ЗМІСТ

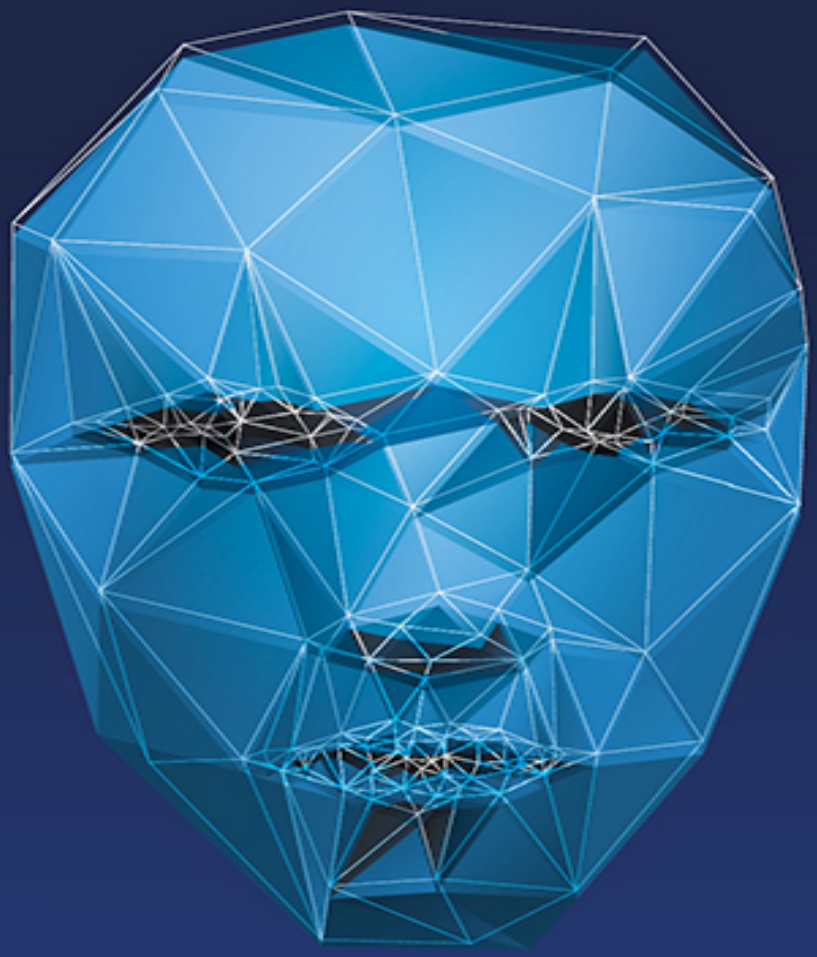
Сумісний з людиною. Штучний інтелект і проблема контролю

Про книгу

У цій книзі провідний дослідник Стюарт Рассел стверджує, що цього сценарію можна уникнути, однак ми маємо переосмислити штучний інтелект. Автор описує короткострокові вигоди, які можна очікувати, і відкриття, які ще потрібно здійснити. Рассел припускає, що ми можемо перебудувати штучний інтелект на новій основі, відповідно до якої машини будуть стриманими, альтруїстичними й прагнутимуть досягати наших цілей, а не власних.

СУМІСНИЙ З ЛЮДИНОЮ

ШТУЧНИЙ ІНТЕЛЕКТ І ПРОБЛЕМА КОНТРОЛЮ



СТЮАРТ РАССЕЛ

Анотація

У нашому уявленні надлюдський штучний інтелект — це потужне цунамі, яке загрожує не лише роботі чи людським стосункам, а й усій цивілізації. Конфлікт між людьми та машинами — неминучий, а його фінал занадто передбачуваний. У цій книзі провідний дослідник Стюарт Рассел стверджує, що цього сценарію можна уникнути, однак ми маємо переосмислити штучний інтелект. Автор описує короткострокові вигоди, які можна очікувати, і відкриття, які ще потрібно здійснити. Рассел припускає, що ми можемо перебудувати штучний інтелект на новій основі, відповідно до якої машини будуть стриманими, альтруїстичними й прагнутимуть досягати наших цілей, а не власних. Та якщо передбачені прориви справдяться і на нас очікує поява надлюдського штучного інтелекту, це створіння буде могутнішим за людину. Чи можемо ми гарантувати, що воно ніколи не матиме влади над нами?

ISBN 978-966-993-502-1

СУМІСНИЙ З ЛЮДИНОЮ. ШТУЧНИЙ ІНТЕЛЕКТ І ПРОБЛЕМА КОНТРОЛЮ

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. Наймогу-тніші 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. Тимча-сом 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 Бостонсько-го 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 «Гаффінґтон Пост» з'явився мій відгук на «Перевагу» в співавторстві з фізиками Максом Теґмарком, Френком Вільчеком та Сті-веном 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 Бер-клі, 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?

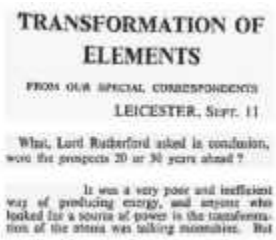
Чи означає цей стрімкий поступ, що нас невдовзі захоплять машини? Ні. Кілька проривів мають статися, перш ніж ми досягнемо чогось, бодай трохи подібного до машини з надлюдським інтелектом.

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]).

(a)



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

(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 штучного інтелекту зароджується своєрідне протиріччя. Воно зайшло навіть аж так далеко, що ставить під сумнів можливість успіху в досягненні мети ШІ в далекій перспективі. Це ніби водій автобуса, пасажирами якого є все людство, каже: «Так, я на повній швидкості мчу до урвища, але, повірте, пальне в нас скінчиться, перш ніж ми туди дістанемося!».

Я не стверджую, що ми *обов'язково* досягнемо успіху в царині штучного інтелекту. І гадаю, він доволі мало ймовірний у наступні кілька років. Та, мабуть, достатньо помірковано принаймні підготуватися до цієї події. Якщо все буде добре, на нас чекає розквіт людства, але мусимо усвідомити, що ми передбачаємо істот, набагато могутніших за людей. Як упевнитися, що вони ніколи- ніколи не намагатимуться нас підкорити?

