UDK 316.647.8:159.95]:007.52
DOI 10.7251/ZRSNG2324048S

Алгоритамска пристрасност: Рефлекција постојеће друштене пристрасности и изазови

Algorithmic Bias: Reflection on Existing Social Biases and Challenges

Љубинко Стојановић, Универзитет Синергија, Бијељина; Славица Савић, Факултет техничких наука, Косовска Митровица

***Сажетак* — Овај научни рад истражује успон вјештачке интелигенције (ВИ) и машинског учења који је довео до незабиљеженог напретка у различитим областима. Међутим, ове технологије нису имуне на пристрасности које постоје у нашем друштву. Термин алгоритамска пристрасност се односи на системске и систематске грешке које се могу јавити у процесима и одлукама које алгоритми доносе. Ове пристрасности настављају због различитих фактора као што су подаци које се користе за обуку алгоритама, изглед и функционалност алгоритма и људи укључени у процес. Алгоритми значајан утицај имају у друштвеним секторима гдје несвјесно уводе дискриминацију и пристрасности у обраду података, која као резултат има негативне посљедице за појединца или групе. Овај рад анализира постојеће друштвене пристрасности, изазове који настају због ових пристрасности и напоре који се улажу да се оне ублаже.**

***Кључне ријечи* – Алгоритам, грешке, подаци, пристрасност, друштвене предрасуде.**

Abstract – This scientific paper explores the rise of artificial intelligence (AI) and machine learning, which has led to unprecedented progress in various fields. However, these technologies are not immune to the biases that exist in our society. The term algorithmic bias refers to systematic and systemic errors that can occur in the processes and decisions made by algorithms. These biases persist due to various factors such as the data used to train algorithms, the appearance and functionality of algorithms, and the individuals involved in the process. Algorithms have a significant impact in societal sectors where they inadvertently introduce discrimination and biases into data processing, resulting in negative consequences for individuals or groups. This paper analyzes existing societal biases, the challenges arising from these biases, and the efforts made to mitigate them.

***Keywords* – Algorithm, errors, data, bias, social prejudice.**

I. Увод

Употреба алгоритама и машинског учења постала је све популарнија у различитим областима попут финансија, здравства, образовања, па чак и кривичног правосуђа. Ове технологије имају потенцијал да побољшају ефикасност, смање трошкове и побољшају доношење одлука. Међутим, алгоритми који се користе у овим областима нису имуни на пристрасности које постоје у нашем друштву.

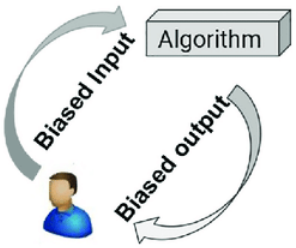
Алгоритамска пристрасност се односи на системске и систематске грешке које се не могу јавити у процесима и одлукама које алгоритми доносе. Ове пристрасности могу имати далекосежне посљедице, доводећи до неправедног третмана одређених појединаца или група и одржавајући постојеће друштвене неједнакости.

II. Постојеће друштвене предрасуде

Алгоритамска пристрасност је често одраз друшвених предрасуда које постоје у нашем друштву. Према Barocas и Selbst „подаци који се користе за обуку система машинског учења су често пристрасни, одржавајући друштвену динамику и неједнакости друштва у којем се производе [1]“. На пример, подаци који се користе за обуку алгоритама могу бити пристрасни према одређеним групама, што доводи до неправедног третмана других. У кривичном правосуђу, на пример, подаци који се користе за обуку алгоритама могу бити пристрасни према одређеним расним групама, што доводи до неправедног третмана појединца из тих група. Слично томе, алгоритми који се користе при запошљавању могу бити пристрасни

Слика 1. Еволуција пристрасности између алгоритма и човека [2]. На слици под (а), пристрасни подаци од човјека могу довести до пристрасног алгоритма – према одређеном образовању или радном искуству, што доводи до искључења квалификованих кандидата из маргинализованих заједница.

