ЗМІСТ

Лою, Ґордону, Люсі, Джорджеві та Ісаакові

Передмова

Чому ця книга? Чому зараз?

Ця книга про минуле, сучасне та майбутнє наших спроб зрозуміти й створити інтелект. Це важливо не тому, що ШІ — швидко проникає в усі сфери сучасності, але тому, що це основна технологія майбутнього. Наймогутніші уряди світу починають усвідомлювати цей факт, а найбільші світові корпорації вже якийсь час це знали. Ми не можемо передбачити, як розвиватимуться технології, чи скласти певну хронологію. Хай там як, маємо врахувати можливість того, що машини насправді набагато випередять людську здатність вирішувати в умовах реального світу. Що тоді?

Цивілізація — продукт нашого інтелекту; доступ до інтелекту, більш розвинутого, стане найвизначнішою подією в історії людства. Мета цієї книги — пояснити, чому це може виявитися ще й останньою подією, та спробувати впевнитися, що такого не станеться.

Огляд

Книга поділяється на три частини. Перша (розділи від першого до третього включно) досліджує поняття інтелекту в людей і машин. Матеріали не потребують технічної підготовки, але для зацікавлених читачів розділи доповнені чотирма додатками з поясненням концепції систем, покладених в основу сучасного штучного інтелекту. У другій частині (розділи від четвертого до шостого включно) обговорюються проблеми, що виникають, внаслідок оснащення машин інтелектом. Зокрема я зосереджуюся на проблемі контролю: як утримувати абсолютну владу над машинами, потужнішими за нас. У третій частині (розділи від сьомого до десятого включно) запропоноване нове бачення штучного інтелекту й способи досягнення впевненості в безпечній для людей роботі машин. Книга призначена для широкої аудиторії, але, сподіваюся, вона змусить фахівців з розробки штучного інтелекту переосмислити основні засади своєї діяльності.

Розділ 1

ЯКЩО НАМ ВДАСТЬСЯ

Колись давно мої батьки жили в будинку біля університету в англійському Бірмінгемі. Вони вирішили виїхати з міста й продали будинок Девідові Лоджу, професорові англійської літератури. На той час Лодж був уже відомим письменником. Я ніколи з ним не зустрічався, але вирішив почитати деякі його книги, наприклад, «Зміна місць» і «Світ тісний». Серед головних героїв були вигадані науковці, що переїжджали з вигаданого Бірмінгема до вигаданого каліфорнійського Берклі. Я був справжнім науковцем, який щойно перебрався зі справжнього Бірмінгема до справжнього Берклі, тож ніби хтось у Міністерстві збігів порадив мені звернути на це увагу.

Одна сцена з «Тісного світу» мене вразила. Головний герой, гонористий літературознавець, під час великої міжнародної конференції запитує у видатних науковців: «Що буде, коли з вами всі погодяться?». Запитання викликає переполох, бо провідні науковці

більше цікавилися інтелектуальним поєдинком, ніж з'ясуванням істини чи досягненням порозуміння. Мені здалося, що те саме можна запитати й у провідних науковців зі сфери ШІ: «А якщо нам вдасться?». Завданням усієї галузі завжди було створення штучного інтелекту людського чи надлюдського рівня, але ми всі мало розуміли чи й зовсім не розуміли, що станеться, коли нам це вдасться.

За кілька років потому ми з Пітером Норвіґом почали працювати над новим підручником зі штучного інтелекту, перша редакція якого з'явилася 1995 року[1]. Заключний розділ книги називався «А якщо нам вдасться?». У розділі вказується на можливі як позитивні, так і негативні наслідки. Але там не пропонувалися ґрунтовні висновки. До виходу третьої редакції, 2010 року, багато хто нарешті почав розмірковувати над імовірністю не надто позитивних наслідків створення надлюдського штучного інтелекту. Але більшість цих людей не були фахівцями з досліджень у галузі ШІ. До 2013 року я переконався, що це не просто нагальна, а, можливо, найважливіша проблема, з якою стикалося людство.

