Ми і Є ШІ № 4:
Усе про ті упередження

Умови використання

Цитувати як:

Джулія Стоянович та Фала Аріф Хан. «Усе про упередженість». Ми і є ШІ. Комікси, том 4 (2021) http://r-ai.co/comics

Поговорімо про те, що ми розуміємо як «упередженість» у ШІ і як вона виникає.

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 Батя Фрідман і Гелен Ніссенбаум визначили три типи упереджень, які можуть виникнути в комп'ютерних системах,

представлені тут як триголовий дракон:

Раніше набуті

Технічні

Виниклі

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

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

Застосуймо цю саму метафору, щоб збагнути упередженість!

Раніше набуті (у даних)

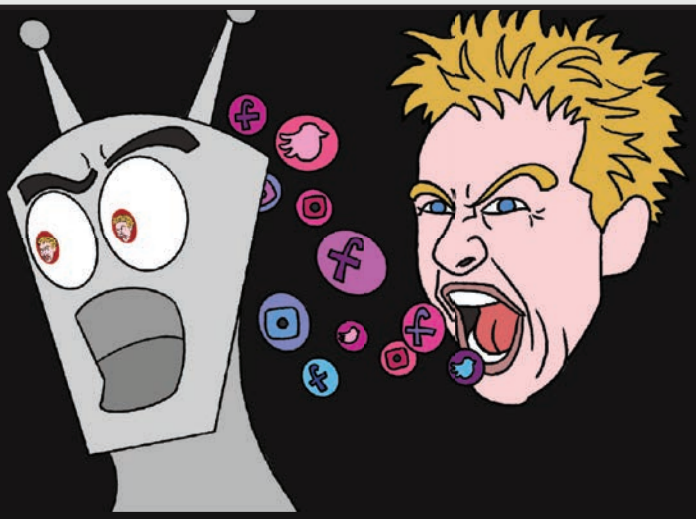
Раніше набуті упередження мають суспільні джерела й існують незалежно від алгоритму.

Це будуть смакові нотки, які просочаться у ваш хліб, якщо не приділите належної уваги чистоті/свіжості інгредієнтів,

або якщо вирішили використати готове тісто.

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.

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.

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.

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!

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

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

Щоб конкретизувати нашу дискусію, розгляньмо реальні приклади алгоритмічного упередження.

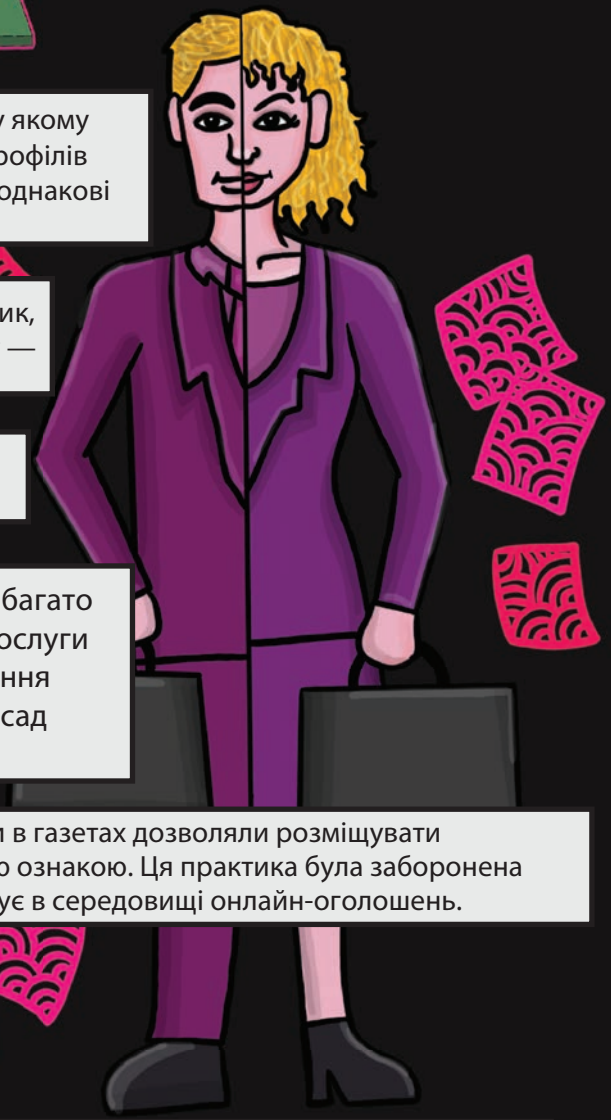
Візьмімо «Найм» за репрезентативну сферу, у якій алгоритми дедалі частіше використовують, щоб приймати критично важливі рішення «ефективніше».

Однією з перших ознак того, що є привід для занепокоєння, стали результати дослідження AdFisher, здійсненого Університетом Карнеґі — 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.



