НІК БОСТРОМ

БЕСТСЕЛЕР NEW YORK TIMES

СУПЕР ІНТЕЛЕКТ

Стратегії і небезпеки
розвитку розумних машин

Суперінтелект

Nick Bostrom

SUPER INTELLIGENCE

Paths, Dangers, Strategies

Oxford University Press · Oxford · 2014

Нік Бостром

СУПЕРІНТЕЛЕКТ

Стратегії і небезпеки розвитку розумних машин

Переклали з англійської
Антон і Антоніна Ящуки

«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)

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 житті 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 слова вжито як систематичне визнання власної непевності та схильності помилятися. І це не лицемірна скромність: я справді вважаю, що моя книжка може містити хибні ідеї чи ідеї, здатні завести читачів на манівці. А втім, альтернативні погляди, які можна знайти в літературі, ще гірші за мої, і серед них найгірша — загальноприйнята «нульова гіпотеза», за якою ми до певного часу можемо спокійно і свідомо ігнорувати перспективу появи штучного інтелекту.

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 густота 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.

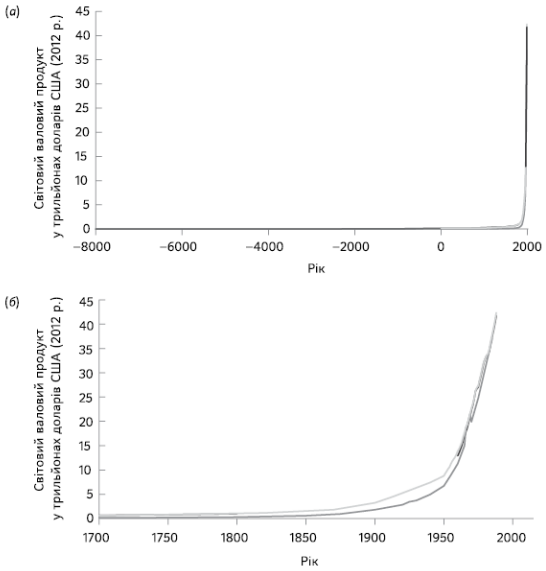


Рисунок 1. **Історія світового валового продукту**
У лінійному масштабі історія світової економіки перетворюється на горизонтальну лінію, яка пролягає дуже близько до осі *x*, а потім різко злітає вертикально вгору (а). Навіть за детальнішого розгляду останніх десяти тисяч років картина майже не міняється. Лише в одному місці графік починає підніматися під кутом 90° (б). Тільки в останні 100 років крива помітно відривається від осі *x*, тобто нуля. (Різні лінії на графіку відповідають різним наборам даних, а отже, демонструють незначні відмінності в оцінках[6]).

Зараз дуже популярною стала теорія про наближення епохи технологічної сингулярності. Уперше про неї заговорив Вернор Віндж, а його концепції підхопили Рей Курцвейл та інші[7]. На жаль, терміну «сингулярність» відтоді часто надавали неправильних значень, і він набув мало не диявольських (хоча разом із тим есхатологічних) техно-утопійних конотацій[8]. А втім, для нашої розмови всі ці значення й конотації не підходять, тому ми не вживатимемо слово «сингулярність», а натомість розробимо точнішу термінологію.

Однак із сингулярністю пов'язане переконання, яке нас тут найбільше цікавитиме: невдовзі може відбутися *вибух інтелекту*, а саме — поява штучного суперінтелекту. До такого переконання можуть привести ілюстрації, наведені на рисунку 1: схоже на те, що незабаром на людство може чекати кардинальна зміна моделі розвитку, яку буде можливо порівнювати з аграрною або промисловою революцією. Можна припустити, що досягти надзвичайно швидкого (за кілька тижнів) подвоєння світової економіки буде нереально, якщо на допомогу людині не прийде значно швидший і ефективніший

розум, ніж той, що міститься у звичайному біологічному мозку. Проте не лише графіки й екстраполяції історичного досвіду спонукають нас очікувати, що революція машинного інтелекту відбудеться вже скоро. Є для цього й серйозніші причини.

Великі сподівання

Що машини стануть гідними суперниками людини в інтелекті — тобто будуть здатні вчитися, матимуть здоровий глузд і зможуть планувати, а отже, виконувати складні завдання, пов'язані з обробленням інформації в різних сферах, конкретних і абстрактних, — очікували вже давно, від моменту винайдення комп'ютерів у 1940-х роках. Тоді, у сорокових, вважали, що розумні машини з'являться через років двадцять[9]. Із плином часу очікування почали відсуватися все далі в майбутнє. Сьогодні футурологи передбачають появу штучного інтелекту через кількадесят років[10].

Двадцять років — улюблений відрізок для пророків значних змін. Цей період достатньо близький у часі, щоб привернути увагу і звучати солідно, але й достатньо віддалений, щоб очікувати запуск необхідних важливих процесів, які сьогодні важко уявити. Говорячи про ближче майбутнє, потрібно врахувати, що технології, які впливатимуть на світ через п'ять-десять років, уже використовують, але тільки в обмеженому колі. А технології, які через п'ятнадцять років змінять наше життя, сьогодні розробляють у лабораторіях. Крім того, якраз двадцять років зазвичай залишається до закінчення кар'єри футурологів, які наважуються пророкувати майбутнє, тому навіть якщо передбачення не збудуться, великої шкоди їхній репутації це не завдасть.

Науковці в 1940-х трохи поспішили з очікуваною датою появи штучного інтелекту; а проте з цього не варто робити висновок, що винайти ШІ неможливо і цього ніколи не станеться[11]. Головна причина, чому прогрес повільніший, ніж очікували: з технічного погляду конструювати такі машини виявилося значно важче, ніж сподівались піонери в цій галузі. Тепер ми бачимо, що технічні перешкоди справді колосальні, і людство ще не дуже наблизилося до їхнього вирішення. Трапляється, звісно, що для проблеми, яка спочатку здається

невирішуваною, раптово знаходиться просте розв'язання (хоча, певно, частіше буває навпаки).