У листопаді 2013-го я читав лекцію в картинній галереї Далвіч, шанованому музеї мистецтв на півдні Лондона. Аудиторія складалася в основному з пенсіонерів, які не були фахівцями, та загалом цікавилися інтелектуальними подіями. Тож мені довелося уникати складних технічних термінів. Ця галерея здавалася цілком придатною для першої спроби донести мої міркування до широкого загалу. Спершу я пояснив, що таке штучний інтелект, а потім висунув п'ять «кандидатів на звання найважливішої події у майбутньому людства»:

1. Ми всі помремо (зіткнення з астероїдом, кліматична катастрофа, епідемія тощо).
2. Ми всі житимемо вічно (медичне вирішення проблеми старіння).
3. Ми винайдемо можливість подорожувати швидше за світло й підкоримо Всесвіт.
4. Нас відвідають представники більш розвиненої інопланетної цивілізації.
5. Ми винайдемо надпотужний ШІ.

Я припустив, що п'ятий «кандидат», надпотужний ШІ, переможе, бо це створить можливості для уникнення фізичних катастроф, досягнення вічного життя та подорожей зі швидкістю, що перевершує швидкість світла, якщо така взагалі можлива. Це стало б гігантським стрибком, навіть розривом нашої цивілізації. Винайдення надпотужного штучного інтелекту багато в чому подібне до прибуття представників розвиненішої інопланетної цивілізації, але набагато вірогідніше. Можливо, найважливіше те, що в ситуації з ШІ, на відміну від візиту інопланетян, ми матимемо право голосу.

Тоді я попросив аудиторію уявити, що станеться, коли ми отримаємо повідомлення від розвиненішої інопланетної цивілізації про прибуття на Землю її представників через тридцять або п'ятдесят років. Словом *сум'яття* навіть приблизно цього не описати. Тимчасом наша реакція на передбачення появи надпотужного штучного інтелекту... Ну, можна сказати, це нікого особливо не вражає. (В іншій лекції я проілюстрував це обміном повідомленнями, показаним на малюнку 1). Зрештою я пояснив значення надпотужного штучного інтелекту так: «Успіх у цій справі стане найбільшою... і, можливо, останньою подією в історії людства».

Від: Вищої інопланетної цивілізації <sac12sirius.canismajor.u>
До: humanity@UN.org
Тема: Контакт
Попереджаємо: ми прибудемо через 30–50 років
Від: humanity@UN.org
До: Вищої інопланетної цивілізації <sac12sirius.canismajor.u>
Тема: Нас немає на місці.
Відповідь: Контакт
Людства зараз немає на місці. Ми відповімо на ваше повідомлення, щойно повернемося. ☺

Малюнок 1. Мабуть, це не належний обмін повідомленнями для першого контакту з вищою інопланетною цивілізацією.

За кілька місяців, у квітні 2014 року, під час моєї участі в конференції у Ісландії, мені зателефонували з Національного громадського радіо й запитали, чи не погоджусь я дати їм інтерв'ю про фільм «Перевага», який щойно вийшов у Сполучених Штатах. Хоча я читав сюжетні анотації та відгуки, самого фільму не бачив, бо на той час жив у Парижі, де його мали показувати лише в червні. Та сталося так, що дорогою додому мені довелося заїхати в Бостон для участі в засіданні Міністерства оборони. Тож по прибутті до Бостонського аеропорту імені Логана, я взяв таксі до найближчого кінотеатру, де показували цей фільм. Сидів у другому ряду й дивився, як у професора ШІ з Берклі, якого грав Джонні Депп, стріляють активісти, стурбовані появою надпотужного штучного інтелекту. Я мимохіть засовався у кріслі. (Ще один дзвінок із Міністерства збігів?). До того, як герой Джонні Деппа помер, його свідомість завантажили на квантовий суперкомп'ютер, і цей комп'ютер невдовзі перевершив людські можливості, загрожуючи захопити світ.

19 квітня 2014 у «Гаффінгтон Пост» з'явився мій відгук на «Перевагу» в співавторстві з фізиками Максом Теґмарком, Френком Вільчеком та Стівеном Гокінгом. Там була цитата з моєї лекції про найвизначнішу подію в історії людства, прочитаної в Далвічській картинній галереї. Відтоді я публічно пов'язаний з цією точкою зору: галузь, у якій я веду дослідження, несе потенційну загрозу моєму видові.