Просто щоб уявити, з яким вогнем граємося, уявімо, як функціонують алгоритми вибору контенту в соціальних мережах. Вони не особливо розумні, але завдяки своєму

безпосередньому впливові на мільярди людей можуть змінити цілий світ. Зазвичай такі алгоритми розроблені для максимального збільшення кількості *натискань*, тобто підвищення імовірності реакції користувача на певне посилання. Рішення начебто просте: презентувати посилання, яким користувач надає перевагу, чи не так? А ось і ні. Рішення полягає в тому, щоб змінити вподобання користувача й зробити їх більш передбачуваними. Передбачуваного користувача можна «годувати» тим, на що він зазвичай натискає, а це — більший зиск. Люди з чіткіше вираженими політичним поглядами, передбачуваніші в тих посиланнях, на які натискають (Можливо, існує категорія статей, на які натиснуть затяті центристи, але непросто уявити, з кого саме ця категорія складається.) Як і кожна розумна істота, алгоритм знає, як моделювати стан свого середовища — у цьому разі свідомості користувача, — щоб отримати більшу вигоду[8]. Серед наслідків – можливе відродження фашизму, розрив суспільного договору, що лежить в основі демократичних держав усього світу та потенційний крах Європейського Союзу й НАТО. Непогано для кількох рядків коду, навіть якщо йому вряди-годи й допомагають люди. А тепер уявіть, що може зробити *справді* розумний алгоритм.

Що пішло не так?

Історія штучного інтелекту рухалася під єдину мантру: «Що розумніший, то кращий». Я переконаний: це помилка не через забобонний острах перед можливістю заміни людини штучним інтелектом, а через хиби нашого розуміння самого інтелекту.

Концепція інтелекту є базою нашої особистості, саме тому ми називаємо себе Homo sapiens або «людиною розумною». Після понад двох тисяч років самопізнання ми дійшли

характеристики інтелекту, яку можна скоротити до такого визначення:

Люди розумні тією мірою, до якої їхні дії свідчать про здійснення намірів.

Решту характеристик розуму — сприйняття, мислен -ня, навчання, вигадування тощо — можна зрозуміти через їхній внесок у нашу здатність успішно діяти. Від самого початку ШІ, інтелект машин визначався так само:

Машини розумні тією мірою, до якої їхні дії свідчать про досягнення їхніх намірів.

Машини, на відміну від людей, не мають власної мети, тож мету їм вказуємо ми. Іншими словами, конструюємо оптимальні машини, ставимо певну мету, й вони працюють над її досягненням.

Це загальний підхід не лише до штучного інтелекту. Він простежується як у технологічних, так і в розрахун -кових основах нашого суспільства. У галузі теорії управління, яка розробляє управлінські системи для всього від реактивних двигунів до інсулінових помп, завдання системи — мінімізація *втрат*, що зазвичай визначаються певними відхиленнями від бажаної поведінки. У галузі економіки, механічних процесах і політиці вони розроблені для максимальної *кори -сті* індивідів, *добробуту* груп і *зиску* корпорацій[9]. В операційних дослідженнях, під час яких розв'язуються складні логістичні та виробничі проблеми, максимізуються очікувані *суми прибутку*. Зрештою навчальні алгоритми в статистиці розроблені для зниження *очікуваних втрат*, а це визначає вартість помилок прогнозування.

Вочевидь ця загальна схема, яку називатиму *стандартною моделлю*, значно поширена й дуже потужна. На жаль, *ми не бажаємо мати машин, розумних у цьому сенсі.*

На недолік стандартної моделі вказав Норберт Вінер, легендарний професор Массачусетського технологічного інституту й один із найвидатніших математиків середини двадцятого століття. Вінер побачив програму Артура Семюеля для гри в шашки, яка навчилася грати набагато краще за своїх творців. Цей досвід спону-кав його написати пророчу, але маловідому статтю «Моральні та технічні наслідки автоматизації»[10]. Ось як він визначив свою головну думку:

«Якщо для досягнення наших намірів ми використовуємо механічні пристрої, у діяльність яких не можемо активно втручатися.., маємо бути абсолютно впевнені, що намір, який вкладаємо в цю машину, — саме те, чого ми справді бажаємо».

«Намір, закладений у машину» і є метою, яку машини оптимізують у стандартній моделі. Якщо ми закладемо хибну мету в машину, розумнішу за нас, вона досягне мети й ми програємо. Криза соціальних мереж, описана вище, — лише квіточки, результат оптиміза-ції процесу досягнення мети, хибної у глобальному масштабі з недостатньо продуманими алгоритмами. У п'ятому розділі я описую набагато гірші наслідки.

Усе це не має нікого дивувати. Тисячі років ми знаємо про ризики отримання бажаного. У кожній історії, де в когось здійснювалися три бажання, третє завжди скасовувало два попередні .

Якщо стисло, здається, поступ у напрямку створення надлюдського інтелекту зупинити неможливо, але успіх у цій справі може знищити людську расу. Однак не все втрачено. Маємо зрозуміти, де помилилися, й це виправити.

Чи можемо ми це виправити?

Проблема криється в самому визначенні штучного інтелекту. Ми говоримо, що машини розумні, бо здатні до здійснення *своїх* намірів, але не маємо надійного способу впевнитися, що *їхні* наміри збігаються з *нашими*.

А якщо замість дозволу машинам на досягати *їхньої* мети, наполягатимемо на досягненні *нашої* ? Така машина, якщо її можливо розробити, виявиться не лише *розумною*, а й *корисною* людям. Тож спробуймо отак.

Машини корисні, коли виконують власні дії для здійснення наших намірів.

Напевно, саме так ми вже давно мали це сформулювати.

Звісно, складність у тому, що наміри маємо ми (всі вісім мільярдів у всьому нашому дивовижному розмаїтті), а не машини. І все-таки можливо побудувати машини, корисні саме в цьому сенсі. Вони неминуче виявляться не впевненими в наших намірах (зрештою ми й самі не завжди в них упевнені), але це перевага, а не вада (так, це справді добре). Невпевненість щодо наших намірів передбачає звернення машин до людей: вони проситимуть дозволу, прийматимуть виправлення й дозволятимуть себе вимикати.

