Ми і є ШІ: беремо технології під контроль

Ми і є ШІ № 4: Усе про ті упередження

Обкладинка

Величне золоте дзеркало стоїть у порожній кімнаті: ми заглядаємо через плече втіленого ШІ/робота, який дивиться в нього. Відображення показує міріади людських облич: спереду троє європеоїдів, які впевнено всміхаються і сяють у віддзеркаленні; за ними — азійці та афроамериканці, чоловіки й жінки із заклеєним ротом, — ці люди відходять на задній план. Дзеркало випромінює кольорове світло, намальоване у стилі лайнарту, що веде від відображення до ШІ/робота. Відображення ШІ як людей у дзеркалі — пригноблених і наділених владою — символізує ідею, що упередженість машини лише віддзеркалює людські упередження, кодовані технологією.

Цитувати як:
Джулія Стоянович та Фала Аріф Хан. «Усе про упередженість». Ми і є ШІ. Комікси, том 4 (2021)
https://dataresponsibly.github.io/we-are-ai/comics/vol4_en.pdf

Сторінка 1
Поговорімо про те, що ми розуміємо як «упередженість» у ШІ і як вона виникає.
Уявіть собі голову робота — з виряченими очима, двома сірими антенами на голові та кнопками клавіатури замість зубів. В очах по колу написано слово bias («упередженість»).

Ми говоримо, що ШІ упереджений, якщо його використання може призвести до систематичного та несправедливого дискримінування одних осіб або груп на користь інших.

Упередженість може виникати через шкідливі стереотипи, узяті із самих даних,
або з того, як розроблений алгоритм,
або з цілей, які перед ним ставимо,
або з того, як його використовуємо.

У фундаментальній роботі 1996 року Батья Фрідман і Гелен Ніссенбаум визначили три типи упереджень, які можуть виникнути в комп'ютерних системах, Карикатурні Батья Фрідман і Гелен Ніссенбаум усміхаються до камери/читача.

представлені тут як триголовий дракон:
Раніше набуті
Технічні
Виниклі
Чорний дракон із трьома головами простягає пазурі до читача, гарчить і роззявляє пащу — крайня ліва голова повернута ліворуч, на її шиї синім кольором написано « раніше набуті». Середня голова повернута догори й зухвало гарчить, на її шиї фіолетовим кольором написано «технічні». Крайня права голова, повернута до читача, роззявила пащу й висолопила слизького зеленого язика, на шиї рожеве слово «виниклі»

[1] Батя Фрідман та Гелен Ніссенбаум. (1996). Упередженість у комп'ютерних системах.

Сторінка 2

Згадайте метафору з випіканням, яку ми використовували, щоб зрозуміти алгоритми на основі даних у першому томі.

Застосуймо цю саму метафору, щоб збагнути упередженість!

І знову наша головна героїня Мо в жовтих рукавицях дістає буханець хліба з духовки. Зовні хлібина має ідеальну золотаво-коричневу скоринку, на якій докладно випечені усміхнені голови роботів.

Раніше набуті (у даних)

Раніше набуті упередження мають суспільні джерела й існують незалежно від алгоритму.

Це будуть смакові нотки, які просочаться у ваш хліб, якщо не приділите належної уваги чистоті/свіжості інгредієнтів, або якщо вирішили використати готове тісто.

Мо в продуктовій крамничці — тримаючи повний кошик, стоїть у відділі «готових до вживання продуктів» і вибирає коробку хлібної суміші.

Ці упередження існують у суспільстві й закладені «перед випіканням» з дискримінаційної системи, що лежить в основі збирання даних, наприклад від гендерних та расових стереотипів, які підхоплюють мовні моделі, коли їх навчають на даних із соціальних мереж.

Голова робота супиться до глядача. Його брови підведені від гніву, а очі червоні та люті. У зіницях бачимо обличчя людини, яка кричить. Позаду робота стоїть той самий чоловік із насупленими від гніву бровами, настовбурченим волоссям і ротом, розтуленим на півслові. З його рота з'являються маленькі бульбашки, які потрапляють у потилицю робота.

