorytmu, inne już troche mniej. Weźmy za przykład algorytm rozpoznający twarze i cechę jaką jest kolor włosów. Wiele takich algorytmów uczonych jest na zbiorach które nadreprezentują osoby z brązowymi i blond włosami, kosztem osób z włosami koloru czarnego czy rudego [5]. Czysto sprawiedliwy system powinien być uczony na zbiorze, który reprezentowałby osoby z każdym kolorem włosów w tym samym stopniu, aby zgodnie z prawdopodobieństwem mieć najlepsze wyniki działania algorytmu dla każdej z grup. Zastanówmy się jednak czy aby na pewno jest to najlepsza droga. Czy aby nie byłoby lepszym rozwiązaniem użycie zbioru, który reprezentowałby w odpowiedni sposób przewidywaną grupę użytkowników naszego systemu. Jeśli nasz system jest planowany do użycia w Polsce, to czy nie byłoby lepszym rozwiązaniem, aby zestaw danych uczących reprezentował rozkład tej cechy w polskim społeczeństwie. Takie podejście powoduje, że może faktycznie osoby z grup mniej reprezentowanych otrzymują słabiej działający algorytm, ale średnio dla wszystkich użytkowników algorytm działa lepiej.

Taki poziom akceptacji pewnego biasu jest aktualnie zauważalny w algorytmach stosowanych przez duże korporacje dla których najważniejsze jest, aby algorytm działał dobrze dla większości użytkowników. Dobrym przykładem mogą być algorytmy kategoryzowania zawartości zdjęć. Za przykład można wziąć dwa poniższe zdjęcia i kategorie, które otrzymały one wykorzystując jeden z bardziej popularnych algorytmów klasyfikujących zdjęcia.



Seasoning
Spice
Spice rack
Ingredient

Product
Yellow
Drink
Bottle

Oba z tych zdjęć [5] przedstawiają przyprawy, jednak jedno z nich wykonane zostało w domu w USA, drugie w domu na Filipinach. Widać tutaj zauważalny bias i problem z rozpoznaniem przypraw na drugim zdjęciu [5]. Czy jednak taki bias stanowi problem? Jeśli algorytm ma być używany na Filipinach to raczej na pewno, jeśli jednak jego zadaniem było używanie w USA lub Europie to taki bias nie stanowi sporego problemu. Zwłaszcza jeśli weźmiemy pod uwagę, że próba pozbycia się tego biasu mogłaby zaburzyć poprawne rozpoznawanie obiektów w kategoriach napoju.

Innym przykładem, na który możemy spojrzeć są policyjne systemy AI, używane do predykcji czy dana osoba jest bardziej lub mniej skłonna do ponownego popełnienia przestępstwa po opuszczeniu zakładu karnego. W niedawnym czasie zrobiło się głośno wobec tych systemów z USA i tego, że rzekomo dyskryminują osoby czarnoskóre, ponieważ częściej stwierdzają, że są oni skłonni do recydywy [6]. Nie zamierzam tutaj zajmować żadnego stanowiska w tej sprawie, chciałbym tylko wskazać problem. Systemy te bazują na policyjnych statystykach, z których wynika, że to osoby czarnoskóre częściej popełniają przestępstwa i są skazywani w USA [7]. Jeśli więc historyczne dane statystyczne pokazują taki trend, można zrozumieć, dlaczego algorytmy AI częściej wskazują osoby czarnoskóre jako potencjalnie skłonne do recydywy. Pytanie, które należy zadać to czy na pewno należy na te systemy wpływać, by spowodować, aby ten bias zniknął. I czy stosowanie tego typu systemów w ogóle powinno mieć miejsce.

Powyżej przytoczone przykłady pokazują, że znamy powody, dlaczego niektóre algorytmy nie są sprawiedliwe, mamy też narzędzia by wpływać na już gotowe algorytmy, które przedstawiają jakiś bias. Problemem natomiast jest zdefiniowanie co jest biasem, a co nie i czy na pewno każdy bias powinien być z algorytmów usuwany. Dlatego dopóki te kwestie nie zostaną rozwiązane, to mimo że mamy techniczne możliwości, uważam, że stworzenie całkowicie sprawiedliwych algorytmów nie jest możliwe.

Źródła

[1] Feuerriegel, Stefan & Dolata, Mateusz & Schwabe, Gerhard. 2020. Fair AI: Challenges and Opportunities. Business & Information Systems Engineering. 62. 10.1007/s12599-020-00650-3.
[2] Joyce Xu. 2019. How to Fix AI: Solutions to ML Bias (And Why They Don't Matter). Strange Loop 2019.
[3] Bender, E. M., & Friedman, B. 2018. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science.
[4] Mary Reagan. 2021. AI Explained: Understanding Bias and Fairness in AI Systems. Fiddler AI. https://blog.fiddler.ai/2021/03/ai-explained-understanding-bias-and-fairness-in-ai-systems/
[5] Alexander Amini, and Ava Soleimany. 6.S191 Introduction to Deep Learning. January IAP 2020. Massachusetts Institute of Technology: MIT OpenCourseWare, https://ocw.mit.edu. License: Creative Commons BY-NC-SA.
[6] Karen Hao. 2019. AI is sending people to jail - and getting it wrong. https://www.technologyreview.com/2019/01/21/137783/algorithms-criminal-justice-ai/
[7] FBI.gov. 2020. Expanded Homicide Data Table 3: Murder Offenders by Age, Sex, Race, and Ethnicity, 2019. https://ucr.fbi.gov/crime-in-the-u.s/2019/crime-in-the-u.s.-2019/tables/expanded-homicide-data-table-3.xls