Як ми сюди дісталися?

Коріння штучного інтелекту відходить до античності, але його «офіційний» початок датується 1956 роком. Двоє молодих математиків, Джон Маккарті та Марвін Мінський, переконали Клода Шеннона, вже відомого винахідника інформаційної теорії, та Натаніеля Рочестера, розробника першого комерційного комп'ютера IBM, долучити їх до організації літньої програми в Дартмутському коледжі. Мета цього заходу визначалася так:

> «Продовжити дослідження на основі припущень: кожен аспект навчання чи будь-яка інша особливість інтелекту в принципі можуть бути описані так точно, що машина виявиться здатною це відтворити. Спробувати знайти спосіб створення машин, які користуватимуться мовою, формуватимуть абстракції та поняття, вирішуватимуть проблеми, якими нині займаються люди, а також самовдосконалюватимуться. Ми вважаємо, що можна досягти значного просування вперед у вирішенні однієї чи кількох із цих проблем за умови ретельного добору групи науковців, яка працюватиме над ними влітку».

Не потрібно нагадувати, що це затягнулося набагато довше; ми досі працюємо над усіма цими проблемами.

У перше десятиліття після Дартмутської зустрічі в галузі штучного інтелекту вдалося досягти деяких значних успіхів, зокрема створити алгоритм загального логічного мислення, розробником якого є Алан Робінсон[2], а також програму для гри в шашки, здатну до самонавчання. Автор останньої — Артур Самюель; ця програма переграла свого творця[3]. Перша бульбашка ШІ луснула наприкінці 1960-х, коли спроби машинного навчання та машинного перекладу не виправдали сподівань. У звіті, наданому урядом Великої Британії, зазначалася: «У жодній галузі досліджень не досягнуто очікуваного значного поступу»[4]. Іншими словами, машини виявилися недостатньо розумними.

На щастя, одинадцятирічний я не читав цього звіту. За два роки, коли мені дали «Сінклер», Кембриджський запрограмований калькулятор, я просто хотів наділити його інтелектом. Однак «Сінклер» із максимальною програмою на тридцять шість клавіш виявився заслабким для ШІ рівня людини. Я палко прагнув доступу до гігантського суперкомп'ютера CDC 6600[5] в Імперському коледжі Лондона й написав програму для гри в шахи — стосик перфокарт два фути заввишки. Програма виявилася не надто вдалою, але це не мало значення. Я знав, чим хочу займатися.

До середини вісімдесятих я став професором у Берклі, а тимчасом розробка штучного інтелекту неабияк пожвавилася завдяки комерційному потенціалові так званих експертних систем. Друга бульбашка штучного інтелекту луснула, коли ці системи виявилися непри-

датними для виконання багатьох покладених на них завдань. Знову машини виявилися просто недостатньо «кмітливими». І для штучного інтелекту настала зима. Мій курс ШІ в Берклі, який нині зібрав понад 900 студентів, 1990 року налічував лише двадцять п'ять слухачів.

Спільнота розробників штучного інтелекту засвоїла урок: розумніший дорівнює кращий, але нам довелося над цим добряче попрацювати. Галузь стала набагато математичнішою. Встановлені зв'язки з такими давно усталеними дисциплінами, як теорія імовірності, статистика й теорія управління. Зерно сьогоднішнього поступу лягло в ґрунт саме тієї зими штучного інтелекту. Зокрема про це свідчать ранні роботи, присвячені широкомасштабним системам імовірнісної логіки, пізніше відомі як *глибинне навчання*.

Від 2011 року технології глибинного навчання неабияк прогресували в удосконаленні способів розпізнавання мови, візуальних об'єктів та машинного перекладу — трьох найважливіших з невирішених проблем галузі. За деякими оцінками нині машини в цих сферах зрівнялися з людьми або навіть перевищують людські можливості. 2016-го й 2017-го «AlphaGo» компанії «DeepMind» перемогла колишнього світового чемпіона з ґо Лі Седоля та нинішнього чемпіона Ке Джі. Ці події, передбачені деякими експертами ще до 2097 року, хоча дехто й сумнівався, що таке взагалі можливе[6].