У наступному розділі ми розглянемо різні підходи до розроблення машинного інтелекту, який не поступатиметься людському. Але одразу наголосимо: хай би скільки зупинок відділяло нас на нашому шляху від нинішньої точки до пункту «машинний інтелект, який не поступатиметься людському», цей пункт не буде кінцевим призначенням. Наступною зупинкою (і не такою вже й далекою) буде «машинний інтелект, який перевершить людський». Цей поїзд не зупиниться, ба навіть не пригальмує, на станції «Місто Людство». Він, найімовірніше, зі свистом промчить далі.

Математик Ірвінг Джон Гуд, який під час Другої світової війни був головним статистиком у Тюрінговій команді зі зламу шифрів, став, мабуть, першим, хто передбачив саме такий розвиток подій. Його слова, сказані ще 1965 року, часто цитують:

> *Введемо поняття ультрарозумної машини, тобто машини, яка зможе перевершити в інтелектуальній діяльності навіть найрозумнішу людину. Оскільки створення машини — це теж інтелектуальна діяльність, ультрарозумна машина зможе створювати ще розумніші машини; а тоді, поза сумнівом, відбудеться «вибух інтелекту», і людський розум залишиться далеко позаду. Тому перша ультрарозумна машина стане останнім винаходом людини — але тільки в тому разі, якщо машина виявиться слухняною і навчить нас керувати собою*[12].

Здається, очевидно, що таке вибухове зростання штучного інтелекту значно загрожує людському існуванню. Тому треба якнайсерйозніше ставитися до такої перспективи й вивчати її, навіть якби ми точно знали (а ми не знаємо), що існує лише мінімальна ймовірність появи штучного інтелекту. Проте першопрохідці в галузі машинного інтелекту, хоча й були переконані, що незабаром буде винайдено штучний розум, який не поступатиметься людському, найчастіше не думали про можливість появи інтелекту, що перевершить людський. Ніби м'язи їхньої фантазії так утомлювалися від уяви абсолютно нових можливостей розумних машин, що уявляти далі — як машини перевершують людину — уже просто не лишалося сил.

Піонери в галузі ШІ майже ніколи навіть думки не припускали, що їхні винаходи можуть якось загрожувати людству[13]. Вони цілковито легковажили питаннями безпеки, їхніх сердець не торкалася ні найменша тривога щодо етичних вимірів майбутнього (потенційного) штучного інтелекту і встановлення панування комп'ютерів. Насправді досить дивно, що науковці так уперто ігнорували ці питання, навіть враховуючи загалом досить скромний рівень критичного осмислення технологій, характерний для тих часів[14]. Залишається тільки сподіватися, що до моменту, коли в нас з'являться можливості створити ШІ, ми не тільки матимемо технічну змогу призупинити вибух інтелекту, а й володітимемо інструментами, які допоможуть цей вибух пережити.

Але перш ніж починати думати про майбутнє, корисно буде поглянути, що ми маємо в царині ШІ зараз.

Між надією і розпачем

Улітку 1956 року в Дартмутському коледжі десятеро науковців, які вивчали нейронні мережі, інтелект і розробляли теорію автоматів, зібралися для участі в науковій програмі, що тривала шість тижнів. Саме на Дартмутському семінарі штучний інтелект виділили як окрему галузь наукових досліджень. Із тих науковців, які брали участь у семінарі, потім багато хто став засновником різних наукових напрямів. Учасники, всі як один, із оптимізмом дивилися в майбутнє, і їхній настрій відображено в поданні, направленому до Фонду Рокфеллера, який мав фінансувати семінар:

> *Подаємо пропозицію на двомісячний семінар із вивчення штучного інтелекту для десяти учасників. (...) Дослідження базуватиметься на гіпотезі, що здатність навчатися і будь-яку іншу властивість інтелекту можна описати достатньо точно, щоб можна було створити машину, яка симулюватиме цей процес. Учасники обміркують можливість проектування машини, здатної говорити, формувати абстрактні поняття, розв'язувати задачі, які нині під силу лише людському розуму, а також удосконалювати саму себе. Ми віримо, що наблизитися до виконання одного або кількох із цих завдань можливо, за умови проведення протягом літніх канікул*

спеціалізованого наукового заходу за спільної участі обраних науковців.

За шість десятиліть, які минули від цього сміливого старту, наука про штучний інтелект встигла пережити не одну хвилю піднесення й великих сподівань — а також розчарувань і відступів.

Перший період захвату від досягнень, який розпочався Дартмутським семінаром, Джон Маккарті (організатор зустрічі) назвав ерою «ось-вам-а-ви-не-вірили». Тоді, на зорі розвитку науки про штучний інтелект, дослідники своїми розробками систем намагалися втерти носа критикам, які заявляли: «Жодна машина не здатна зробити *X*!». Такі скептичні твердження лунали тоді досить часто. І науковці створювали системи, які робили *X* у «мікросвіті» (чітко визначених обмежених середовищах, де можна було спрощено продемонструвати виконання завдання). Так науковці доводили спроможність ШІ й показували, що машина фактично здатна зробити *X*. Одна з таких ранніх систем, «Логічний теоретик» (Logic Theorist), могла доводити теореми з другого розділу Principia Mathematica Вайтгеда й Рассела, і одне з машинних доведень було навіть значно елегантнішим за оригінальне. Так було розвінчано стереотип, що машини можуть «мислити тільки цифрами», і показано, що вони здатні застосовувати дедукцію й знаходити логічні доведення[15]. Іще одна програма, «Розв'язання загальних задач» (General Problem Solver), могла розв'язувати широкий спектр формально визначених задач[16]. З'явилися також програми, які могли розв'язувати задачі на обчислення (такі, з якими зазвичай працювали студенти першого курсу коледжу), задачі на пошук візуальних аналогій (як у тестах на визначення IQ), а також прості вербальні алгебраїчні задачі[17]. Робот Трусько (Shakey) (який отримав своє ім'я тому, що тремтів під час виконання операцій) демонстрував, як логічні доведення можна поєднати з фізіологічними реакціями, і вмів планувати й контролювати фізичні дії[18]. Програма «Еліза» (ELIZA) демонструвала, як комп'ютер може імітувати психотерапевта[19]. У середині 1970-х було розроблено програму SHRDLU — симульована рука робота в симульованому світі геометричних блоків могла виконувати накази й

відповідати англійською на питання, які користувач друкував на комп'ютері[20].