Аби позбавити машини можливості визначати власні наміри, нам доведеться видаляти й замінювати частину того, на чому базується штучний інтелект, — конкретні визначення потрібних нам операцій. Це також означатиме зміни значної частини надбудови — купи ідей і методик створення штучного інтелекту. Результатом стануть нові стосунки між людьми й машинами. Сподіваюся, вони дадуть нам змогу спокійно прожити кілька наступних десятиліть.

Розділ 2. Інтелект у людей та машин

Коли потрапляєш у глухий кут, найкраще простежити свій шлях і зрозуміти, де помилився. Я певен, що стандартна модель, за якої машини оптимізують фіксовану мету, встановлену людьми, — це глухий кут. Проблема не в тому, що ми *не зможемо* вдало сконструювати системи ІІІ; тут якраз *успіх* дуже можливий. Саме визначення успіху ІІІ — хибне.

Тож простежимо наш шлях до самого початку. Спробуймо збагнути витоки нашої концепції інтелекту і з'ясуймо, яким чином це застосовне до машин. Тоді матимемо шанс на ліпше визначення того, що слід вважати гарною системою штучного інтелекту.

Інтелект

Як працює Всесвіт? Із чого почалося життя? Де мої ключі? Це основні запитання, варті того, щоб над ними замислитися. Але хто запитує? Як я відповідаю? Як може жменька матерії, кілька фунтів рожевувато-сірої маси, що її ми називаємо мозком, сприймати, розуміти, передбачати й маніпулювати немислимим, неосяжнним світом? Розум уже давно сам себе вивчає.

Тисячі років ми намагаємося зрозуміти, як працює наш інтелект. Спочатку з цікавості, потім для самоконтролю, пояснення власної віри та, з доволі прагматичною метою, математичного аналізу. Хоча кожен крок до розгадування принципів роботи розуму водночас є кроком до штучного відтворення можливостей мозку й формування штучного інтелекту.

Перш ніж зможемо створити інтелект, мусимо зрозуміти, що це таке. Відповідь не в тестах IQ і навіть не в тестові Тюрінґа, а в простому взаємозв'язку нашого сприйняття з нашими бажаннями та вчинками. Спрощено це звучатиме так:

сутність розумна тією мірою, якою її діяльність допомагає в досягненні бажаного залежно від її сприйняття.

Еволюційні витоки

Уявімо нижчу бактерію, таку як E. coli. Вона має приблизно шість джгутиків — довгих, схожих на волосини щупальців. Вони обертаються навколо тіла за годинниковою стрілкою або проти неї. (Сам обертовий механізм — дивовижний, але це вже інша історія.) Коли E. coli пливе у своєму рідкому середовищі — нижньому кишковику — вона чергує обертання джгутиків за годинниковою стрілкою, що змушує її перекидатися на місці, та проти годинникової стрілки, завдяки якому джгутики крутяться, як пропелер, і бактерія пливе по прямій. Таким чином, E. coli здійснює щось подібне до вільної ходьби — пливе, перекидається, пливе, пере-кидається. Це дає їй змогу знайти й спожити глюкозу: так вона не залишиться в незмінному положенні й не загине без живлення.

Якби цим і обмежилося, ми б не сказали, що E. coli володіє чимсь подібним до інтелекту, адже її рух ніяк не залежить від середовища. Бактерія нічого не вирішує, лише виконує одні й ті самі рухи, яку еволюція заклала їй генетично. Але це ще не все. Коли E. coli відчуває підвищення концентрації глюкози, вона пливе довше й менше перевертається, і все відбувається навпаки, коли вона відчуває зменшення кількості глюкози. Тож її поведінка (просування до глюкози) допомагає їй досягти необхідного (скажімо, більше глюкози), залежно від її сприйняття (підвищення концентрації речовини).

Можливо, ви думаєте: «Але еволюція й це заклала їй генетично! Як вона може виявляти інтелект?». Це небезпечне міркування, адже еволюція і вам генетично заклала базовий проект мозку, та вочевидь на цій підставі ви не заперечуватимете власного інтелекту. Суть у тому, що еволюція генетично заклала E. coli, так само, як вам, механізм зміни поведінки залежно від зміни умов. Еволюції не відомо, де має

бути глюкоза чи де лежать ваші ключі. Тож закласти в організм здатність знайти їх — найкраще, що вона могла запропонувати натомість.

Нині E. coli далебі не вважається гігантом думки. Як ми знаємо, вона не пам'ятає, де була. Коли бактерія переміщується від точки A до точку B й не знаходить глюкози, вона, швидше за все, повернеться в точку A. Якщо ми сконструюємо середовище, де дуже привабливий градієнт глюкози вестиме до краплі фенолу (отрути для E. coli), бактерія рухатиметься за градієнтом. Вона ніколи не навчиться. У неї немає мозку — лише кілька простих хімічних реакцій для здійснення життєдіяльності.

Значним поступом виявився *потенціал* у формі електричних сигналів одноклітинних організмів, яким вони почали користуватися близько мільярда років тому. Пізніше багатоклітинні організми залучили до цього спеціальні клітини, що нині називаються *нейронами.* Електричний потенціал використовувався для швидкого передавання сигналу в організмі — до 120 метрів за секунду чи 270 миль за годину. Зв'язок між нейронами називається *синапсом*. Сила зв'язку синапсів диктує інтенсивність електричного збудження, що передається від нейрона до нейрона. Завдяки зміні зв'язку синапсів тварини навчаються[11]. Навчання надає величезну еволюційну перевагу, адже тварини можуть адаптуватися до різноманітних умов. Навчання також пришвидшує саму еволюцію.

Спочатку нейрони організувалися у *нервові сітки,* які пронизують організм і слугують для координації таких дій, як харчування, травлення чи відповідне скорочення м'язових клітин на широких ділянках. Граційний рух медуз — результат дії нервової мережі. Медузи зовсім не мають мозку.