На сьогодні штучний інтелект майже щодня — в перших заголовках усіх медіа. Потоки фінансів із венчурних фондів стимулювали тисячі стартапів. Мільйони студентів проходять онлайн-курси з ШІ та ма-

шинного навчання, експерти в галузі отримують зарплатню. Інвестиції надходять не лише з венчурних фондів, а й від національних урядів та корпорацій і оцінюються в десятки мільярдів доларів щороку; за останнє п'ятиріччя в галузь вкладалося більше коштів, ніж за всю її попередню історію. Приблизно вже в наступному десятилітті досягнення, які чекають на нас найближчим часом, такі як самокеровані авто й розумні особисті помічники, справлятимуть істотний вплив на світ. Потенційні економічні та соціальні вигоди від штучного інтелекту є потужним імпульсом для подальших досліджень.

Що станеться потім?

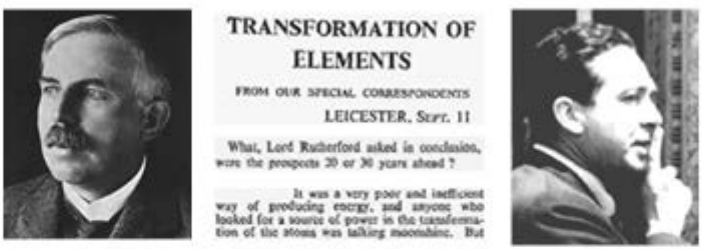
Чи означає цей стрімкий поступ, що нас невдовзі захоплять машини? Ні. Кілька проривів мають статися, перш ніж ми досягнемо чогось, бодай трохи подібного до машини з надлюдським інтелектом.

Наукові прориви відомі своєю непередбачуваністю. Щоб це собі уявити, можемо озирнутися на історію іншої галузі, що загрожує кінцем цивілізації, — ядерної фізики.

На початку двадцятого століття, можливо, жоден ядерний фізик не відзначився так, як Ернест Резерфорд, дослідник протона, «людина, яка розщепила атом» (малюнок 2[a]). Як і його колеги, Резерфорд розумів, що в атомних ядрах накопичено безліч енергії, хоча й переважала думка, що вивільнити її неможливо.

11 вересні 1933 року Британська наукова асоціація проводила щорічну зустріч у Лестері. Лорд Резерфорд звернувся до вечірнього зібрання. Як уже кілька разів

раніше, Резерфорд охолодив запал тих, хто змальовував блискучі перспективи атомної енергетики: «Усі, хто шукає джерело енергії у трансформації атомів, верзуть нісенітниці». Промова Резерфорда наступного ранку з'явилася в лондонській «Таймз» (малюнок 2[b]).



TRANSFORMATION OF ELEMENTS

FROM OUR SPECIAL CORRESPONDENTS

LEICESTER, SEPT. 11

What, Lord Rutherford asked in conclusion, were the prospects 20 or 30 years ahead?

It was a very poor and inefficient way of producing energy, and anyone who looked for a source of power in the transformation of the atoms was talking moonshine. But

(а) (б) (в)

Малюнок 2. (а) Лорд Резерфорд, ядерний фізик. (б) Уривок зі статті в «Таймз» від 11 вересня 1933 року про доповідь Резерфорда, виголошену напередодні. (в) Лео Сілард, ядерний фізик.

Тим часом Лео Сілард (малюнок 2[c]), угорський фізик, який нещодавно втік із нацистської Німеччини, зупинився в готелі «Imperial» на площі Рассел у Лондоні. За сніданком він прочитав статтю в «Таймз». Замислився над прочитаним — і під час прогулянки винайшов індуковану нейроном ядерну ланцюгову реакцію[7]. Проблема вивільнення ядерної енергії пройшла шлях від немислимої до, по суті, вирішеної менш ніж за двадцять чотири години. Впродовж наступного року Сілард оформлював таємний патент на ядерний реактор. Перший патент на ядерну зброю видали у Франції 1939 року.

Мораль цієї історії в тому, що закладатися супроти людської винахідливості — нерозумна ідея, особливо коли на кону наше майбутнє. У спільноті розробників