Потім було створено кілька програм, які писали музику в стилі різних класичних композиторів; ставили точніші діагнози, ніж молоді лікарі, керували автомобілями й створювали винаходи, що підлягали патентуванню[21]. Була навіть система ШІ, яка вигадувала оригінальні жарти[22]. (Якщо чесно, не дуже смішні: «Що буде, як схрестити *оптику* з *кондитерською справою*? Ш*око*лад» — а втім, є свідчення, що дітям ці жарти подобалися[1]).

Але методи, які добре працювали в ранніх системах і були по суті демонстраційними, було важко поширити на весь спектр такого типу задач або застосувати до складніших випадків. Одна з причин — «комбінаторний вибух» можливих варіантів, вибирати з яких доводиться методом повного перебору. Наприклад, щоб довести теорему, доведення якої є дедуктивним ланцюжком із п'яти тверджень з одним правилом висновування і п'ятьма аксіомами, можна просто перелічити 3125 можливих комбінацій і по черзі перевірити кожну: чи приводить та до очікуваного результату. Вичерпний пошук можна застосувати і для доведень із ланцюгом із шести чи семи тверджень. Але що складнішим стає завдання, то важче застосувати повний перебір. Адже у випадку теореми, доведення якої складається з 50 послідовних умовиводів, треба витратити часу не в десять разів більше, ніж для доведення з 5: якщо використовувати вичерпний пошук, доведеться обробити $5^{50} \approx 8{,}9 \cdot 10^{34}$ можливих результатів. З таким завданням не впорається навіть найшвидший суперкомп'ютер.

Щоб подолати труднощі, що виникають унаслідок «комбінаторного вибуху», потрібен алгоритм, який зможе зрозуміти природу цільової галузі й використати всі переваги цього знання через евристичний пошук, планування і гнучкі абстрактні представлення — такого ранні системи ШІ не вміли. Також вони часто бували неефективні через малий арсенал методів у роботі з невизначеністю, недостатньо продумані й коректні символічні записи, малу кількість даних. Існували й технічні обмеження: для оброблення певних завдань просто не вистачало пам'яті та продуктивності процесора. До середини 1970-х науковці почали розуміти, що проблеми справді серйозні, і багато які проекти в галузі ШІ просто неможливо буде

здійснити. Після цього настала перша «зима штучного інтелекту» — період згортання програм: фінансування ставало дедалі меншим, а скептицизм усе більшим. Штучний інтелект виходив з моди.

Нова «весна» розпочалася на початку 1980-х, коли в Японії запустили Програму комп'ютерів п'ятого покоління, щедро профінансовану державою і приватними спонсорами. Її мета була дуже амбітна: здійснити прорив у галузі розроблення комп'ютерів (забезпечити паралельну роботу великої кількості мікропроцесорів). Суперкомп'ютер мав стати платформою для роботи штучного інтелекту. Програму запустили якраз тоді, коли світ зачудовано спостерігав за японським повоєнним «економічним дивом». Західні очільники влади й бізнесу гарячково намагалися вивести формулу економічних успіхів Японії, бо хотіли відтворити їх у себе. Тож коли японці вирішили спрямувати серйозні інвестиції у ШІ, деякі інші країни теж це зробили.

У наступні кілька років відбувся справжній розквіт «*експертних систем*». Це були програми, основані на виконанні наборів правил і розроблені спершу для допомоги людям, які ухвалюють рішення. Вони базувалися на людських знаннях і досвіді, вручну описаних формальною мовою. Було створено сотні таких експертних систем. Але малі системи приносили мало користі, а на створення, перевірку й оновлення великих доводилося витрачати занадто багато ресурсів, до того ж їх було складно застосовувати. Рішення кинути всі потужності цілого комп'ютера тільки на виконання однієї програми здавалося не дуже практичним. До кінця 1980-х весняний розквіт помалу зів'яв.

Програма розробки комп'ютерів п'ятого покоління так і не досягла мети, ні в Японії, ні в західних країнах. Настала друга зима штучного інтелекту. Один критик справедливо оплакував «історію досліджень ШІ, які досягали успіху тільки в дуже обмежених царинах, а більших цілей, до яких, здавалося, було рукою подати від перших перемог, так і не вдавалося досягти»[23]. Приватні благодійники сахалися, як вогню, будь-яких проектів, пов'язаних зі штучним інтелектом. Навіть науковці уникали цього словосполучення в розмовах зі своїми меценатами[24].

А втім, техніка розвивалася й далі, і в 1990-х крига другої зими ШІ поволі скресла. Оптимізм повернувся разом із новими

інтелектуальними підходами, які, здавалося, мали стати альтернативою традиційній логічній парадигмі (її часто називали «Старим добрим штучним інтелектом», СДШІ, англійською — GOFAI). Вона полягала у високорівневому оперуванні символами, і її золота доба припала на 1980-ті, час розроблення експертних систем. Очікували, що за допомогою нових популярних методів, зокрема нейронних мереж і генетичних алгоритмів, будуть усунені деякі слабкі сторони СДШІ, наприклад, їхня «крихкість» (якщо програміст робив бодай найменше помилкове припущення, програма видавала цілковиту нісенітницю). Нові методи працювали органічніше. Наприклад, нейронні мережі могли похвалитися «стійкістю до відмов»: якщо траплялися незначні пошкодження, це спричиняло лише невелике погіршення якості, а не повний крах. Але важливіше було те, що нейронні мережі могли вчитися на досвіді, природним способом виводячи узагальнення з окремих випадків, а також знаходячи неявні статистичні залежності у вхідних даних[25]. Тож мережі могли ефективно розв'язувати задачі класифікації та пошуку закономірностей. Наприклад, нейронну мережу вчили розпізнавати набір звукових сигналів, і після цього вона вміла розрізняти акустичні особливості підводних човнів, мін, а також морських мешканців. Водночас така система була точнішою, ніж фахівці-люди, і для цього не треба було наперед прописувати, як розрізняти категорії, а також якої ваги надавати різним характеристикам.