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

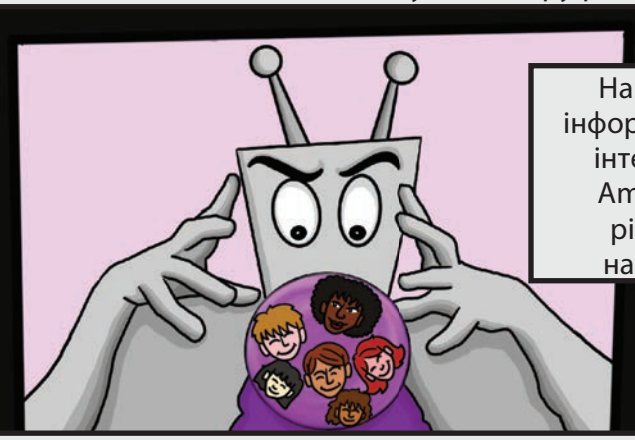
Це повертає нас до часів, коли в газетах дозволяли розміщувати оголошення про вакансії за статевою ознакою. Ця практика була заборонена у США 1964 року, але й далі панує в середовищі онлайн-оголошень.

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)

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



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.

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

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

Ось ще один приклад, на пізнішому етапі найму, може, тоді, коли потенційний роботодавець перевіряє претендента після співбесіди

Латанія Свіні, професорка інформатики в Гарварді,

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!

Це раніше набута упередженість —

вияв давних расових упереджень суспільства.

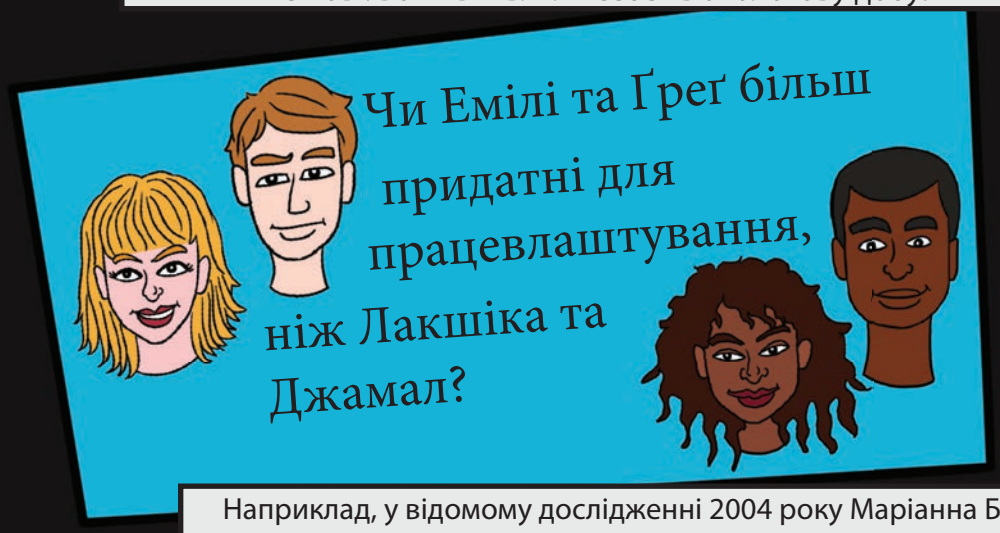
[4] Расизм отруює поширення реклами в Інтернеті, — стверджує професорка Гарварду. MIT Technology Review (2013).

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».

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].

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.

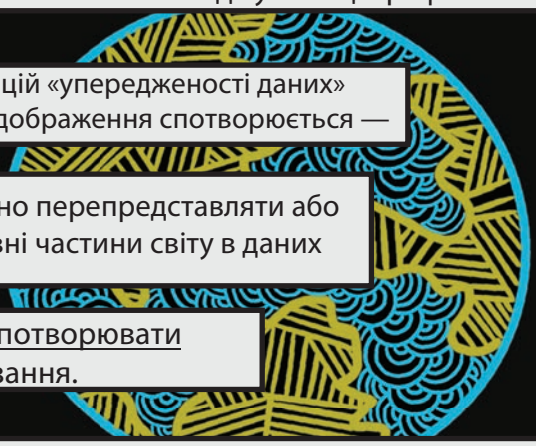
Цей випадок показує, що упередженість може бути зумовлена людськими рішеннями.

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

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.

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?

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,

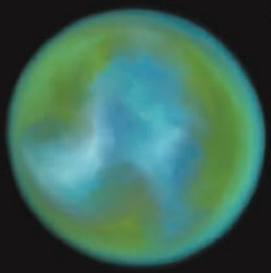
і не обов'язково таким, як міг би бути або мав би бути.