предалгоритамска пристрасност. На слици под (b), пристрасни алгоритамски излаз може утицати на људско понашање: на пример,



(c)

На слици под (c), се одвија континуирана интеракција између алгоритма који пристрасност коју називамо итерираном пристрашношћу, односно пристрасношћу која је резултат поновљене интеракције између људи и алгоритама.

скривањем одређених ставки од људи, алгоритми могу дугорочно утицати на људско мишљење, учење и свијест.

Слика 1.1 Еволуција пристрасности између алгоритма и човека [2].

III. Утицаји на различите секторе

Друштвене науке имају становиште да је дискриминација „категоризација различитих друштвених група са различитим позицијама [3]“ која може бити коришћена као оправдање за лош третман особе. Прави се разлика између директне дискриминације, тј. Третмана особе који директно зависи о њеним карактеристикама (пол, године, изглед и сл.), и индиректне дискриминације, тј. третмана особе која није директно повезан са њеним карактеристикама већ је повезан са корелацијом са њима. Индиректну дискримиацију такође можемо назвати и систематском или ненамјерном дискриминацијом [4]. Зато што алгоритамска дискриминација може да се јави ненамјерно путем корелација са карактеристикама особе, а детектовање дискриминације може бити екрстремно тешко. Комуникација између љиди се драстично промијенила и прешла са традиционалних медија као што су лице у лице, телефон, телевизија или штампа на интернет платформе друштвених мрежа као што је Facebook, Instagram, Twitter итд. За разлику од претходних традиционалних медија, друштвене мреже на интернету контролишу инфомације које корисници виде и шаљу једним другима путем алгоритама за филтрирање. Ови алхоритми биљеже индивидуалне информације о корисничким преференцама и затим филтрирају податке које приказују према тим референцама. Као резултат тога, људи су склони излагању мишљењима с којима се већ слажу, што доводи до већ поменуте алгоритамске пристрасности. Ова појава гдје се људи дијеле у групе са супротним ставовима које се ријетко међусобно срећу – „филтер балон“ или „комора еха“ је све учесталија. Употреба алгоритамских система у процесима може довести до појачане индиректне дискриминације, а као доказ за то јесу ADM системи који могу да функционишу на основу пристрасних података [5].

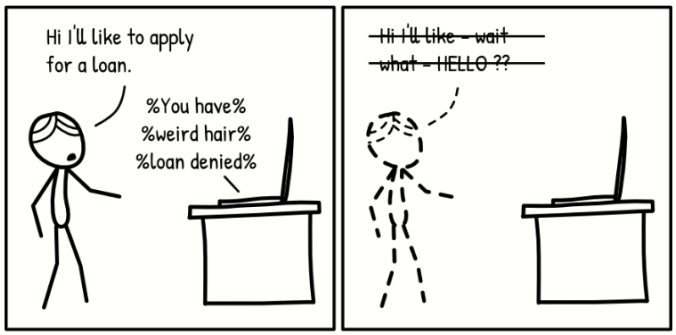
У здравственом систему љекари све више користе здравствене алгоритме базиране на математичким моделима који помажу истима да се поставе дијагнозе и донесу одлуке о лијечењу и вјештачку интелигенцију да дијагностикују болести, предложе третмане, предвиђају ризике по здравље, живот и сл. Међутим у неким случајевима кориштење истих може довести до погоршања стања пацијента због погрешног резултата због тога што се алгоритми базирају на подацима једне групе људи. VBAC алгоритам 2007. године дизајниран је у циљу процјене вјероватноће безбиједног природног порођаја након царског реза. Алгоритам узима у обзир велики број иформација попут старости жене, разлога претходног царског реза и времена који је прошао између порођаја. Међутим 2017. године, у студији која је спроведена утврђено је да алгоритам није био тачан, односно да је грешком предвиђао да ће црнкиње, хиспаноамериканске и латиноамериканке имати малу шансу за успијешан порођај након царског реза, што је довело до тога да се на истима више изводе царски резови него код бијелих жена [6].