Прості нейронні моделі були розроблені ще наприкінці 1950-х, але справжній розквіт у цій галузі відбувся, коли в роботі з нейронними мережами застосували метод зворотного поширення помилки. Після цього багатошарові нейронні мережі були здатні до навчання[26]. Такі багатошарові мережі, у яких між входом і виходом є один або більше внутрішніх («прихованих») шарів нейронів, можуть навчитися виконувати набагато більше функцій, ніж попередні простіші мережі[27]. А оскільки комп'ютери стали потужнішими, інженери почали розробляти справді функціональні нейронні мережі, які могли приносити реальну практичну користь у різних галузях.

Нейронні мережі мали властивості, співвідносні з мозком, і це вигідно відрізняло їхню діяльність від строго логічної, але вразливої роботи систем СДШІ. На честь цього навіть з'явився новий «-ізм» —

конекціонізм, підхід, що наголошував на важливості паралельної підсимвольної обробки даних. Відтоді про штучні нейронні системи опублікували більше ніж 150 000 наукових праць. Конекціонізм досі залишається важливим підходом у галузі машинного навчання.

Розтопити сніг другої зими штучного інтелекту допомогла також поява ще одного підходу еволюційних методів, а саме: генетичного алгоритму й генетичного програмування. Вони мали, можливо, трохи меншу наукову цінність, ніж нейронні мережі, але теж стали широковідомими. Еволюційні моделі працюють так: будується популяція потенційних кандидатів у розв'язки (це можуть бути структури даних або програми), і далі на їхній основі генеруються нові кандидати за допомогою випадкового мутування або розмноження наявних. Періодично популяцію проріджують, установлюючи критерій відбору (функція виживання), і в результаті лише найкращі кандидати залишаються в наступній популяції. Така процедура повторюється для кількох тисяч поколінь, і в середньому якість розв'язків у популяції поступово підвищується. У сферах, де такі алгоритми застосовні, вони можуть знаходити ефективні розв'язки для дуже широкого спектра задач — розв'язки, на диво, новаторські й неінтуїтивні. Часто вони більше схожі на природні структури, ніж на речі, створені інженерами-людьми. У теорії, генетичні алгоритми можуть працювати майже без втручання людини — після того, як для них буде задано функцію виживання. Але на практиці, для того щоб еволюційні методи добре працювали, їхні розробники мусять докласти багато зусиль, умінь і винахідливості, особливо для чіткого визначення формату представлення даних. Адже якщо кандидати будуть закодовані неправильно (за допомогою генетичної мови, яка описує латентну структуру цільової галузі), еволюційний пошук перетвориться на нескінченне блукання у просторі можливих розв'язків або застрягне в локальному оптимумі. Але навіть із правильним форматом представлення даних еволюційний метод потребує великого обсягу обчислень і стає значно менш ефективним внаслідок комбінаторного вибуху.

Нейронні мережі та генетичні алгоритми — це ті методи, які обнадіювали в середині 1990-х, адже, здавалося, пропонували альтернативу застарілій парадигмі СДШІ. Але я пишу про них не для

того, щоб співати їм дифірамби й підносити над іншими техніками машинного навчання. Тому що насправді найбільший теоретичний прорив останніх двадцяти років — розуміння, що незіставні на перший погляд методи можна розглядати як часткові випадки в межах спільної математичної парадигми. Наприклад, багато видів штучних нейронних мереж можна вважати класифікаторами, які виконують певний тип статистичних розрахунків (наприклад, оцінку максимальної правдоподібності)[28]. Із цього погляду нейронні мережі можна порівняти з ширшим класом алгоритмів побудови класифікаторів на основі даних: «деревом рішень», «логістичною регресією», «методом опорних векторів», «наївним Баєсом», «методом k найближчих сусідів» та інших[29].

Також генетичні алгоритми можна вважати алгоритмами стохастичного сходження на вершину, а отже, підрозділом ширшого класу алгоритмів оптимізації. Кожен із таких алгоритмів побудови класифікаторів або пошуку простору рішень має власні сильні та слабкі сторони, які можна описати математично. Ці алгоритми висувають різні вимоги до процесорного часу й обсягу пам'яті комп'ютера, до свого початкового стану, різняться в тому, наскільки легко вони можуть працювати із зовнішніми даними і наскільки зрозуміла їхня робота для людини.

За всією цією кухнею машинного навчання та креативного розв'язання задач стоять конкретні математичні компроміси. Ідеал — досконалий баєсів механізм, який найоптимальніше (у ймовірнісному сенсі) обробляє наявну інформацію. Цього ідеалу досягти неможливо, бо він ставить надто високі вимоги до фізичних параметрів машини, на якій працюватиме (див. додаток 1). Тому штучний інтелект можна вважати таким собі квестом із пошуку найкращої спрощеної версії: Баєсів ідеал зводиться до стану, у якому зберігається висока ефективність у потрібній сфері досліджень, але певною мірою втрачається оптимальність роботи чи здатність до узагальнення.

Додаток 1. Оптимальний баєсів агент

Ідеальний баєсів агент, насамперед, має функцію «апріорної ймовірності», тобто здатен приписувати ймовірності кожному «можливому світу» (тобто максимально конкретному стану, у

якому може опинитися світ)[30]. Ця апріорна ймовірність містить індуктивне зміщення так, що простіші можливі світи мають вищу ймовірність появи. (Один зі способів формально визначити простоту ймовірного світу — скористатися критерієм «колмогоровської складності», який оцінює довжину найкоротшої комп'ютерної програми, що генерує повний опис світу)[31]. Також апріорна ймовірність враховує все попереднє знання, що його програмісти захочуть передати агенту.

Коли агент отримує нову інформацію із сенсорів, він оновлює ймовірнісний розподіл, обумовлюючи його відповідно до нової інформації, згідно з теоремою Баєса[32]. Обумовлення — це математична операція, яка встановлює нульову ймовірність світам, які не узгоджуються з отриманою інформацією, і нормує імовірності для решти світів. Результатом є «апостеріорний розподіл імовірностей» (що його механізм може використати як апріорний у наступній ітерації). Після кожного нового спостереження ймовірнісна міра перерозподіляється між дедалі меншою кількістю світів, які не суперечать спостереженням. Серед них прості світи завжди мають більшу ймовірність існування. Результатом є «апостеріорний розподіл імовірностей» (що його агент може використати як апріорний у наступній ітерації).

