Ми і є ШІ: Беремо технології під контроль

Ми і є ШІ № 3: Хто живе, а хто вмирає — хто визначає?

Обкладинка

Величезна золота консервна банка стоїть відкрита — кришка банки піднята догори, і на ній у відображенні ми бачимо класичну проблему вагонетки: вагонетка наближається до колії, яка розгалужується на два шляхи. На одному з них лежить одна людина, прив'язана до рейок, а на іншому — група з чотирьох людей, прив'язаних до рейок. Поряд стоїть чоловік, який керує важелем, визначаючи, яким шляхом поїде вагонетка. Банка переповнена рожевими хробаками, деякі з яких упали на підлогу. На хробаках є абстрактні візерунки і слова «нерівність», «цінності», «утилітаризм», «невизначеність» і «етика». Це символізує ідею про те, що проблема вагонетки відкриває скриньку Пандори етичних питань в штучному інтелекті.

Цитувати за:
Джулія Стоянович та Фала Аріф Хан. «Хто живе, а хто вмирає — хто вирішує?». Ми і є ШІ
Комікси, том 3 (2021)
https://dataresponsibly.github.io/we-are- ai/comics/vol3_en.pdf

Сторінка 1

Прогнозувати складно, особливо майбутнє.

Прогнозування майбутнього, хоч яке воно нелегке — через невизначеність і багатогранність світу, — часто стає роботою ШІ.

Розмахуючи руками, ШІ-віщун сидить перед кришталевою кулею: вираз його обличчя суворий, а брови задерті.

Це завдання складне, а іноді навіть неможливе, тож системи штучного інтелекту припускатимуться помилок.

Наприклад, ШІ смартсвітла може неправильно вгадати, має бути воно вимкненим чи увімкненим.

Увімкнення світла посеред ночі розбудить вас.

Сплячу жінку зненацька пробуджує сліпуче світло — бачимо очі як у зомбі, розкинуті руки й розпатлане волосся — вона гнівається.

Якщо довіряти його систематичності, вам надходитимуть великі рахунки за електроенергію.

Жінка тримає в руці довжелезний рахунок за електроенергію й вигукує, побачивши суму. Жінка тримається за голову й здається засмученою.

Інший приклад: штучний інтелект для обслуговування клієнтів у вашому улюбленому взуттєвому магазині може неправильно зрозуміти замовлення, ...і вам надішлють не ту пару взуття.

Ми бачимо розмову між клієнткою та ШІ з клієнтської підтримки. ШІ має навушники й мовить у мікрофон, а насуплена жінка тримає в руках розфарбовані туфлі з візерунками.

Таке дратує, але це помилки з низькими ставками.

Наслідки таких помилок несуттєві й оборотні.

Сторінка 2

Однак трапляється, що помилки призводять до катастрофічної, незворотної шкоди,
навіть до людських жертв.

Розгляньмо безпілотний автомобіль — Жінка відпочиває в безпілотному автомобілі: вона відкинулася на водійському сидінні, поклавши ноги на приладову панель, аж раптом засвічується кермо й машина переходить у режим самостійного руху.

ШІ збирається перетнути перехрестя і не розпізнає людини на велосипеді як один із типів об’єктів, які очікує побачити на дорозі.

Тоді автомобіль продовжить рух, переїхавши велосипедиста.

Ми бачимо погляд водія безпілотного автомобіля. Навпроти на вулиці стоять два автомобілі, на тротуарі — дерева, а попереду — велосипедистка, яка перетинає дорогу.

Алгоритм сегментації зображень ідентифікує різні об’єкти на місцевості: він виявив дерево та автомобілі, позначив їх позиційними квадратиками. Він не помітив велосипедистку як об’єкт на дорозі.

Інший приклад: безпілотний автомобіль не виявляє людини, яка перетинає перехрестя в інвалідному візку.

Це може статися, якщо, наприклад, людина перетинає перехрестя задкуючи, і штучний інтелект самокеровного автомобіля хибно розраховує траєкторію руху пішохода.

Жінка у візку перетинає перехрестя. На світлофорі — червоне світло, тож вона розвертає візок і перетинає перехрестя задом наперед.

Однак люди-водії теж спричиняють аварії!

