Нік Бостром

СУПЕРІНТЕЛЕКТ

Стратегії і небезпеки розвитку розумних машин

Переклали з англійської
Антон Ящук,
Антоніна Ящук

«Наш Формат» · Київ · 2020

УДК 517(0.062)
Б85

Бостром Нік

Б85 Суперінтелект. Стратегії і небезпеки розвитку розумних машин / пер. з англ. Антон Ящук, Антоніна Ящук. — К. : Наш Формат, 2020. — 408 с.

ISBN 978-617-7866-31-1 (паперове видання)
ISBN 978-617-7866-32-8 (електронне видання)

Штучний інтелект — це бомба в руках дитини. Що як одного дня з'явиться розум, який перевершить наш — досі найбільший на планеті? Чи стане він руйнівною загрозою, що змінить історію людства?

У цій книжці професор Оксфорду Нік Бостром досліджує наукові теорії, що стали передумовою відкриття штучного інтелекту, та наслідки його впливу. Автор розглядає важливі аспекти: швидкість поширення штучного розуму, його форми і здібності, варіанти стратегічного вибору, перед якими опиниться суперінтелект, щойно отримає вирішальну перевагу. Та найголовніше, Бостром не кидає людство напризволяще, а пропонує конкретні запобіжні заходи, які допоможуть контролювати штучний інтелект у майбутньому.

УДК 517(0.062)

Перекладено за виданням: Nick Bostrom. *Superintelligence: Paths, Dangers, Strategies* (Oxford, Oxford University Press, 2014, ISBN 978-0-19-873983-8).

Superintelligence was originally published in English in 2014. This translation is published by arrangement with Oxford University Press. Nash Format is solely responsible for this translation from the original work and Oxford University Press shall have no liability for any errors, omissions or inaccuracies or ambiguities in such translation or for any losses caused by reliance thereon.

ISBN 978-617-7866-31-1 (паперове видання)
ISBN 978-617-7866-32-8 (електронне видання)

ЗМІСТ

Історія про зграю горобців (із відкритим фіналом) 7
Передмова 9

1. Що ми маємо і на що здатні 12
2. Шлях до суперінтелекту 37
3. Форми суперінтелекту 70
4. Кінетика інтелектуального вибуху 82
5. Вирішальна стратегічна перевага 100
6. Розумові суперздібності 114
7. Воля суперінтелекту 130
8. То ми всі приречені? 141
9. Проблема контролю 155
10. Оракули, джини, суверени, інструменти 175
11. Багатополярні сценарії 190
12. Формування цінностей 220
13. Обираємо критерій відбору 247
14. Стратегічна картина 269
15. Переломний момент 301

Післямова 307
Подяки 311
Словник термінів і понять 313
Бібліографія 318
Примітки 346

ІСТОРІЯ ПРО ЗГРАЮ ГОРОБЦІВ (ІЗ ВІДКРИТИМ ФІНАЛОМ)

Якось надвечір горобці обсіли гілки дерева, щоб разом поцвірінькати й відпочити після дня важкої праці. Був сезон будування гнізд, і пташки добряче натомилися.

«Ми такі малі та слабкі, — писнув один. — Уявіть, наскільки легше нам велося б, якби гнізда нашій зграї допомагала будувати сова!»

«Еге ж, — погодився другий, — а ще сова могла б доглядати наших старих і малюків».

А третій підхопив: «Вона могла б давати нам мудрі поради і до того ж стежити за сусідським котом».

Тоді старий поважний горобець Правічник промовив: «Розішлімо в усі усюди розвідників і спробуймо знайти покинуте совеня, а ще краще — яйце. Для наших потреб згодиться і вороненя або дитинча ласки. Якщо нам вдасться знайти когось із них, це буде неймовірна удача! Навіть більша, ніж відкриття Павільйону з нескінченним запасом зерна в сусідньому саду!».

Почувши це, горобці так зраділи, що розверещалися на всю околицю.

І тільки Цінь-Цвірінь,одноокий горобець, який постійно все робив усім наперекір, засумнівався, чи варто починати цю справу. Серед гамору пролунав його скрипучий голос: «На погибіль собі придумали ми таку затію. Хіба не варто нам спершу опанувати мистецтво вирощування та приручення сов, а потім уже селити цих істот посеред нас?».

На це Правічник відказав: «Знайти сову — то вже нелегка справа. Украй важко знайти також і яйце. Почнімо з цього. А ось коли зможемо виростити сову, тоді подумаємо, як розв'язати наступну проблему — приручити її».