Щоб краще зрозуміти цей процес, уявімо тонкий шар піску на великому аркуші паперу. Аркуш розкреслено на різного розміру ділянки, що відповідають можливим світам, а пісок рівномірно розподілено на поверхні паперу. Із кожним новим спостереженням ми забираємо пісок зі «світів», які на цьому етапі забираються, і розподіляємо порівну між іншими ділянками. Так загальна кількість піску не змінюється, з кожним новим спостереженням пісок лише збирається на все меншій площі. Це і є найпростіший прототип процесу навчання. (Щоб визначити ймовірність гіпотези, треба лише виміряти кількість піску в тих «світах», у яких ця гіпотеза істинна).

Отже, ми визначили критерій навчання. Щоб отримати агента, також потрібне правило ухвалення рішення. Для цього агента наділяють «функцією корисності», яка ставить у відповідність числове значення кожному можливому світу. Це значення

відображає міру бажаності світу з погляду критеріїв корисності агента. Тепер на кожній ітерації агент виконує дію, що має максимальну очікувану корисність[33]. (Щоб знайти її, потрібно визначити множину всіх можливих дій, і в кожній з них перерахувати розподіл імовірностей, за умови, що вона відбулася. Тоді корисність кожної дії можна обчислити як суму добутків умовних імовірностей світів і значень їхньої корисності[34]).

Правило навчання та правило ухвалення рішення разом формують «поняття оптимальності» агента. (Його широко використовують науковці в галузі штучного інтелекту, епістемології, філософії науки, економіки та статистики[35]). Однак побудувати такий алгоритм на практиці неможливо, адже для його роботи необхідно виконати значний обсяг розрахунків. За будь-якої спроби їх здійснити відбувається комбінаторний вибух, як у СДШІ. Щоб зрозуміти причину, уявіть невелику підмножину можливих світів: комп'ютерний монітор, що пропливає в порожнечі безкінечного вакууму. Монітор має 1000 × 1000 пікселів, кожен із яких завжди ввімкнений або вимкнений. Навіть ця підмножина виявляється неймовірно великою, адже $2^{(1000 \times 1000)}$ можливих станів цього монітора кількісно перевищують усі обчислення, які будь-коли можуть бути виконані у відомому нам Всесвіті. Тому ми не можемо навіть просто перелічити цю підмножину можливих світів, а тим паче провести будь-які розрахунки над кожним із них.

Утім розрахунок оптимальності, незважаючи на практичну нездійсненність, має цінність із теоретичного погляду. Він пропонує стандартну модель, відносно якої можна оцінювати евристичні наближення, іноді міркуючи, як діяв би цей агент в окремих випадках. Альтернативні методи розрахунку оптимальності для штучних агентів буде розглянуто в розділі 12.

Кілька останніх десятиліть науковці працюють саме в цьому напрямі, наприклад, під час розроблення імовірнісних графових моделей, зокрема баєсової мережі. Баєсові мережі стисло репрезентують імовірнісні та умовні незалежні зв'язки, що існують у певних галузях досліджень. (Такі зв'язки необхідно використовувати для подолання

комбінаторного вибуху — проблеми, спільної для ймовірнісного методу і для логічної дедукції). Також вони допомагають краще зрозуміти поняття причинності[36].

Завдяки встановленню зв'язку між загальним завданням баєсового висновування та завданнями з навчання в конкретних сферах, поява нових досконаліших алгоритмів реалізації баєсового висновування одразу дає нагоду покращити реалізації похідних алгоритмів в інших сферах. Наприклад, досягнення в реалізаціях методу Монте-Карло для апроксимації відразу було застосовано в комп'ютерному аналізі зображень, робототехніці, обчислювальній геноміці. Іще одна перевага — можливість об'єднувати й узагальнювати результати досліджень із різних наукових галузей. Графічне моделювання та баєсова статистика перебувають у центрі уваги науковців із найрізноманітніших сфер, як-от машинне навчання, статистична фізика, біоінформатика, комбінаторна оптимізація та теорія комунікації[37]. Розвиток машинного навчання останніми роками відбувається ще й завдяки залученню результатів із інших галузей науки. (Цьому також сприяла поява швидших комп'ютерів і наявність великих обсягів даних).

Штучний інтелект сьогодні

Уже зараз штучний інтелект обійшов людину в багатьох сферах. У таблиці 1 згадано деякі ігрові комп'ютери та програми. Як видно, у деяких іграх ШІ вже переміг людину[38].

Зараз нас це не дуже дивує, але насправді, тільки тому, що сьогодні досягненнями штучного інтелекту людей уже складніше вразити. Колись вважали, що першість у шахах — це вершина людського інтелекту. У кінці 1950-х років експерти писали: «Той, хто створить успішну машину для гри в шахи, проникне до самих основ людського інтелекту»[39]. Зараз так би вже ніхто не сказав. Як тут не згадати слова Джона Маккарті, який нарікав: «Щойно ШІ запрацює, ніхто більше не називатиме його штучним інтелектом»[40].

Утім тут варто назвати важливу причину, чому виявилося, що комп'ютер, який добре грає в шахи, уже не є аж таким великим досягненням. Раніше всі вважали (можливо, не без підстав): для гри в шахи на рівні гросмейстера необхідно мати високий *загальний*

інтелект[41]. Усі думали, що шахи потребують абстрактного мислення, стратегічного планування, вміння будувати гнучкі стратегії та складні логічні судження, можливо, навіть моделювати мислення суперника. Але виявилося, що можна створити цілком успішний шаховий комп'ютер на основі вузькоспеціального алгоритму[42]. Коли наприкінці XX століття комп'ютерні процесори досягли певного рівня розвитку,- рівень майстерності цього алгоритму став дуже пристойний. Але ШІ такого типу обмежений. Усе, що він може, — це грати в шахи[43].