Бульбашки мають логотипи/символи популярних соціальних медіаплатформ: векторизовану камеру та пташку, що щебече.

Сторінка 3
Технічні (у технічній системі)

Технічні упередження вносить сама система — через те, як вона розроблена або працює.

Це ті дефекти, які просочаться у ваш хліб, якщо використовуєте неправильне обладнання, ідеться, наприклад про нерівномірне пропікання кексів, якщо температура духовки неправильно відкалібрована, або розтікання тіста, якщо ваше обладнання для випікання не годиться за розміром.

Два знімки з пекарських фіаско Мо. Ліворуч зображено переповнену форму для випікання, де Мо явно переборщила з тістом, що призвело до його розтікання в духовці. Праворуч — знімок деяких цікавих кексів: усі кекси мають розрив у верхній частині й піднялися асиметрично — з одного боку сплющені та підгорілі, а з другого — на диво пухкенькі й припідняті. Тісто всередині розриву не пропеклося!

Повернімося до комп’ютерних систем:
яскравий приклад — платформи соціальних мереж, розроблені для оптимізації взаємодії (а не для безпеки чи автентичності) які зрештою просувають роз’єднувальні статті та фейкові новини.

Чоловік сидить занурений у мрії, ліктями спираючись на стіл, а долонями підперши обличчя. Над його головою — хмара хвилястого абстрактного мистецтва, що виходить з лівої руки штучного інтелекту. ШІ-джин створює мережу голів, вивергаючи такі самі хвилясті хмари, які привертають увагу чоловіка — кожна з голів ідентична і, заплющивши очі, насупивши брови й гнівно гарчучи, кричить на іншу.

Сторінка 4

Виниклі (у зв’язку з рішеннями)

Виникла упередженість з’являється з часом, адже рішення, прийняті за допомогою системи, змінюють світ, що, своєю чергою, впливає на роботу системи в майбутньому.

Подумайте про поведінкові зміни, які відбудуться внаслідок вашого випікання:

може, ви станете таким майстром, що мимоволі зробите хліб постійною частиною свого раціону!

В одній руці Мо тримає тарілку зі свіжоспеченими булочками, а другою наминає їх за обидві щоки.

Або робитимете це так часто, що відвернете всіх навколо від думки про те, щоб з'їсти ще один шматочок!

Молода дівчина хмуриться, дивлячись на простягнуту тарілку з хлібом на заквасці, і відштовхує її рукою.

Або поміркуйте, як ваше уявлення про те, «яким має бути хліб на смак», формується під впливом популярності таких продуктів, як «Диво-хліб».

Рум’янощокий хлопчик із темним кучерявим волоссям усміхається до читача, підносячи до рота шматочок білого хліба. Перед ним лежить хлібина, ідеально розташована, щоб показати читачеві етикетку «Диво-хліб».

У цьому самому ключі подумайте, як ваш доступ до новин та інформації загалом

формується алгоритмами, що курують соціальні стрічки з популярними та «трендовими» публікаціями.

Мо вдивляється в небо, повне даних, хештегів, прогнозів, передбачень та аналітики.

Сторінка 5
Щоб конкретизувати нашу дискусію, розгляньмо реальні приклади алгоритмічного упередження.

Візьмімо «Найм» за репрезентативну сферу, у якій алгоритми дедалі частіше використовують, щоб приймати критично важливі рішення «ефективніше».

Алгоритмічний інструмент найму / ШІ проводить співбесіду: троє нервово ввічливих претендентів сидять навпроти робота за столом і прагнуть справити враження

Однією з перших ознак того, що є привід для занепокоєння, стали результати дослідження AdFisher, здійсненого Університетом Карнегі — Меллона 2015 року [2].