Тож навіщо віддавати добре на поталу досконалого?

Карикатурний Ілон Маск знизує плечима і вказує праворуч від екрана, аргументуючи потребу в безпілотних автомобілях.

Чи не маємо ми стерпіти кілька помилок безпілотних авто, щоб поліпшити загальну безпеку нашої транспортної системи та зручність водіїв?

Сторінка 3

Насправді чи не можемо ми закодувати судження про те, яких помилок важливіше уникати, і дозволити штучному інтелекту знаходити компроміси?

Чи не можемо наділити наш ШІ цінностями?
Відомий приклад, який змушує замислитися про наші цінності та породжувані ними компроміси, — проблема вагонетки.

Цей уявний експеримент порушує етичну дилему:
чи можемо пожертвувати життям однієї людини, щоб урятувати життя багатьох?

Цікаво, що однозначної відповіді немає, як показали експерименти в етиці та психології.

Наше рішення залежить від наших цінностей — від того, що вважаємо правильним чи неправильним, від різних елементів нашої ідентичності та культурного досвіду, а також від конкретного формулювання проблеми — контексту, у якому приймаємо рішення.

Уявіть велику, на всю сторінку, мережу вагонеток і шляхів (згори вниз): вагонетка під'їжджає до перехрестя. На верхній колії лежить одна людина, прив'язана до рейок, а на нижній — чотири особи. На перехресті стоїть чоловік, який керує важелем, скеровуючи вагонетку на певну колію. Нижня

колія веде до іншої колії з аналогічними верхньою й нижньою коліями, до яких прив’язані одна людина і група людей. Кожна колія й далі розділена на інші, демонструючи взаємопов’язану природу цієї дилеми, а також сукупний вплив індивідуальних рішень.

Цікаво, що проблема вагонетки досі залишається уявним експериментом, і її критикують як обурливу й нереалістичну.

Сторінка 4

Однак безпілотні автомобілі зараз створюють для нас життєву версію цієї дилеми.

Якщо вирішимо широко впроваджувати безпілотні автомобілі, то як працюватимемо з неминучими помилками, навіть якщо таких помилок буде відносно мало?
...А як щодо всієї транспортної системи, яку утворюють безпілотні автомобілі, люди, погода й різні дорожні умови?

Як нам одночасно розв’язувати сотні взаємозалежних проблем вагонетки?

Важлива додаткова складність — те, що, на відміну від класичної проблеми вагонетки, де відомо, скільки людей на якій колії, безпілотний автомобіль та інші типи технологій працюють в умовах високого ступеня невизначеності.

Може бути невідомо, чи є взагалі люди на дорозі, не кажучи вже про те, скільки їх там і які групи вони можуть представляти.

Як ми робимо оцінні судження в умовах невизначеності?
Уявіть варіант класичної проблеми вагонетки в умовах неповної інформації: вагонетка під'їжджає до роздоріжжя. На перехресті стоїть чоловік і керує важелем, що відповідає за наступну колію, якою рушить вагонетка. Чоловік підозрює, що до рейок прив’язані люди, а отже, людські життя будуть втрачені, якщо вагонетка поїде далі, але чоловік не певен, скільки саме людей, хто вони тощо.

Сторінка 5

Проблема вагонетки ілюструє конкретний напрям моральної філософії — Утилітаризм.

Може, ця доктрина запропонує нам якісь вказівки?

Утилітаризм — це моральний принцип, який стверджує, що правильний спосіб дій (за будь-яких умов) той, що забезпечує найбільший баланс між вигодами та шкодою всім, кого це стосується.

Утилітаризм походить від Джеремі Бентама та Джона Стюарта Мілла — англійських філософів та економістів кінця XVIII — XIX ст.

Карикатурний Джеремі Бентам позує перед камерою/глядачем.

Бентам відомий висловом: «Найбільше щастя найбільшої кількості людей — мірило добра та зла».

Ніби чудово!

Утілений ШІ/робот «осипає» людей щастям: багато людей танцюють/стрибають під час святкування, а над ними стоїть масивний робот і розкидає частинки емодзі «усміхасика» та емодзі «любові» у стилі інтернет-сенсації Salt Bae.

На жаль, застосування цих ідей до безпілотних автомобілів — і до розроблення та експлуатування технологій загалом — відкриває банку з хробаками.