Мозок виникне пізніше, разом зі складними органами відчуттів, такими як очі й вуха. За кілька сотень мільйонів років після медуз із їхніми нервовими мережами, прийшли ми,

люди, з нашим великим мозком — сотнею мільярдів (10^{11}) нейронів та квадрильйоном (10^{15}) синапсів. Доволі повільний порівняно з електронною схемою «час циклу» від кількох мілісекунд для зміни стану — це швидко, якщо порівняти з більшістю біологічних процесів. Людський мозок часто описується його носіями як «найскладніший об'єкт у Всесвіті». Насправді це не так, хоча й здається прийнятним виправданням нашого недостатнього розуміння процесів, які в ньому відбуваються. Хоча ми багато знаємо про біохімію нейронів і синапсів та анатомічну структуру мозку, нейронна організація *когнітивного* рівня — навчання, знання, запам'ятовування, причино-наслідкові зв'язки, планування, прийняття рішень тощо — достеменно досі невідома[12]. (Можливо, все зміниться, коли ми більше знатимемо про штучний інтелект або винайдемо точні інструменти для вимірювання мозкової активності.) Тож коли читаєте в медіа, що така-то й така-то технологія штучного інтелекту працює «точно як людський мозок», — це чийсь здогад або просто вигадка.

Про *свідомість* ми нічого не знаємо, тож я нічого й не говоритиму про неї. Ніхто в галузі штучного інтелекту не намагається зробити машини свідомими. Ніхто навіть не знає, звідки почати, бо жодна поведінка не передбачає свідомість як обов'язкову умову. Уявіть, я даю вам програму і запитую: «Чи становить це загрозу людству?». Ви аналізуєте код і справді: задіяний, він сформує і здійснить план, результатом якого стане знищення людської раси. Це просто як програма з гри в шахи формулює і здійснює план, результатом якого буде перемога над будь-якою людиною-партнером. Отже, я кажу, уявіть, що код у разі його запуску так само створить форму машинної свідомості. Ви зміните ваше передбачення? Зовсім ні. Тут немає *жодної різниці*[13]. Ваше передбачення поведінки програми незмінне, тому що передбачення базується на кодові. Автори всіх цих голлівудських сюжетів про машини, які

загадковим чином здобувають свідомість і ненавидять людей, насправді випускають з поля зору найголовніше: має значення лише компетентність, а не свідомість.

Однак існує один когнітивний аспект роботи мозку, який ми починаємо розуміти, а саме: *система винагороди*. Це внутрішня сигнальна система, яка за посередництва дофаміну поєднує позитивні й негативні стимули поведінки. Дію цієї речовини дослідили шведський нейробіолог Нілсо-Еке Хілларп і його помічники наприкінці 1950-х років. Дофамін змушує нас шукати позитивних стимулів, таких, як підсолоджена їжа, що підвищує його рівень, та уникати негативних — голоду й болю, бо вони цей рівень знижують. У певному сенсі це досить подібне до механізму пошуку глюкози бактерією E. Coli, але набагато складніше. Процес пов'язаний із вбудованими методами навчання, тож наша поведінка з плином часу стає ефективнішою після отримання винагороди. Це також дає змогу затримувати задоволення, тож ми навчаємося бажати таких речей, як гроші, тобто проміжної винагороди, а не негайної. Розуміти систему винагороди мозку необхідно вже хоча б тому, що вона нагадує систему *навчання з підкріпленням*, розроблену в ШІ, для якої у нас є серйозна теорія[14].

З еволюційної точки зору можемо вважати систему винагороди мозку, як і механізм пошуку глюкози E. Coli, способом удосконалення еволюційної придатності. Організми, ефективніші в пошуку винагороди — смачної їжі, уникнення болю, сексуальної активності тощо, — імовірніше передадуть свої гени. Організмові вкрай складно вирішити, яка дія в далекій перспективі завершиться успішним передаванням генів, тож еволюція спростила процес: забезпечила нас вбудованими дороговказами.

Проте ці дороговкази не ідеальні. Існують способи отримання винагороди, що *зменшують* імовірність передавання генів. Наприклад, вживання наркотиків, надміру газованих напоїв із цукром, відеоігри вісімнадцять годин на

добу — усе це, здається, суперечить ставкам на розмноження. Ба більше, якщо вам дати доступ до електричної стимуляції системи винагороди, ви, напевно, безперервно стимулюватимете себе, поки не помрете[15].

Випадки, коли сигнали винагороди й еволюційної придатності не збігаються, можуть згубно впливати не лише на окремих індивідів. На маленькому острові біля берегів Панами живе популяція трипалих лінивців, які через залежність від субстанції, подібної до валіуму, присутньої в їхньому харчуванні, — листі червоного мангрового дерева, опинилася на межі вимирання[16]. Тож, здається, увесь вид може вимерти, якщо знайде екологічну нішу, де задовольнятиме свою систему винагороди в дезаптативний спосіб.

Але крім таких випадкових помилок, навчання, спрямоване на максимальну активацію системи винагороди, в природному середовищі поліпшує шанси на передавання генів та виживання.

Еволюційний прискорювач

Навчання добре не лише для виживання й процвітання. Воно також *пришвидшує еволюцію.* Як це можливо? Зрештою навчання не змінює ДНК, а еволюція зосереджена на зміні ДНК поколінь. Визначення зв'язку між навчанням та еволюцією запропоноване 1896 року американським психологом Джеймсом Бо-лдвіном[17]. Незалежно від нього ту саму теорію висунув британський етолог Конві Ллойд Морґан[18], але в той час вона не здобула загального визнання.

Ефект, зараз відомий під назвою ефекту Болдвіна, стане зрозумілим, якщо уявивти, ніби еволюція має вибір між створенням *інстинктивного* організму, кожна реакція якого буде фіксованою, і *адаптивним* — який навчається. Тепер уявіть, наприклад, що оптимальний інстинктивний організм може бути закодований у шестизначному числі, скажімо, 472116, тимчасом як у випадку адаптивного організму еволюція

вказує лише 472***. І організм, завдяки навчанню, має самотужки впродовж свого життя додати останні три цифри. Зрозуміло, якщо еволюції необхідні лише перші три цифри, її робота набагато простіша: адаптивний організм за допомогою навчання за своє життя добере останні знаки коду, а еволюції на це знадобилося б чимало поколінь. Тож за умови виживання адаптивних організмів у процесі навчання можна припустити, що здатність до нього — найраціональні-ша з точки зору еволюції. Розробники цифрового моделювання твердять, що ефект Болдвіна реально працює[19]. Культурні чинники лише пришвидшують процес, тобто організовують цивілізаційний захист організму в процесі його навчання та забезпечують інформацією, яку в іншому разі індивід мусив би шукати самотужки.