Дослідники виконали експеримент, у якому створили два набори синтетичних профілів веб-користувачів, які були абсолютно однакові — з погляду демографічних характеристик, заявлених інтересів і моделей перегляду — за єдиним винятком: вказаною статтю — чоловічою або жіночою.

Уявіть собі зображення потенційного кандидата на роздільному екрані: на лівій панелі бізнесмен в елегантному костюмі з портфелем, на правій — бізнесвумен у строгому костюмі з ідентичним портфелем. Обидві людини на панелях ідентичні за кольором волосся, відтінком шкіри, відтінком костюмів, кольором і розміром/формою портфелів.

Дослідники показали, що Google набагато частіше показує оголошення про послуги кар'єрного коучингу для отримання високооплачуваних керівних посад чоловікам, ніж жінкам.

Це повертає нас до часів, коли в газетах дозволяли розміщувати оголошення про вакансії за статевою ознакою. Ця практика була заборонена у США 1964 року, але й далі панує в середовищі онлайн-оголошень.

Пізніше довели, що частково це стається через механіку самої системи таргетингу реклами як артефакту аукціону.
Це технічне упередження в дії!

[2] Жінкам рідше показують оголошення про високооплачувану роботу в Google — показує дослідження. Guardian (2015)

Сторінка 6
Перескочмо вперед у часі, а також перейдімо до наступного етапу лійки найму — добору резюме.

Утілений ШІ/робот вдивляється в кришталеву кулю, намагаючись прочитати майбутнє групи претендентів: їхні голови кружляють у

кришталевій кулі, а ШІ не на жарт зосереджений, розкинувши руки навколо неї.

Наприкінці 2018 року з’явилася інформація, що інструмент штучного інтелекту для добору персоналу Amazon, покликаний збільшити різноманітність робочої сили, насправді зробив протилежне:

система навчила себе, що кандидати-чоловіки кращі за кандидатів-жінок. Вона «штрафувала» резюме, які містили слово «жіночий», наприклад «капітан жіночого шахового клубу».

І це знизило рейтинг випускниць двох жіночих коледжів.

Результати підтвердили й посилили разючий гендерний дисбаланс у робочій силі.
Це виникле упередження в дії – HR-менеджер, якому ШІ-інструмент неодноразово пропонує такий самий тип претендента на посаду, що й той, який добре підходить,
з часом переконається, що саме так виглядає перспективний працівник.

Двоє європеоїдів позують перед читачем: той, що ліворуч, опускає великий палець униз, а той, що праворуч, — підіймає його догори. У лівого на великому пальці символ жінки — коло з маленьким хрестиком під ним, у правого — символ чоловіка — коло зі стрілкою, що вказує на північний схід, на великому пальці. Позаду них — дзеркальні відображення двох чоловіків, але замість їхніх силуетів втілені ШІ/роботи витягують ті самі великі пальці догори й донизу. Є три набори таких відображень, а на самому верху — відображення двох чоловіків, які мають однакові пози. Це символізує, що раніше набута упередженість кодується, а потім посилюється технологіями, а відтак подається як «нейтральна» людська інтуїція/судження для особи, яка приймає рішення.

У цьому прикладі ми також бачимо вже наявну упередженість: інструмент штучного інтелекту навчався на історичних даних про минулих працівників, і це були переважно чоловіки.

[3] Amazon відмовляється від секретного інструмента рекрутингу зі штучним інтелектом, який демонстрував упереджене ставлення до жінок. Reuters (2018)

Сторінка 7

Ось ще один приклад, на пізнішому етапі найму, може, тоді, коли потенційний роботодавець перевіряє претендента після співбесіди - Латанія Свіні, професорка інформатики в Гарварді, Карикатурна Латанія Свіні навмисно всміхається до читача й вказує вгору на результати свого дослідження.

показала, що пошук імен у Google, які звучать як імена афроамериканців, з більшою ймовірністю приводить до появи реклами, що вказує на колишню судимість, ніж пошук імен, які звучать як імена білих, навіть якщо контролюють, чи справді людина має судимість!