Таблиця 1. **Ігровий ШІ**

Гра	**Рівень**	**Деталі**
Шашки	Перевершив людину	Програма, яку написав Артур Семюел у 1952 році, а потім удосконалив (1955 року вона отримала можливість навчатися), стала першою програмою гри в шашки, що грала краще за свого творця[44]. 1994 року програма CHINOOK перемогла тодішнього чемпіона. Тоді комп'ютер уперше переміг в офіційному ігровому чемпіонаті світової першості. А 2002 року Джонатан Шейфер із колегами «розв'язали» шашки, тобто створили комп'ютерну програму, яка завжди робить найкращий хід (на основі альфа-бета пошуку в базі даних із тридцяти дев'яти трильйонів можливих фінальних позицій). Сама з собою така програма завжди грає внічию[45]
Нарди	Перевершив людину	1979 року програма Ганса Берлінера BKG перемогла чемпіона світу. Це була перша в історії перемога програми над чемпіоном світу з ігор. Проте пізніше Берлінер твердив, що комп'ютеру просто пощастило[46]. 1992 року програма Джеррі Тезауро TD-Gammon досягла майстерності рівня чемпіона завдяки алгоритму навчання на базі методу часових відхилень (тип навчання із підкріпленням) і постійній грі проти самої себе[47]. Відтоді програми гри в нарди набагато випередили людей[48]
Traveller TCS	Перевершив людський рівень у команді з людьми[49]	Програма Дугласа Лената Eurisco 1981 та 1982 років виграла першість Сполучених Штатів із Traveller TCS (футуристична військово-морська гра). Комп'ютерна програма використала таку незвичну стратегію, що творці гри змушені були змінити правила[50]. Eurisco використовувала евристику для побудови флоту та евристику для зміни евристики

Гра	Рівень	Деталі
Реверсі	Перевершив людину	1997 року програма Logistello виграла всі шість турів гри проти тодішнього чемпіона світу Такеші Муракамі[51]
Шахи	Перевершив людину	1997 рік: шаховий комп'ютер Deep Blue переміг чемпіона світу Гаррі Каспарова. Каспаров стверджував, що бачив проблиски справжнього розуму та творчості в деяких ходах комп'ютера[52]. Відтоді шахові програми розвивалися далі[53]
Кросворди	Рівень експерта	1999 рік: програма розв'язування кросвордів Proverb показує вищий за середній результат[54]. 2012 рік: програма Dr. Fill Мета Гінсберга виходить у першу четвірку американського турніру з кросвордів. (Dr. Fill показує чудові результати в найважчих завданнях, але зазнає невдачі в нестандартних випадках, коли треба написати слово навпаки або по діагоналі)[55]
Ерудит	Перевершив людину	Від 2002 року програми гри в «Ерудит» перемагають гравців-людей[56]
Бридж	На рівні з найкращими гравцями	Із 2005 року програми гри у бридж грають на рівні з найкращими гравцями-людьми[57]
Jeopardy!	Перевершив людину	2010 рік: суперкомп'ютер IBM Watson переміг двох найкращих гравців усіх часів Кена Дженнінгса та Бреда Раттера[58]. Jeopardy! — це телевікторина із завданнями в галузі історії, літератури, спорту, географії, поп-культури, науки тощо. Завдання подають у формі натяку, і часто використовують гру слів
Покер	Різний	У варіанті Full-Ring Texas Hold'em комп'ютер трохи не дотягує до рівня найкращих гравців-людей, але в деяких інших варіантах гри перемагає[59]
FreeCell	Перевершив людину	Для створення програми для гри в пасьянс FreeCell (яка загалом є NP-повною задачею), здатної перемогти найкращих гравців-людей, було використано евристику, покращену за допомогою генетичних алгоритмів[60]

Гра	Рівень	Деталі
Го	Дуже сильний початковий рівень	Станом на 2012 рік покоління програм гри в го під назвою «Дзен» досягло шостого дану у швидкій грі (рівень сильного початківця). Ця програма використовує метод Монте-Карло (пошук по дереву) та алгоритми машинного навчання[61]. Раніше програми гри в го покращували свій рівень на один дан за рік, тож із такими темпами можуть вийти на один рівень з людиною років через десять

Досягти подібних результатів в інших сферах виявилося набагато складніше. Науковець у галузі комп'ютерної техніки Дональд Кнут зазначив, що «ШІ зараз може майже все, що вимагає "думання", але не може майже нічого, що люди і тварини роблять "не замислюючись" — досягнути цього виявилося набагато важче![62]». Візуальний аналіз, розпізнавання об'єктів, керування поведінкою робота під час взаємодії з природним середовищем — це, як виявилося, досить складні завдання. Але прогрес триває, тим більше, що апаратне забезпечення також постійно розвивається.

Дати комп'ютеру загальні знання про світ і навчити розуміти мовлення — теж виявилося нелегко. Часто тепер ознакою справжньої «інтелектуальності» ШІ вважають саме здатність розуміти мову на рівні з людиною — реалізація цієї функції вважається найскладнішим завданням у створенні по-людськи розумних машин[63]. Якби хтось навчив ШІ розуміти мовлення не гірше за дорослу людину, то розробити ШІ, здатний робити все інше, стало б якщо не легко, то принаймні нескладно[64].

Уміння грати в шахи, як виявилося, можна вкласти в нескладний алгоритм. Отже, можна припустити, що алгоритми, здатні робити інші речі — наприклад, міркувати або навіть складати комп'ютерні програми — будуть не такими вже й складними. Точно можна сказати одне: те, що зараз роблять складні програми, колись, можливо, стане під силу значно простішим. Можливо, їх просто ще не винайшли. Космогонічна система Птолемея (та, де земля була в центрі, а Сонце, Місяць, планети й зірки оберталися навколо) була останнім словом в астрономії понад тисячу років. Її століттями вдосконалювали, покращували точність прогнозування, додавали епіцикл за епіциклом,

щоб пояснити рух космічних тіл. А потім просто відкинули на користь геліоцентричної системи Коперника, яка була простішою та більш прогностично точною (проте тільки після вдосконалення Кеплера)[65].