Історія відкриття ефекту Болдвіна захоплива, але неповна: вона підсумовує, що навчання й еволюція обов'язково вказують в одному напрямку. Так, вона підводить до певного висновку. Але це не буде висновком про внутрішній сигнал зворотного зв'язку як напрям навчання. Це означатиме звичайну здатність еволюціонувати. Як бачимо, у випадку з трипалим лінивцем, не завжди ця здатність спрацьовує. У найкращому разі механізми, вбудовані для навчання, здатні зберегти результат будь-якої дії, що сприяє еволюції. Ба більше, постає запитання: «Як взагалі виникла система винагороди?». Відповідь, звичайно, криється в еволюційному процесі: засвоєння механізму зворотного зв'язку, який принаймні певною мірою узгоджується зі здатністю еволюціонувати[20]. Звісно, механізм, який навчав би втікати від потенційної пари й кидатися назустріч хижакам, не протримався б довго.

Таким чином, маємо подякувати природі за ефект Болдвіна: за те, що нейрони з їхньою здатністю до навчання та вирішення проблем поширені в світі тварин. Водночас важливо розуміти, що еволюція насправді не переймається наявністю мозку чи складністю мислення. В її процесі ви лише *агент*, тобто маєте

діяти. Такі важливі інтелектуальні характеристики, як логічне міркування, цілеспрямоване планування, мудрість, дотепність, уява та креативність, можуть зробити агента розумним, а можуть і не зробити. Одна з причин надмірного захоплення штучним інтелектом у тому, що він пропонує потенційний шлях до відповіді на ці запитання. Ми можемо дійти розуміння того, як ці інтелектуальні характеристики спричинюють розумну поведінку. І завдяки цьому виразно бачимо, чому без них неможливо її сформувати.

Раціональність для одного

Від самих витоків давньогрецької філософії концепція інтелекту прив'язана до можливості *успішно*[21] сприймати, міркувати й діяти. За століття дана концепція набула точнішого визначення й ширшого застосування.

Серед інших Арістотель вивчав поняття успішного міркування або методи логічної дедукції, що за відповідних передумов привели б до правильних висновків. Він також вивчав процес прийняття рішень стосовно поведінки, який іноді називається *практичним міркуванням.* Він передбачає певну послідовність дій, яка за підтримки дедукції приведе до бажаної мети:

«Ми розмірковуємо не про результат, а про засоби. Адже лікар не розмірковує про те, чи має він лікувати, й оратор не сумнівається в тому, що має когось переконати… Вони припускають результат і розмірковують, яким чином і за допомогою яких засобів його досягти, чи легко це буде, чи найкращий це шлях. Якщо це досягається певними засобами, вони міркують, як саме їх використати й що вкаже на досягнення мети, аж поки не доходять першопричини… І те, що є останнім у плані аналізу, напевне, стане першим у плані здійснення. І якщо доходимо висновку, що завдання виконати неможливо, припиняємо пошук, наприклад, якщо нам потрібні гроші й ми не

можемо їх дістати. Але якщо виконати завдання, на нашу думку, можливо, ми намагаємося досягти мети»[22].

Цей пасаж, із яким можна сперечатися, запо чат-кував напрям розвитку західної концепції раціональності на дві тисячі років. Мислитель стверджує, що «результат», якого прагне індивід, — фіксований; і тут говориться, що раціональна дія — це та, що відповідно до логічної дедукції, через послідовність дій «найпростіше та найшвидше» приводить до результату.

Позиція Арістотеля видається обґрунтованою, але це неповне визначення раціональної поведінки. Зокрема вона нехтує проблемою невпевненості. Реальний світ часто втручається в нашу діяльність, і кілька вчинків чи їх послідовність насправді не гарантують досягнення передбачуваного результату. Наприклад, я пишу це речення дощової неділі в Парижі, а у вівторок о 2:15 мій рейс до Риму вирушає з аеропорту Чарльза де Голля. До цього аеропорту від мого будинку потрібно їхати сорок п'ять хвилин. Я планую вийти з дому об 11:30 ранку, тобто матиму достатньо часу, але, напевно, це означає, що я годину сидітиму в залі очікування. Чи я *точно* впевнений, що потраплю на рейс? Зовсім ні. Можливо, я натраплю на велетенський затор, можливо, страйкуватимуть водії таксі, машина, в якій їхатиму, може зламатися, або водія заарештують за перевищення швидкості тощо. Тож я можу вирушити до аеропорту в понеділок. Це дуже знижує шанси на запізнення, але перспектива ночі в залі очікування не надто приємна. Іншими словами, мій план містить *компроміс* між упевненістю в успіхові та ціною такої впевненості. Наступний план купівлі будинку теж містить подібний компроміс: якщо купити лотерейний квиток і виграти мільйон доларів, можна придбати будинок. Цей план «найпростіше та найкраще» приведе до результату, але він не надто здійсненний. Різниця між необдуманим планом придбання будинку й моїм тверезим

та розумним планом поїздки до аеропорту — у ступені вірогідності. І там і там — лотерея, але перший план видається раціональнішим за другий.

Виявляється, за врахування невпевненості випадковість відіграє центральну роль в узагальнені Арістотеля. 1560 року італійський математик Джироламо Кардано використав гру в кості як свій основний приклад і розвинув першу математично точну теорію імовірності. (На жаль, його роботу не публікували до 1663 року[23]). У сімнадцятому столітті французькі мислителі, зокрема Антуан Арно та Блез Паскаль (без сумніву, лише з цікавості) вивчали питання раціональних рішень в азартних іграх[24]. Уявіть собі такі ставки.

А: 20% імовірності виграти $10
Б: 5% імовірності виграти $100

Математики запропонували те, що, напевно, запропонували б і ви: порівняти *математичний розрахунок сподіваних ставок*, що означає середню суму, яку ви очікували отримати від кожної. Для ставки А за розрахунком сподіваєтеся на 20% зі $10, тобто $2. Для ставки Б — 5% зі $100 або $5. Тож відповідно до цієї теорії ставка Б краща. Теорія має сенс за постійного повторення однакових ставок. Тоді той, хто ретельно дотримується цього правила здобуває більше грошей, ніж той, хто не дотримується.