Є два набори жіночих облич: у верхньому лівому кутку — набір облич білих жінок — зі світлим волоссям і світлою шкірою. Під ними лупа, що символізує пошук, і ім'я Крістен, написане поряд. Унизу праворуч — набір облич афроамериканок — з темним кучерявим волоссям і смаглявим кольором шкіри. Під ними — символ пошуку та ім'я Латанія, написане поряд.

Це раніше набута упередженість — вияв давних расових упереджень суспільства.

[4] Расизм отруює поширення реклами в Інтернеті, — стверджує професорка Гарварду. MIT Technology Review (2013)

Сторінка 8

Подані випадки мають одну спільну рису: вони показують, що ШІ може посилювати й загострювати незаконну дискримінацію щодо меншин та історично незахищених груп.
Часто це називають «упередженістю в ШІ».

Та сама роботизована голова — з виряченими очима, двома сірими антенами та клавіатурними кнопками замість зубів — знову з’явилася в нашій бесіді про упередженість. У її очах по колу написано слово bias («упередженість»).

Чому ж складні системи, які мають на меті підвищити ефективність найму, не справляються з цим завданням і, певне, навіть погіршують ситуацію?

Звісно, проблеми упередженості під час працевлаштування не нові. Вони виявляли себе і в аналогову добу.

Наприклад, у відомому дослідженні 2004 року Маріанна Бертран та Сендгіл Муллайнатан надіслали фіктивні резюме до оголошень про пошук роботи в бостонські та чиказькі газети. [5]

Чи Емілі та Грег більш придатні для працевлаштування, ніж Лакшіка та Джамал?
Щоб маніпулювати сприйняттям раси, вони навмання зазначали в резюме імена, що звучать як імена афроамериканців або білих.

Імена білих отримують на 50 % більше дзвінків із запрошенням на співбесіди.

Карикатурні Маріанна Бертран і Сендгіл Муллайнатан дивляться на читача, вказуючи на велику рамку над ними. У ній написано: «Чи є Емілі та Грег більш придатні для працевлаштування, ніж Лакіша та Джамал?» Ліворуч від тексту — обличчя білого чоловіка та жінки. Праворуч — обличчя афроамериканця та афроамериканки.

Цей випадок показує, що упередженість може бути зумовлена людськими рішеннями.

[5] Чи є Емілі та Грег більш працездатні, ніж Лакіша та Джамал? Польовий експеримент з дискримінації на ринку праці. Маріанна Бертран та Сендгіл Муллайнатан (2003)

Сторінка 9
Повернімося до раніше набутих упереджень, які часто виявляють себе в даних.

Дані — це образ світу, його дзеркальне відображення.

Думаючи про упередженість даних, ставимо під сумнів цю рефлексію.

Одна з інтерпретацій «упередженості даних» полягає в тому, що відображення спотворюється — ми можемо систематично перепредставляти або недопредставляти певні частини світу в даних або інакше спотворювати показання.

Жінка сидить за ноутбуком, і ми бачимо її через плече: Перед нею електронна таблиця, а з ноутбука з'являються фрагменти даних. Дані об'єднані в круглу фігуру, яка лежить на рожевому дзеркалі. Кругла фігура — це піксельне відображення, а над нею ми бачимо первісний образ світу — круглу земну кулю, намальовану у стилі абстрактного лайнарту.

Згадаймо провал рекрутингового штучного інтелекту Amazon, який не зміг поліпшити різноманітність робочої сили.

Цей інструмент навчався на історичних даних: резюме людей, яких наймали на роботу в минулому.

Таке навчання було упереджене.

Ці дані свідчать про недостатнє представництво жінок серед робочої сили й на технічних посадах.

Тонший момент — спотворення.

Коли розглядаємо такі характеристики, як результат людини на стандартизованому тесті, чи приймаємо їх за чисту монету?