Зараз засоби ШІ застосовують у такій кількості сфер, що й не перелічити. Наведу лише кілька прикладів. Окрім ігрових алгоритмів, про які згадано в таблиці 1, існують слухові апарати, які фільтрують сторонні шуми; навігатори з автоматичним прокладанням маршруту; рекомендаційні системи, що допомагають вибирати книжки й музику за історією попередніх покупок та оцінок; системи, що допомагають лікарям ухвалювати рішення під час діагностування хвороб, призначати лікування, інтерпретувати результати досліджень. Існують роботи-домашні улюбленці, роботи-прибиральники, роботи-газонокосарки, роботи-рятувальники, роботи-хірурги та більше ніж мільйон промислових роботів[66]. Популяція роботів на Землі вже досягла десяти мільйонів[67].

Сучасні системи розпізнавання мовлення, що базуються на методах статистичного аналізу, як-от прихована марковська модель, досягли точності, достатньої для практичного застосування (деякі фрагменти цієї книжки записано за допомогою програми розпізнавання мовлення). Особисті цифрові помічники, на кшалт Сірі компанії Apple, здатні виконувати голосові команди й відповідати на прості запитання. Засоби розпізнавання друкованого та рукописного тексту давно застосовують у сортуванні пошти й оцифруванні документів[68].

Засоби машинного перекладу, хоч і недосконалі, та їх уже можна застосовувати для деяких завдань. Раніше такі системи йшли шляхом СДШІ: великі колективи фахових лінгвістів вручну наповнювали програми на основі правил і словників для кожної мови окремо. Зараз використовують алгоритми машинного навчання, які, аналізуючи великі обсяги можливих варіантів тексту, автоматично будують статистичні моделі вживання слів. Механізм підбирає параметри роботи моделей на основі аналізу корпусів паралельних текстів. Так участь лінгвістів зводиться до мінімуму: програмістам, які створюють такі системи, не потрібно знати всі мови, з якими працює їхня програма[69].

Можливості систем розпізнавання облич останнім часом демонструють таке дивовижне зростання, що їх уже використовують в

автоматизованих пунктах перетину кордону в Європі й Австралії. Державний департамент США для видачі віз використовує систему розпізнавання облич з базою даних, що налічує більше ніж 75 мільйонів фотографій. Системи спостереження використовують усе складніші алгоритми ШІ та технології виявлення інформації в даних (data-mining) для аналізу величезних обсягів голосових повідомлень, відеопотоків, текстів, перехоплених із комунікаційних каналів і збережених у гігантських дата-центрах.

Програми для доведення теорем та розв'язання рівнянь є настільки звичною справою зараз, що їх уже й не вважають ШІ. Засоби розв'язання рівнянь вбудовано в програми для наукових розрахунків, як-от Mathematica. Формальні методи перевірки, які містять автоматизоване доведення теорем, використовують виробники мікрочипів для детальної симуляції роботи схем перед запуском у виробництво.

Американські військові й розвідувальні установи торують шлях до масштабного виробництва роботів-саперів, розвідувальних і бойових дронів та інших безпілотних засобів. Звісно, усі вони працюють переважно з ручним дистанційним керуванням, але науковці постійно намагаються розширити їхню автономність.

Вражають також досягнення у сфері автоматизованого планування. Наприклад, програма автоматизованої оптимізації та планування постачання DART, яку вперше застосували в операції «Буря в пустелі» 1991 року. За словами представників DARPA (the Defence Advanced Research Projects Agency in the US — Агентство передових оборонних дослідницьких проектів США), лише вона одна принесла вже більше доходу, ніж було інвестовано в проекти ШІ за попередні тридцять років[70]. Сервіси бронювання авіаквитків використовують складні системи планування й визначення ціни. Засоби ШІ широко використовують у промислових системах управління запасами. Також різні служби часто користуються системами-автовідповідачами телефонних замовлень із голосовим меню та розпізнаванням мовлення, які потім водять нещасних клієнтів заплутаними лабіринтами своїх опцій.

Багато сервісів у всесвітній павутині працюють завдяки ШІ. Програми фільтрують потоки електронних листів цілого світу. Незважаючи на вигадливість спамерів, які намагаються обійти

блокування, баєсові фільтри успішно стримують натиск хвиль небажаної пошти. Програми із ШІ дозволяють чи відхиляють платіжні транзакції за кредитними картками. А також постійно відстежують активність банківських рахунків, щоб запобігти шахрайству. Системи пошуку інформації також використовують машинне навчання. Пошукова система Google — мабуть, найвизначніша створена система ШІ.

Варто сказати, що чіткого розділення між штучним інтелектом і просто програмним забезпеченням немає. Серед згаданих застосунків багато можна вважати просто програмним забезпеченням, а не ШІ. Однак як тут не згадати тезу Маккарті: якщо щось працює, його ніхто вже не називає ШІ. Але нам буде краще розділяти системи з конкретними вузькими інтелектуальними можливостями (незалежно від того, чи називають їх «ШІ», чи ні) та системами, які можна застосовувати для ширшого спектра проблем. Усі системи, що зараз використовують, є, по суті, системами першого типу. Проте часто вони містять компоненти, які можна безпосередньо використати в штучному інтелекті або які знадобляться для його створення. Наприклад, алгоритми класифікації, пошуку, планування, розрахунків і системи представлень.

Ще одна сфера застосування ШІ: світовий фінансовий ринок, середовище з високою конкуренцією та високими ставками. Автоматизовані фондові торгові системи використовують усі великі інвестиційні компанії. Деякі з них лише спрощують фондовому менеджеру процедури купівлі або продажу, тоді як інші автономно діють у межах певної торговельної стратегії й адаптують її до мінливих умов ринку. Фінансові аналітичні системи застосовують різноманітні засоби, щоб отримувати інформацію з даних, аналізувати часові ряди для виявлення патернів і тенденцій у коливаннях ринків цінних паперів або виявляти зв'язок між змінами ринкових цін і появою тих чи тих ключових слів у новинах. Для цього служби новин навіть надають спеціально відформатовані платні рядки фінансових новин. Деякі системи допомагають знаходити можливості арбітражу на ринку чи між ринками або дають змогу вести високочастотну торгівлю на швидких цінових коливаннях (коли відлік часу міжцінових змін іде на мілісекунди і навіть затримки передавання інформації в мережі мають

значення, тому такі системи розміщують якнайближче до торгового майданчика). Більш ніж половину всіх акцій, що продаються на біржах США, купують такі алгоритмічні високочастотні трейдери[71]. Алгоритмічна торгівля стала однією з причин обвалу американського ринку акцій, що має назву «Миттєвий крах-2010» (2010 Flash Crash) (див. додаток 2).