У образовном систему алгоритми су се протеклих година почели примјењивати у великом обиму за различите примјене, аутоматско оцјењивање радова, статистички прорачуни предвиђања напуштања школовања, уписивања виших степена образова и сл. У Великој Британији 2020. године ручним алгоритмом су додјељиване оцјене на основу процјена наставника. Алгоритам је додјељивао лошије оцјене ученицима у школама финансираним од стране државе, док је приватним школама додјељивао боље оцјене [7] (чак у неким случајевима доста више од процјене наставника) [8].

Финансијски сектор за разлику од других индустрија које теже правичности као основном принципу балансира између ризика и награде. Здравствени систем, на пример, се односи према људима на основу потреба, а не на основу тога да ли могу да плате услугу. Међутим, у банкарском сектору, употреба података за разликовање добрих од лоших кредитних ризика омогућава финансијским институцијама да приближно процјене ризик који преузимају са одређеним корисником за подизање кредита, али и понекад да поставе камате засноване на ризику који преузимају. Повећана тачност у овим процесима повећава ефикасност процеса давања кредита, што доприноси повећаном конкурентношћу банке. Дискриминација се дешава када одређене приоритизоване групе добијају систематску предност, док друге групе систематски ставља у неповољан положај, а та пристрасност обично јесте произашла из свјесних или несвјесних предрасуда које уносе појединци који стварају исте алгоритме. Постоје двије ствари гдје алгоритми могу изразити пристрасности или предрасуде појединаца:

1. Алгоритми су написани од стране људи, а људи долазе са пристрасностима и предрасудама,
2. Појединац може унијети пристрасности јер користи непотпуне, нетачне или пристрасне скупове података за обуку алгоритма (врсте пристраности у наставку).

Пристрасност при узорковању се појављује када је једна популација претјерано заступљена или недовољно заступљена у скупу података за обуку. Примјер овога био би дигитална кредитна апликација гдје су мушкарци доминантнији у поређењу са женама. Уколико се подаци о корисницима користе за обуку алгоритма, алгоритам ће се више ослањати на податке мушкараца него податке жена.



Слика 2. Алгоритамска пристрасност – аплицирање за кредит [9]

Означавање, процес којим се одређене особе обиљежавају и класификују по особинама и карактеристичним тачкама како би омогућили лакше проналажење помоћу алгоритма. Као примјер можемо навести означавање занимања клијената за кредит – директор насупрот професорка умјесто радник на универзитету. Ректор и професорка би убрзо постали замјене за пол међу клијентима који су аплицирали за кредит, док би радник на универзитету склонио исту пристрасност.

V. Напори да се ублажи алгоритамска пристрасност

Учињено је неколико напора да се ублажи алгоритамска пристрасност. Један приступ јесте да се осигурају подаци који се користе за обуку алгоритама разнолики и репрезантативни за различите групе. Ово се може постићи кориштењем података из различитих извора и обезбиједивањем да подаци нису пристрасни према одређеним групама. Према O'Neil „кључ за избјегавање пристрасних података је имати различите податке и бити транспарентан о томе како су подаци прикупљени и како су очишћени [13]“. Други приступ је укључивање људи из различитих средина у изглед и развој алгоритма. Ово може помоћи у идентификацији потенционалних пристрасности и осигурати да су алгоритми праведни и инклузивни. Према Crawford „различити тимови су бољи у идентификацији потенционалних предрасуда и развоју алгоритама који су инклузивнији [12]“. Коначно, важно је редовно ревидирати алгоритме како би се осигурало да нису пристрасни према одређеним групама. Barocas и Selbst сугеришу да „алгоритми ревизије за пристрасност могу бити ефикасан начин да се идентификују и исправе грешке прије него што постану укоријењене [1]“.

VI. Закључак

Алгоритамска пристрасност је озбиљан проблем који треба да се ријеши како би се осигурало да се алгоритми и машинско учење користе на поштен и инклузиван начин. Постојеће друштвене предрасуде које доводе до алгоритамске пристрасности треба да се идентификују и адресирају кроз различите и репрезантативне податке, укључујући људе различитог поријекла и редовну ревизију алгоритама. Ублажавајући алгоритамску пристрасност, можемо осигурати да се алгоритми и машинско учење користе за побољшање ефикасности, смањење трошкова и побољшање доношења одлука на поштен и праведан начин. Учињено је неколико напора да се ублажи алгоритамска пристрасност као што је обезбеђивања података из различитих извора који нису пристрасни према одређеним групама.