Чи зважаємо на відмінності в доступі до освітніх можливостей, наприклад чи може людина піти у кращу школу або займатися з платними репетиторами?

Сторінка 10

Інша інтерпретація «упередженості даних» полягає в тому, що, навіть якби ми змогли ідеально відобразити світ у даних, це однаково було б

відображення світу таким, яким він є, і не обов'язково таким, як міг би бути або мав би бути.

Важливо пам'ятати, що відображення не може знати, чи воно викривлене. Дані самі собою не можуть сказати нам, чи вони викривлено відображають досконалий світ, чи досконало відображають викривлений світ, чи ці викривлення доповнюють одне одного.

Жінка дивиться на те, що виглядає як портрет / мистецька інсталяція. Це тривимірна півсфера із зелених і синіх пікселів, встановлену на рожевій сітці. Позаду й поза поглядом жінки — глобус, намальований у стилі абстрактного лайнарту. Половина глобуса — та, що звернена до жінки, намальована синім і зеленим кольорами, тоді як інша половина виконана чорним кольором поверх зелених і синіх ліній. Посередині глобуса — великий червоний знак питання.

Другий момент полягає в тому, що не від даних чи алгоритмів, а від людей — окремих осіб, груп та суспільства загалом — залежить консенсус щодо того, чи світ такий, яким має бути, чи його потрібно поліпшити.

І якщо так, то як можемо його поліпшити.

Ми бачимо силует групи людей, які дивляться на щось схоже на мистецьку інсталяцію — 4 різні зображення глобуса. Зліва направо:

реалістичне тривимірне зображення різних рельєфів і географічних елементів Землі, потім абстрактний малюнок із використанням лише концентричних багатокутників у яскравих фіолетових і рожевих кольорах, потім ескіз, складений із абстрактних ліній у синьо-зелених тонах, і, нарешті, ще абстрактніша композиція з різноколірних багатокутників.

Сторінка 11
Останній момент тут полягає в тому, що зміна відображення не може змінити світ.
Якщо саме відображення використовують, щоб приймати важливі рішення,
наприклад, кого найняти або яку зарплату запропонувати людині, що наймається на роботу, тоді можна компенсувати спотворення.

Однак метафора із дзеркалом веде нас лише так далеко.

Ми повинні працювати набагато більше — зазвичай виходячи далеко за межі технологічних рішень, — щоб досягти тривалих змін у світі, а не просто почистити відображення.

Знову уявіть тривимірну півсферу із зелених і синіх пікселів, розташовану на рожевій сітці. Позаду неї, праворуч, земна куля, намальована у стилі абстрактного лайнарту.

Половина глобуса — та, що ближче до півсфери, намальована синьо-зеленою лінією, залитою чорним кольором, тоді як інша половина намальована білим кольором. Жінка тримає в одній руці мийний засіб, а у другій — ганчірку для посуду, і чистить півсферу (відображення) абстрактного глобуса (світу). Чоловік стоїть поруч, тримає блокнот і ручку, дивиться на півсферу й робить нотатки.

Повертаймося до триголового Дракона Упередження.

Говорячи про боротьбу з упередженістю в ШІ, ми переважно формулюємо проблему як пошук способу вбити дракона упереджень.

Триголовий дракон упереджень з’явився знову. Цього разу він дихає вогнем через дві з трьох голів. Відважний лицар бореться з драконом і простромлює мечем середню голову. Однак дві інші обливають лицаря полум'ям.

Але обговорюючи незворотний зв'язок між упередженістю людини та упередженістю машини, ми ставимо під сумнів саму природу цієї оповідки.

Зрештою, може, питання не в тому, як убити дракона і врятувати принцесу?

З вікна свого замку принцеса визирає вниз на галантного лицаря, який стоїть на голові щойно вбитого дракона Питання, яке ми має собі поставити, таке:
що ми робимо з суспільством, яке зачиняє принцес у замках?