Важливо пам'ятати, що відображення не може знати, чи воно викривлене.

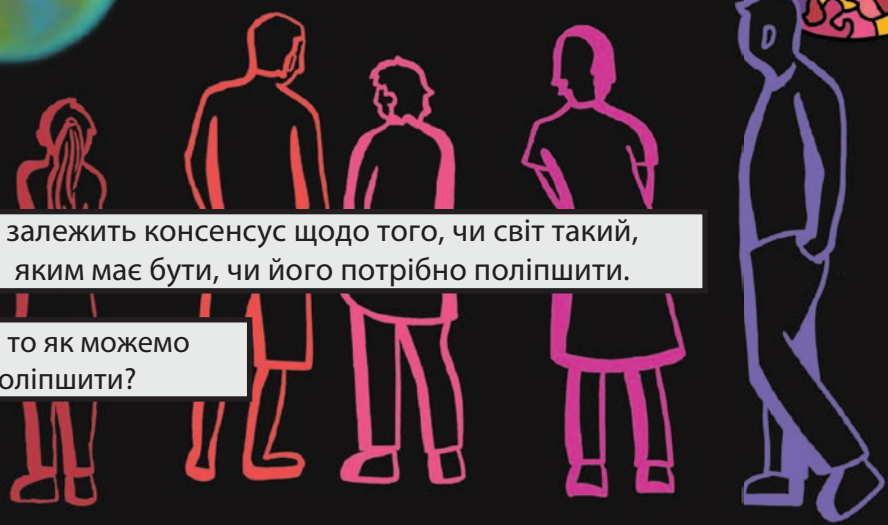
Дані самі собою не можуть сказати нам, чи вони викривлено відображають досконалий світ, чи досконало відображають викривлений світ,

чи ці викривлення доповнюють одне одного.

Другий момент полягає в тому, що не від даних чи алгоритмів, а від людей



— окремих осіб, груп та суспільства загалом —

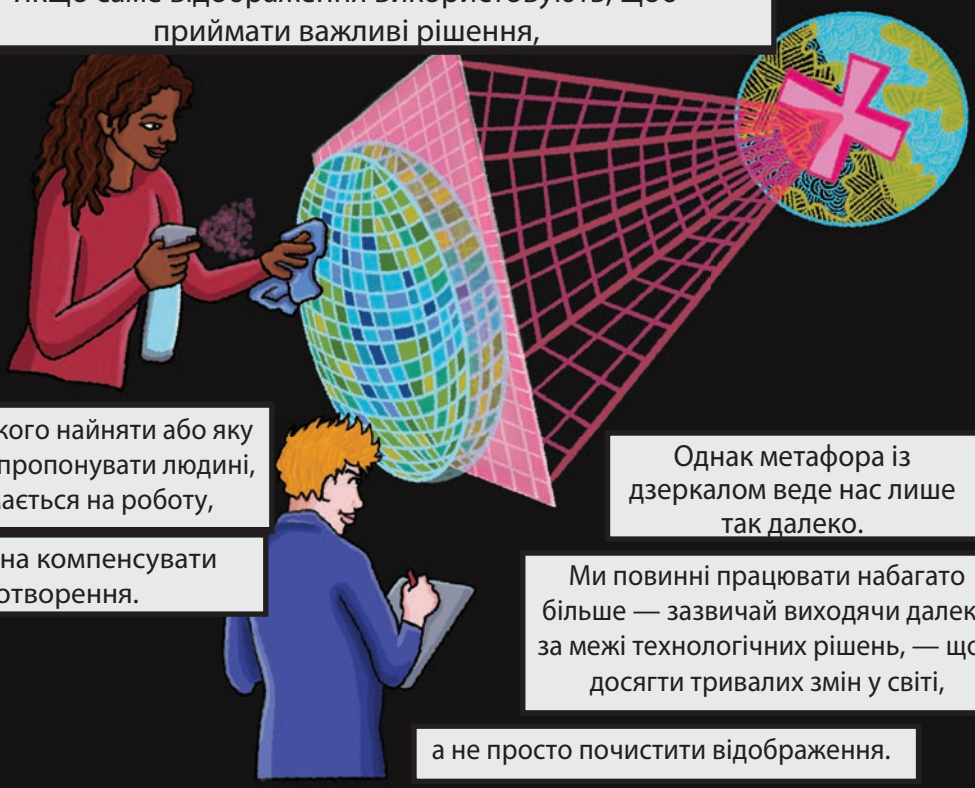


залежить консенсус щодо того, чи світ такий, яким має бути, чи його потрібно поліпшити.

І якщо так, то як можемо його поліпшити?

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.

Повертаймося до триголового Дракона Упередження.

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

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?

Питання, яке ми має собі поставити, таке:

що ми робимо з суспільством, яке зачиняє принцес у замках?