У вісімнадцятому столітті швейцарський математик Деніель Бернуллі помітив, що це правило не надто спрацьовує для більших сум[25]. Наприклад, уявіть такі дві ставки.

А: 100%-ий шанс отримати $10 000 000 (сподівання за розрахунком — $10 000 000).
Б. 1 шанс отримати $10 000 000 100 ($10 000 001).

Більшість читачів цієї книги так само, як і автор, бажатимуть ставки А, а не Б, хоча правила розрахунку доводять

протилежне! Бернуллі вважав, що ставки оцінюються не відповідно до очікуваних сум, а з огляду на *користь*. Користь або здатність забезпечити певний зиск, на його думку, була внутрішньою суб'єктивною величиною пов'язаною з грошовою вартістю, але не тотожною їй. Зокрема користь теж змінюється за певним законом *що більше бажання отримати гроші, то нижча вірогідність виграшу.* Це означає, що користь від виграшу не прямо пропорційна сумі: вона зростає повільніше. Наприклад, користь від $1 000 000 100 у сто разів менша за користь від $10 000 000. На скільки менша? Запитайте у себе! Які шанси виграти мільярд доларів, якщо вам гарантовано дають десять мільйонів? Я запитав про це в своїх студентів-випускників, і їхня відповідь була: «Приблизно 50%». Це означає, що заклад Б за розрахунками сподіватиметься на 500 мільйонів доларів, тобто такою є бажана ставка А. Дозвольте мені знову сказати: ставка Б матиме очікувану вартість, у п'ятдесят разів вищу за ставку А. Проте обидві ставки будуть однаково корисні.

Запровадження принципу корисності Бернуллі або невидимого надбання для пояснення людської поведінки відносно математичної теорії стало визначною подією свого часу. Це було ще дивовижніше з огляду на те, що, на відміну від грошових сум, споживацька цінність різноманітних закладів і виграшів не може визначатися безпосередньо. В*исновку про користь можна дійти* з *уподобань* окремої особи. Минуло ще два століття, перш ніж сенс цього відкриття повністю опрацювали та почали широко застосовувати в статистиці й економіці.

У середині двадцятого століття Джон фон Нейман (визначний математик, ім'ям якого назвали «архітектуру фон Неймана» для комп'ютерів[26]), і Оскар Морґенштерн опублікували *аксіоматичну* основу теорії корисності[27]. Це означає ось що: поки вподобання демонструють доведення певних базових аксіом, які мають задовольнити будь-якого раціонального агента, вибір цього індивіда *обов'язково*

описується як максимізація розрахованої функції корисності. Якщо висловлюватися стисло, *раціональний агент працює на максимальну очікувану користь*.

Важко перебільшити значення цього висновку. У багатьох сенсах штучний інтелект — це здебільшого розробка деталей, що дадуть змогу сконструювати раціональні машини.

Уважніше розгляньмо аксіоми, що мають задовольнити раціональні утворення. Ось одна з них, яка називається *транзитивністю*. Якщо ви надаєте перевагу A перед B та B перед C, — перевага на боці A перед C. Це здається доволі розважливим! (Якщо ви любите піцу з ковбасою більше, ніж звичайну, а звичайну — більше за піцу з ананасами, здається, цілком розсудливим передбачення: ви оберете піцу з ковбасою, а не з ананасами.) А ось і ще одна так звана *монотонна сентенція*: якщо ви надаєте перевагу виграшеві A перед виграшем B і маєте обрати з лотерей, де A і B — лише два можливі результати, надасте перевагу лотереї з найвищою імовірністю виграшу A, а не B. І знову достатньо розважливо.

Тут ідеться не лише про піцу чи лотереї з грошовими виграшами. Може йтися про будь-що, зокрема про майбутнє ваше життя й життя інших. Коли йдеться про вибір, що передбачає послідовність подій у часі, часто додатково припускається те, що ми звемо *стаціонарністю*. Два різні варіанти майбутнього A і B починаються з однакової події. Якщо надамо перевагу A перед B, й далі надаватимемо перевагу A перед B,

коли станеться передбачена подія. Це звучить помірковано, але наслідок несподіваний: користь будь-якої послідовності подій — це сума вигоди. І вона пов'язана з кожною подією (напевне, з плином часу вигода здається меншою — за певним рейтингом ментальної зацікавленості)[28]. Хоча це припущення про «користь як суму винагород» значно поширене (принаймні сягає вісімнадцятого століття, часів «гедонічного обчислення» Джеремі Бентрама, засновника утилітаризму), припущення

стаціонарності, на якому воно базувалося, не обов'язкове для раціональних агентів. Стаціонарність також виключає вірогідність того, що вподобання можуть змінюватися з плином часу, хоча наш досвід вказує на протилежне.

Попри обґрунтованість аксіом та важливість вис нов-ків із них, теорію корисності від початку її виник нення постійно критикували. Дехто зневажає її за зведення всього до грошей і егоїстичного розрахунку. (Деякі французькі автори[29] глузливо називали цю теорію «американською», хоча її коріння у Франції.) Насправді це ідеально раціональне прагнення — жити в самозреченні, намагатися послабити страждання інших. Альтруїзм просто означає надання значної ваги добробутові інших за оцінювання визначеного майбутнього.

Інші заперечення пов'язані з труднощами у вираховуванні відповідних імовірностей і споживацької цінності та перемножування їх. Саме так можна визначити очікувану користь. Ті, хто висував ці заперечення, просто плутали різні поняття: обрання просто раціональної дії та обрання раціональної дії з *вирахуванням очікуваної корисності користі* . Наприклад, якщо притиснете пальцем своє очне яблуко, повіка заплющиться, щоб захистити око. Це раціонально, але тут не задіяні жодні розрахунки з очікуваної корисності. Або уявіть, що їдете на велосипеді без гальм униз пагорбом. У вас є вибір: врізатися в бетонну стіну на швидкості десять миль на годину чи в іншу стіну на швидкості двадцять миль на годину. Що ви оберете? Якщо обираєте десять миль на годину, вітаю! Ви розрахували очікувану користь (корисність)? Напевне, ні. Але вибір десяти миль на годину все одно раціональний. Це випливає з двох головних припущень: по-перше, ви сподіваєтеся на менш серйозні ушкодження. По-друге, підвищення швидкості руху так само підвищує вірогідність будь- якого рівня ушкоджень під час зіткнення. З цих двох припущень у разі застосування математичного розрахун ку випливає (без наведення жодних цифр), що удар за десяти миль

на годину підвищує очікувану корисність порівняно з ударом за двадцяти[30]. Підіб'ємо підсумок: для максимізації очікуваної корисності не обов'язковий розрахунок певних очікувань чи певної корисності. Це суто *зовнішня* констатація раціональної суті.