IV. Изазови

Алгоритамска пристрасност представља неколико изазова. Прво, може продужити постојеће друштвене неједнакости, што доводи до неправедног третмана одређених појединаца или група. Kleinberg, Mullainathan, и Raghavan тврде да „правичност може бити у супротности са тачношћу, тако да се фокусирањем на правичност неизбјежно може жртвовати одређени степен тачности [10]“. Друго, то може довести до искључења квалификованих кандидата из маргинализованих заједница. Buolamwini и Gebru су открили да „комерцијални системи родне класификације имају веће стопе грешака за тамнопуте појединце и за жене, посебно за жене тамније пути [11]“. Треће, може нарушити повјерење у алгоритме и машинско учење, што доводи до отпора њиховој употреби у различитим областима. Crawford напомиње да је „алгоритамско доношење одлука често нетранспарентно и неодговорно, што резултира да они на које одлуке утичу немају начина да их разумију или оспоре [12]“. Четврто, то може довести до правних изазова, јер појединци из маргинализованих заједница могу оспорити одлуке које доносе алгоритми који су пристрасни против њих.

Пристрасност посредства исхода се јавља када машинско учење није добро дефинисано. На пример ако алгоритам користи адресу пребивалишта као посредство за предвиђање вјероватноће задужења за кредит. Алгоритамска пристрасност се односи на системске и систематске грешке које се не могу јавити у процесима и одлукама које алгоритми доносе.

Литература

[1] Barocas, S., & Selbst, A. D. (2016). Big Data's Disparate Impact. California Law Review, 104(3), 671-732.

[2] Sun, Wenlong & Nasraoui, Olfa & Shafto, Patrick. (2020). Evolution and impact of bias in human and machine learning algorithm interaction, 7.

[3] Kolleck, A.; Orwat, C. (2020). Mögliche Diskriminierung durch algorithmische Entscheidungssysteme und maschinelles Lernen – ein Überblick. In (Büro für Technikfolgen-Abschätzung beim Deutschen Bundestag, ed.): Hintergrundpapier Nr. 24.

[4] Orwat, C. (2020). Risk of Discrimination through the Use of Algorithms. Federal Anti-Discrimination Agency, ed.

[5] Aysolmaz B., Dau N., Iren D., (2020). Preventing Algorithmic Bias in the Development of Algorithmic Decision-Making Systems: A Delphi Study“, Proceedings of the 53rd Hawaii International Conference on System Sciences, 5268.

[6] Vyas, D. A., Jones, D. S., Meadows, A. R., Diouf, K., Nour, N. M., & Schantz-Dunn, J. (2019). Challenging the Use of Race in the Vaginal Birth after Cesarean Section Calculator. Women's health issues : official publication of the Jacobs Institute of Women's Health, 29(3), 201–204.

[7] Duncan P., McIntyre N., Storer R., Levett C. (2020). Who won and who lost: when A-levels meet the algorithm. The Guardian August 13. Доступно: https://www.theguardian.com/education/2020/aug/13/who-won-and-who-lost-when-a-levels-meet-the-algorithm.

[8] Bedingfield W. (2020). "Everything that went wrong with the botched A-Levels algorithm", Wired, 19 September. Доступно: https://www.wired.co.uk/article/alevel-exam-algorithm.

[9] Understanding Bias Part I, Machines Gone Wrong. Доступно: https://greentfrapp.github.io/project-asimov/guide/bias_i/.

[10] Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores.

[11] Buolamwini, J., & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. Proceedings of the 1st Conference on Fairness, Accountability and Transparency, 77.

[12] Crawford, K. (2016). Artificial intelligence's white guy problem. The New York Times, 25.

[13] O'Neil, C. (2016). Weapons of math destruction: How big data increases inequality and threatens democracy. Broadway Books.