«Але це поганий план!» — кричав Цінь-Цвірінь, та ніхто вже його не слухав: птахи хутко розлетілися виконувати вказівку Правічника.

Біля нього залишилося лише кілька горобців. Гуртом вони спробували придумати, як приручити сов. Та скоро зрозуміли, що Правічник мав рацію: це було надзвичайно складне завдання, тим більше, що в них не було сови, на якій можна було б тренуватися. А втім, горобці й далі мужньо розмірковували над розв'язанням проблеми і все оглядалися, чи не летить хтось з їхньої зграї із совиним яйцем у лапах. Бо ж рішення проблеми вони поки що не знайшли...

Чим закінчилася ця історія — невідомо, але автор присвячує цю книжку Цінь-Цвіріневі та його побратимам.

ПЕРЕДМОВА

У вашому черепі є орган, який читає цю книжку. Це мозок, і в нього є властивості, яких немає в мізках інших тварин. Саме завдяки цьому людина домінує на планеті Земля. В інших тварин сильніші м'язи й гостріші пазурі, зате в нас розумніший мозок. І завдяки цій скромній інтелектуальній перевазі ми створили мову, розробили різні технології й розвинули складну соціальну організацію. Інтелектуальні переваги людини накопичуються з часом, адже кожне покоління будує власні досягнення на досягненнях попередників.

Якщо ми колись збудуємо машину з мозком, який інтелектуально переважатиме над людським, то цей новостворений штучний інтелект може стати неймовірно потужним. І доля нашого виду залежатиме від дій супермашини так, як горили нині більше залежать від діяльності людини, ніж від інших горил.

Та все-таки одна перевага в нас є: збудувати цю розумну машину маємо ми. Насправді люди могли б створити штучний інтелект, який захищав би гуманістичні цінності. І нам таки варто було б створити його саме таким. Адже на практиці проблему контролю — як керувати діяльністю штучного інтелекту — розв'язати дуже непросто. Схоже, у нас буде лиш один шанс на спробу. Бо якщо штучний інтелект виявиться до нас ворожим, він легко перепинить наші намагання змінити його преференції. І невідомо, що з нами станеться потім.

У цій книжці я намагаюся осмислити, які випробування можуть на нас чекати, якщо буде винайдено штучний інтелект, і якими мають бути наші дії. Ці випробування будуть, напевно, найважливішими та найстрашнішими за всю історію людства. Й останніми — незалежно від того, успішно вони завершаться для нас чи ні.

Я не стверджуватиму, що ми на порозі винайдення штучного інтелекту, і не передбачатиму, бодай приблизно, коли він з'явиться в жит-

ті людства. Є підстави вважати, що штучний інтелект буде винайдено вже в цьому столітті, але ніхто цього не знає напевно. У перших кількох розділах ми розглянемо способи, як рухатися в цьому напрямі, а також поміркуємо про часові межі. Але здебільшого у книжці йтиметься про те, що станеться потім. Ми розглянемо кінетику вибуху штучного інтелекту, його форми й здібності, а також варіанти стратегічного вибору, перед якими опиниться суперінтелектуальний агент, щойно дістане вирішальну перевагу. Після цього ми зупинимося на питанні контролю й подумаємо, що можна зробити, щоб штучний інтелект не загрожував нашому виживанню, а, навпаки, приносив користь. Ближче до кінця книжки спробуємо оглянути майбутнє, що постає з цього дослідження, під ширшим кутом. І висловимо припущення, що можна зробити зараз, щоб уникнути екзистенційної катастрофи потім.

Ця праця далася мені нелегко. Маю надію, що я сяк-так розчистив стежку, якою до нових горизонтів упевнено пройде ще не один науковець, і що до своєї цілі він дістанеться бадьорим і сповненим сили працювати над розв'язанням нашої спільної проблеми. (А якщо стежка моя виявиться надто кам'янистою та звивистою, сподіваюся, що критики, оцінюючи результат, врахують, що прокладати шлях ex ante на цій території було дуже складно!)

Писати книжку було непросто, але я доклав усіх зусиль, щоб її було легко читати. Не певен, що мені це вдалося. Коли я думав про ідеального читача цієї книжки, то уявляв себе рік чи два тому й намагався писати так, щоб мені тодішньому сподобалося це читати. Визнаю, демографічна вибірка невелика. Але, думаю, багатьом людям буде до снаги зрозуміти написане в цій книжці, особливо якщо вони докладуть певних зусиль і не піддадуться спокусі миттєво й помилково заміняти наведені нові ідеї на схожі кліше, узяті з комор власної культури. Читачів, які не мають достатніх знань у сфері технологій, прошу не покидати книжки при появі математичних обчислень чи спеціальної термінології. Адже головну думку будь-якого уривка можна зрозуміти з текстових пояснень, якщо формули не проясняють ситуацію. (А для тих читачів, які, навпаки, хочуть більше «м'яса», є примітки[1]).