Інші критики теорії раціональності наполягають на ідентифікації локусу прийняття рішень. Що саме вважається агентами? Здається, зрозуміло: агенти — це люди. Та як щодо родин, корпорацій, культур і національних країн? Якщо ми досліджуємо соціальних комах, таких як мурахи, чи має сенс припущення, що одна мураха є розумним агентом? Чи насправді розумом наділена ціла колонія з її спільним мозком, який вміщує численні мізки й тіла мурах, пов'язані феромоновими сигналами замість електричних? З еволюційної точки зору розглядати мурах загалом, а не одну комаху продуктивніше, адже мурахи в колонії зазвичай тісно пов'язані. Як індивідам, мурахам та іншим соціальним комахам, як на мене, бракує інстинкту самозбереження, а не збереження колонії. Вони завжди кидатимуться в бій проти нападників, навіть якщо це просто самогубство. Хоч іноді люди роблять те саме, коли захищають навіть зовсім чужих людей. Здається, вид виграє від існування певної кількості особин, які воліють пожертвувати собою у бою, вирушити в небезпечні дослідницькі експедиції чи годувати чужих нащадків. У таких випадках чогось істотно бракує аналізові раціональності, що повністю зосереджується на індивідові.

Інші принципові заперечення теорії корисності емпіричні; вони базуються на експериментальних свідченнях стосовно людської ірраціональності. Ми практично не відповідаємо аксіомам[31]. Нині не маю на меті обстоювати теорію корисності як формальну модель людської поведінки. Насправді люди не можуть поводитися раціонально. Наші вподобання тяжіють над усіма нашими майбутніми життями, над життями наших дітей та онуків, долею людей, які живуть зараз чи житимуть у

майбутньому. І все- таки ми не можемо зробити правильний рух навіть на шаховій дошці, маленькій, простій території з чітко визначеними правилами та вкрай обмеженим полем можливостей. Це не тому, що наш *вибір* нераціональний, а через *складність* вирішення проблеми. Значна частина нашої когнітивної структури існує для компенсації невідповідності між нашими маленькими повільними мізками та незбагненною складністю проблем, з якими постійно стикаємося.

Тому нерозважливо базувати теорію корисного штуч ного інтелекту на припущенні, буцімто люди раціональні. Достатньо обґрунтоване припущення: доросла людина має в основному послідовні вподобання щодо майбутнього життя. Це означає — *якщо вам дивом вдасться дивитися два фільми, кожен із яких достатньо деталізовано описуватиме майбутнє, спричинене вашими діями, ви можете через віртуальний досвід визначити більш бажаний варіант або оголосити нейтралітет*[32].

Можливо, це твердження сміливіше, ніж потрібно, якщо наша єдина мета — впевнитися в безпечності розумних машин та в тому, що вони не стануть катастрофою для людської раси. Саме поняття *катастрофи* передбачає однозначно небажані події. Для уникнення їх нам потрібно лише проголосити, що дорослі люди можуть розпізнати катастрофічне майбутнє, якщо воно в деталях постане перед ними. Звісно, людські уподобання мають різноманітнішу й, імовірно, визначенішу структуру, ніж твердження «відсутність катастроф ліпша за катастрофи».

Теорія корисного штучного інтелекту насправді може примирити неузгодженість людських бажань. Але непослідовна частина ваших уподобань ніколи не буде задоволена й жоден штучний інтелект тут не зарадить. Уявіть, наприклад, що ваші вподобання піци порушують аксіому транзитивності.

РОБОТ: Вітаю вдома! Хочете ананасової піци?

ВИ: Ні, ти мав би знати, я надаю перевагу звичайній піці, а не ананасовій.

РОБОТ: Добре, невдовзі буде звичайна піца!

ВИ: Ні, дякую, я більше люблю піцу з ковбасою.

РОБОТ: Вибачте, несу піцу з ковбасою!

ВИ: Насправді я надаю перевагу ананасовій піці, а не піці з ковбасою.

РОБОТ: Я помилився, ось піца з ананасом!

ВИ: Я уже сказав, що люблю звичайну піцу більше, ніж ананасову.

Піци, яку робот може дати, щоб задовольнити вас, не існує, бо ви завжди надаватимете перевагу іншій. Робот може задовольнити лише послідовну частину ваших уподобань. Скажімо, ви надаєте перевагу бодай якійсь піці перед її цілковитою відсутністю. У цьому випадку робот може дати вам будь-яку з трьох піц, бо ви волієте уникати відсутності піци, та дасть вам змогу на дозвіллі розмірковувати про свої химерні вподобання.

Раціональність для двох

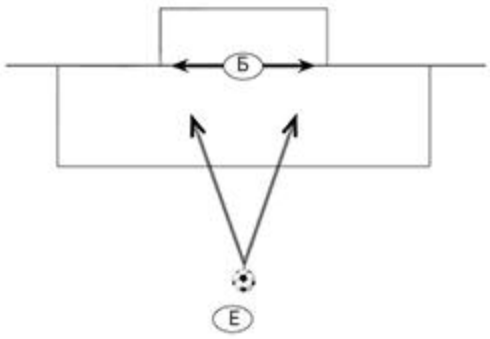
Головна думка проста: раціональний агент намагається максимізувати очікувану корисність, навіть якщо насправді це дуже важко. Але теорія застосовується лише тоді, коли агент діє сам. Більш ніж з одним агентом проблематичним стає припущення, щодо можливого в принципі визначення різних результатів. Причина в наявності іншого агента, який намагається вгадати ваші дії. Тому складно передбачити його можливу поведінку. А без визначення ступеню імовірності того чи того раціонального вчинку максимізація очікуваної корисності неможлива.