Її ім’я:
Алгоритмічна мораль

Відкрита банка лежить на боці, її вміст висипається. Кришка банки відкрита праворуч, і в ній відображено проблему вагонетки. У банці черв’ячки, які вивалюються з банки.

Сторінка 6

Алгоритмічна мораль — це акт приписування моральних суджень алгоритмам.

Дві жінки стоять по обидва боки від втіленого ШІ/робота. Та, що ліворуч, одягнена з голови до п’ят у червоному, як чорт — з вилами в руці, рогами на голові. Та, що праворуч, повністю вбрана в біле, як янгол — із німбом та крилами. Обидві жінки жестикулюють і розмовляють зі штучним інтелектом. ШІ спантеличений: підвів брови, стиснув зуби.
Реалізація проблемна. І ось чому.

Спочатку з’ясуймо, як вимірюємо щастя і нещастя?

Жінка широко всміхається в камеру/глядачеві. Пара рук за допомогою рулетки вимірює ширину її усмішки як показник рівня щастя.

І як нам потім закодувати ці вимірювання в набір цілей, які зрозуміє алгоритм?

Математична формула або логічне твердження, яке може відобразити баланс між користю і шкодою, — це щось рідкісне.

Науковиця несамовито пише формули на дошці. Жінка в окулярах і тримає в руках маркер, яким виводить формули для щастя, смутку, кількості користі й рівня шкоди.
Інакше кажучи, формули «добре» чи «погано» просто не існує.

Не існує формули для цінностей і для того, як цінності виникають і змінюються в складних соціальних ситуаціях.

Ще одна причина, чому алгоритмічна мораль проблематична, полягає в тому, що
коли стається помилка в судженні про те, що правильне чи неправильне,

— а, як уже знаємо, помилки траплятимуться, тому що світ складний, мінливий і, мабуть, навіть непередбачуваний — алгоритмічна мораль передбачає, щоб алгоритм брав на себе відповідальність за помилку.

На фотокартках втіленого ШІ/робота ми бачимо його у двох різних позах із чорною дошкою в руках: одна — обличчям до глядача/камери, друга — збоку, він винувато дивиться на глядача.

Сторінка 7

Однак покладати на алгоритм відповідальність за помилку немає сенсу:

Утілений ШІ/робот ляльководить людиною: він тримає людську ляльку в лівій руці й смикає за ниточки, змушуючи людину махати рукою й дивитися на нього, тоді як його права рука зігнута на стегні.

Алгоритм не має свідомості чи свободи волі, він не робить навмисного вибору, який призводить до помилки, тому не може бути притягнутий до відповідальності.

Що це нам дає?

Проблема вагонетки показала нам, що ми не можемо делегувати етику машинам.
Що саме ми, люди, маємо здійснювати вибір і діяти (або вибирати не діяти) відповідно до наших цінностей та чинного законодавства.

А далі вже нам самим брати на себе відповідальність за наслідки будь-яких помилок.
Усміхнена людина тримає в руках робота-ляльку, смикаючи за ниточки, щоб робот трохи потанцював.

Ми не можемо віддати робо́ту бути людиною на аутсорс машині.

Сторінка 8

Підсумовуючи, скажемо: щоб впровадити етику в соціально-технічні системи, як-от ШІ,
ми повинні обміркувати, які цінності закладені в цих системах, хто виграє, коли системи працюють добре, і кому шкодять їхні помилки.

І ми повинні колективно взяти на себе відповідальність за рішення про баланс між користю і шкодою, щоб «найбільше щастя», яке Джеремі Бентам обіцяв найбільшій кількості людей, було доступне і для найбільшого розмаїття зацікавлених сторін.

Ця робота з колективного розуміння і узгодження компромісів — те, що вкорінює проєктування технологій у потребах людей.

Уявіть собі велике «дерево штучного інтелекту»: листя з розгалуженої мережі/графа — з різноколірними і щільно пов’язаними між собою вузлами і ребрами. Коріння охоплює велику територію під землею, де група людей тримається за нього й «заземлює» дерево.

Люди намальовані в техніці абстрактного лайнарту, і їх можна ідентифікувати лише по силуету. Вони виконані в різних кольорах.