Чимало тверджень із цієї книжки можуть виявитися хибними[2]. Також є імовірність, що я не взяв до уваги якихось важливих міркувань, і через це деякі мої висновки неправильні. Я намагався підкреслити власну непевність (від найменшої до найбільшої) словами: «ймовірно», «можливо», «може», «мабуть», «схоже», «певно», «найімовірніше», «майже напевно». Кожне таке слово я підбирав дуже ретельно і свідомо. Це не просто «топос епістемологічної скромності». Ці сло-

1 ЩО МИ МАЄМО І НА ЩО ЗДАТНІ

Спершу озирнемося назад. Історія з висоти пташиного польоту — це послідовність окремих моделей розвитку, і в межах кожної наступної моделі прогрес відбувався щоразу швидше. Відповідно, можна припустити, що на нас чекає ще один (стрімкіший) період розвитку. Але не про нього тут ітиметься — наша книжка не про «бурхливий прогрес технологій», «експоненційне зростання» чи ще якісь поняття, що їх часто разом називають «сингулярність». Натомість ми оглянемо історію розвитку штучного інтелекту. Потім детально проаналізуємо, які можливості в цій царині має людство нині. І насамкінець згадаємо про найновіші фахові дослідження. А також поміркуємо про те, що знати, як розвиватиметься наше майбутнє, нам поки що не дано.

Моделі розвитку та велика історія

Кілька мільйонів років тому наші предки гойдалися на ліанах в африканських джунглях. Із погляду геології або навіть еволюції відокремлення Homo sapiens від спільного з великими мавпами предка відбулося просто блискавично. У нас розвинулася пряма статура, великий палець на руках розташувався навпроти інших чотирьох, а найголовніше — трохи збільшився об'єм мозку та змінилась організація нейронів. Завдяки цьому люди зробили справжній стрибок у розвитку мислення. Тепер ми можемо міркувати абстрактно, обмінюватися складними думками й накопичувати культурний досвід поколінь набагато краще, ніж інші види істот, що населяють нашу планету.

Завдяки цим здібностям людство змогло створити ефективні знаряддя праці, а відтак розійтися по всій планеті, далеко за межі джунглів і саван. Із розвитком сільського господарства почала зростати гус-

тота й загальна кількість населення на Землі. А що більше людей — то більше ідей. Що більша густота населення — то швидше ці ідеї могли поширюватися. Люди могли присвятити всі сили розвиткові окремих навичок. Відповідно, *швидше зростала* економічна продуктивність і можливості технологій. Пізніше, під час промислової революції, відбувся дальший розвиток і наступний, не менш важливий, крок у прискоренні економічної продуктивності.

Темпи зростання пришвидшувалися, і це мало важливі наслідки. Кілька сотень тисяч років тому, у доісторичні часи, розвиток тривав надзвичайно повільно: знадобився приблизно мільйон років, аби виробничі потужності людства зросли достатньо, щоб кількість населення збільшилася на один мільйон й існувала на межі виживання. Близько 5000 року до н. е. внаслідок аграрної революції темпи зростання пришвидшилися так, що того самого приросту населення вдалося досягти лише за двісті років. Тепер, після промислової революції, світова економіка дає такі темпи зростання в середньому кожні дев'яносто хвилин[3].

Якщо нинішні темпи зросту зберігатимуться порівняно тривалий час, приріст населення й економіки буде колосальним. Якщо економічні показники зростання будуть такими, як останні 50 років, до 2050 року світ стане в 4,8 раза багатшим, а до 2100 року — у 34 рази[4].

Але перспективи стабільного експоненційного приросту бліднуть порівняно з перспективою пережити стрибок у *темпах зростання населення* (згадаймо про стрибок після аграрної та промислової революцій). Спираючись на історичні дані про економіку й кількість населення, економіст Робін Генсон припускає, що двократне зростання світової економіки для мисливців-збирачів плейстоцену відбулося за 224 000 років; для суспільства землеробів — за 909 років; а для промислового суспільства — за 6,3 року[5]. (За нинішньої епохи, згідно з Генсоновою моделлю, поєднання землеробської та індустріальної моделей розвитку: сучасна світова економіка поки що не здатна подвоюватися за 6,3 року). Якби невдовзі відбувся б перехід до іншої моделі розвитку і якби його масштаби були співмірні з масштабами попередніх двох переходів, ми б опинилися в новому режимі зростання світової економіки: там подвоєння тривало б два тижні.