Щойно з'являється ще хтось, агент потребує іншого способу пошуку раціональних рішень. Тут і вступає в силу *теорія гри*.

Попри свою назву ця теорія не обов'язково передбачає ігри в звичному сенсі; це лише спроба розширити поняття раціональності в ситуації з багатьма агентами. Безперечно, це важливо для досягнення мети, бо ми не плануємо (поки що) створювати роботів для застосування їх на ненаселених планетах інших зоряних систем, — просто хочемо привести роботів у світ, населений нами.

Для пояснення нашої потреби щодо теорії ігор роз-гляньмо простий приклад: Еліс і Боб грають у футбол у саду (малюнок 3). Еліс має бити пенальті по воротах Боба.

Еліс битиме праворуч або ліворуч. Їй зручніше послуговуватися правою ногою, тож удар буде точні-шим, якщо вдарити праворуч, з точки зору Боба. Еліс б'є дуже сильно, й Боб знає, що має відразу стрибнути в той чи той бік. Йому нема коли чекати, щоби поди-витися, в який бік летітиме м'яч. Боб може міркувати так: «У Еліс більші шанси забити, якщо вона вдарить праворуч, бо їй зручніше бити правою ногою. Отже, вона битиме праворуч, і мені варто стрибнути туди». Але Еліс розумна й може здогадатися, що Боб міркуватиме саме так. Тому вона битиме ліворуч від Боба. Але Боб також розумний; він здогадається, що Еліс думає саме так, тому стрибне ліворуч. Але й Еліс може здогадатися, що Боб здогадається про це... Гаразд, ви зрозуміли. Підемо іншим шляхом: якщо йдеться про раціональний вибір для Еліс, і Боб також може його з'ясувати, передбачити й завадити їй, — її вибір більше не буде раціональним.



Малюнок 3. Еліс готується забити пенальті Бобові.

Лише 1713 року (і знову завдяки аналізові азартних ігор) знайшли розв'язання для цієї головкрутки[33]. Воно полягає в тому, що не слід обирати певну дію — достатньо обрати *випадкову стратегію.* Наприклад, Еліс може обрати стратегію «з вірогідністю 55% успішного удару праворуч і 45% — ліворуч». Боб може обрати стратегію «із вірогідністю 60% вдалого стрибка праворуч і 40% — ліворуч». Перш ніж кожен почне діяти, він подумки мовби кидає монету, тому й не викаже своїх намірів. Якщо Еліс і Боб діятимуть *непередбачувано,* їм вдасться уникнути описаних вище суперечностей. Навіть коли Боб з'ясує, що Еліс обрала випадкову стратегію, це йому ніяк не допоможе, якщо він не має кришталевої кулі.

Наступне запитання: якою *має* бути ймовірність? Вибір Еліс раціональний на 55 чи на 45%? Точно оцінити його можна залежно від точності удару Еліс праворуч та від здатності Боба врятувати ворота, якщо він стрибне теж праворуч, і так далі. (Дивіться примітки для повного аналізу[34].) Загальний критерій, однак, дуже простий.

1. Стратегія Еліс буде найкращою з усіх, на які вона здатна, за припущення, що стратегія Боба фіксована.

2. Стратегія Боба виявиться найкращою з усіх, на які він здатний, за припущення, що стратегія Еліс фіксована.

Якщо обидві умови виконані, говоримо про стратегії в стані рівноваги. Цей вид рівноваги названий *рівновагою Неша* на честь Джона Неша, який 1950 року в свої двадцять два довів, що така рівновага можлива для будь-якого числа агентів із будь-яким раціональним вибором попри всі правила гри. За кілька десятиліть боротьби з шизофренією Неш зрештою одужав і 1994 року нагороджений за цю роботу Нобелівською премією з економіки.

Для футболу Еліс і Боба можлива єдина рівновага. В інших випадках вона може мати варіанти, тож концеп ція рівноваги

Неша, на відміну від рішень з приводу очікуваної користі, не завжди може дати рекомендації щодо поведінки. Ба більше, в деяких ситуаціях рівновага Неша, здається, має небажані наслідки. Одна з таких ситуацій — відома *дилема в'язня,* названа так 1950 року науковим керівником Неша Альбертом Такером[35]. Гра — абстрактна модель реальних ситуацій, у яких взаємна співпраця часто була б вигіднішою для всіх зацікавлених осіб. Та попри це люди обирають взаємне нищення.

Дилема в'язня працює так: Еліс та Боб — підозрювані в скоєнні злочину. Їх допитують кожного окремо. Кожен має вибір: зізнатися й наговорити на свого спільника або відмовитися від свідчень[36]. Якщо обоє відмовляться, їх засудять за менш тяжкий злочин і вони відбудуться двома роками позбавлення волі. Якщо обоє зізнаються, їм висунуть звинувачення в тяжчому злочині й засудять на десять років. Якщо один зізнається, а інший відмовиться свідчити, того, хто зізнався, відпустять, а спільник отримає двадцять років.

Тож Еліс міркує так: «Якщо Боб зізнається, я теж маю зізнатися (десять років ліпше, ніж двадцять). Якщо він збирається мовчати, я маю зізнатися (вийти на волю краще, ніж провести два роки у в'язниці). Тож у будь-якому разі я маю зізнатися». Боб міркує так само. Таким чином, обоє зізнаються в злочині й отримують по десять років. Якби обоє відмовилися говорити, могли б відсидіти лише по два роки. Проблема в тому, що обопільна відмова — не рівновага Неша, бо кожен має стимул до зради, адже відпустять того, хто зізнається, якщо спільник промовчить.

Зважте, що Еліс може розмірковувати так: «Хоч би як міркувала я, Боб думатиме так само. Тож усе скінчиться тим, що ми вирішимо однаково одна. Наша спільна відмова ліпша, ніж зізнання, значить, маємо відмовитися від свідчень». Таке міркування було би природним для Еліс і Боба як раціональних агентів, які зроблять корелятивний вибір, враховуючи інтереси одне одного. Це лише один із багатьох підходів, які

випробували прихильники теорії гри, намагаючись отримати не такі гнітючі розв'язки дилеми в'язня у своєму намаганні домогтися простішого розв'язку дилеми в'язня[37].