За нинішніх обставин такі темпи зростання видаються фантастичними. А проте людям із минулих епох теж важко було повірити, що світова економіка подвоюватиметься кілька разів за життя однієї особи. Те, що тоді здавалося неймовірним, сьогодні має цілком звичний вигляд.

2 ШЛЯХ ДО СУПЕРІНТЕЛЕКТУ

Зараз машини поступаються людському розуму в загальних завданнях. Але колись (як ми передбачили) вони перевершать його. Як це може трапитися? У цьому розділі спробуємо дослідити кілька технічно можливих способів. Ми розглянемо штучний інтелект, емуляцію цілого мозку, удосконалення біологічного мозку, покращення способів взаємодії людини й машини за допомогою нейроінтерфейсу, потенціалу мережевих та організаційних утворень. Спробуємо оцінити ймовірність досягнення розумових надможливостей кожним із цих варіантів. Адже існування кількох способів досягнення мети значно збільшує сумарну ймовірність її реалізації хоча б одним з них.

Спробуємо дати суперінтелекту таке попереднє означення: *інтелект, розумові можливості якого в більшості важливих для людини сфер діяльності перевищують людські можливості*[87]. Детальний «спектральний аналіз» поняття суперінтелекту для окреслення деяких можливих форм і втілень цієї сутності ми проведемо в наступному розділі. Зараз нам достатньо наведеного означення. До того ж воно не обмежує способу творення суперінтелекту та деталей його влаштування. Безумовно цікаво (зокрема і з погляду моралі), наприклад, чи перебуватиме він у стані суб'єктивної свідомості, чи ні, але поки зосередимося на передумовах і наслідках появи суперінтелекту, а не на його метафізиці[88].

Шахова програма Deep Fritz за визначенням не є суперінтелектом, адже її «розумність» обмежується лише шахами. А втім, деякі випадки вузькоспеціалізованої суперінтелектуальності теж можуть бути цікавими. У таких випадках ми вказуватимемо сферу спеціалізації суперінтелектуальності. Наприклад, розум, що здатний перевершити найкращих інженерів у їхній діяльності, називатимемо «інженерний

суперінтелект». Без цього уточнення термін стосуватиметься *загальної* розумності.

Отже, як би ми могли створити суперінтелект? Нумо розглянемо можливі способи.

Штучний інтелект

Звісно, не варто очікувати тут інструкції зі створення штучного інтелекту. Її поки що не існує. А якби вона в мене була, то я б її точно не публікував у книжці. (Якщо причини такого мого рішення для вас не очевидні, то в наступних розділах ви знайдете його вичерпну аргументацію).

Утім уже зараз ми можемо окреслити деякі загальні риси майбутньої системи. Зокрема, зрозуміло, що уміння навчатися має бути однією з основних властивостей системи, а не чимось, що можна додати потім. Те саме стосується можливості працювати з невизначеністю та з ймовірнісними величинами. Також для досягнення загальної розумності потрібна базова здібність отримувати дані з органів чуттів чи внутрішнього стану системи, перетворювати їх на певні цілісні концептуальні представлення й використовувати під час логічних та інтуїтивних міркувань.

Свого часу системи Старого Доброго Штучного Інтелекту майже не приділяли уваги процесам навчання, врахуванню невизначеності та концептуалізації через слабку розвиненість апарату роботи із цими поняттями. Водночас самі ідеї не такі вже й нові. Застосовувати процес навчання для базової машини, щоб вона могла досягнути розумності, зіставної з людською, пропонував ще Алан Тюрінг у своєму описі «машини-дитини» в 1950 році:

> Чому б замість спроб створити програмну симуляцію розуму дорослого не спробувати створити симуляцію розуму дитини? Адже якщо його правильно навчати, можна згодом отримати дорослий[89].

А так Тюрінг уявляв собі покроковий процес створення такої «машини-дитини»:

> Не варто сподіватися, що вдасться швидко віднайти добру «машину-дитину». Доведеться експериментувати з навчанням і оцінювати успіхи. Якщо одна погано навчатиметься, доведеться спробувати іншу. Цей процес безсумнівно чимось схожий на еволюцію… Про-