Додаток 2. «Миттєвий крах-2010»

До кінця обіду 6 травня 2010 року через Європейську боргову кризу американські ринки акцій уже просіли на чотири відсотки. О 14:32 продавець (комплекс взаємних фондів) ініціював алгоритм продажу великого обсягу ф'ючерсів E-mini S&P 500 з прив'язкою курсу до похвилинного рівня ліквідності на біржі. Ці контракти викупили високочастотні алгоритмічні торговці, які відразу перепродували їх, оскільки були запрограмовані позбавлятися довгих позицій. Поки попит основних покупців падав, алгоритмічні торговці почали перепродавати ці ф'ючерси один одному, формуючи ефект «гарячої картоплини» та збільшуючи обсяг торгів. Для алгоритму продавця це було свідченням зростання ліквідності ринку, тож він почав нарощувати швидкість розміщення ф'ючерсів, посилюючи ефект. У якийсь момент високочастотні торговці почали залишати ринок, виводячи ліквідність, тоді як ціни падали й далі. О 14:45 торгівлю E-mini зупинив автоматичний запобіжний механізм, і робота біржі перервалася. За якісь секунди, після перезапуску біржі, торгівля відновилася, ціни стабілізувалися і втрати почали відновлюватися. Але в момент найбільшого падіння ринок утратив близько трильйона доларів, а суми деяких угод виявилися просто абсурдними, наприклад 1 цент або 100 тисяч доларів. Того дня, після закриття торгів, представники бірж зустрілися з регулятором і вирішили скасувати угоди, проведені за цінами, які відрізнялися від докризових більше ніж на 60 відсотків (вважаючи такі транзакції «безсумнівно помилковими» і такими, що можуть бути скасовані пост-фактум)[72].

Переказ цього епізоду тут є відступом, адже програми, через які стався «Миттєвий крах», були не надто інтелектуальні чи складні, а

загрози, створені ними, принципово відрізняються від проблем, які ми розглядатимемо далі в цій книжці в контексті машинного розуму. Проте з цього можна винести кілька важливих уроків. Найперше: взаємодія між окремими простими компонентами системи може мати складні й непередбачувані наслідки. Із додаванням нових компонентів системні ризики можуть зростати і виявити їх заздалегідь важко, а іноді — неможливо (навіть після відмови системи)[73].

Іншим уроком є те, що фахівець може дати вказівку програмі, спираючись на цілком притомні і зазвичай правильні припущення (наприклад, що обсяг торгів — це хороший показник ліквідності ринку) і програма слухняно й безжально виконуватиме цю вказівку, навіть коли насправді припущення помилкове і це призведе до катастрофічних результатів. Програма просто працює за алгоритмом, навіть коли ми хапаємося за голову в німому заціпенінні від абсурдності результатів її роботи. До цього аспекту ми ще повернемося згодом.

І насамкінець варто зауважити, що хоч причиною «Миттєвого краху» стала автоматизація торгів, саме вона допомогла швидко зупинити й відновити роботу. Розробники системи цілком слушно передбачили, що події можуть почати розгортатися занадто швидко, щоб людина могла втрутитися. Саме тому і було встановлено запобіжний код, який зупинив торги, коли ціна почала падати надто стрімко.

Потреба у вбудованих механізмах захисту в разі непередбачених подій (замість того щоб покладатися на можливість ручного контролю в реальному часі) — ще один аспект проблем, які обіцяє нам машинний розум. Ми ще повернемося до цієї теми згодом[74].

Деякі думки про майбутнє штучного розуму

Дослідження ШІ почали потрохи відвойовувати позиції — завдяки успіхам у двох основних напрямах: зміцненні теоретичного підґрунтя для машинного навчання ідеями з галузей статистики та теорії інформації і створенні практичних та комерційно успішних прикладів використання машинного навчання для розв'язання низки

практичних проблем у деяких важливих сферах. Проте наслідки попередніх невдач усе ще впливають на галузь й утримують авторів прогресивних розробок від надто стрімкого руху до амбітної мети. Нільс Нільссон, ветеран цього напряму, скаржиться на нестачу в сучасних науковців тієї сміливості духу, яка підштовхувала його покоління:

> *Занепокоєння власною «солідністю» спричиняє руйнівний, на мою думку, ефект на деяких дослідників ШІ. Вони кажуть: «ШІ вважають показухою і штукарством. Хоч я і досяг неабияких результатів, проте не хочу мати несолідний вигляд». Результатом такого консерватизму стала зосередженість на «слабкому ШІ», що може лише допомагати людині в її діяльності, а не «сильному ШІ», який покликаний замінити людський розум*[75].

Почуття Нільссона поділили кілька інших засновників, як-от Марвін Мінскі, Джон Маккарті та Патрік Вінстон[76].

Відновлення інтересу до ШІ останніх років може перерости в нові спроби створення справжнього штучного розуму, який Нільссон називає «сильним ШІ». На додаток до швидшого апаратного забезпечення, неабияк посприяти сучасному проекту створення ШІ можуть успіхи в багатьох супутніх до ШІ напрямах, зокрема у сфері розроблення програмного забезпечення взагалі та в обчислювальній нейрології.

Ознакою підвищеної цікавості до знань про ШІ є те, що на безкоштовний вступний онлайн-курс із ШІ у Стенфордському університеті, який організували восени 2011 року Себастьян Трун і Пітер Норвіг, записалося аж 160 тисяч студентів з усього світу (і 23 тисячі успішно його завершили)[77].

Існують різні прогнози про майбутнє ШІ. Немає єдиної думки щодо того, коли та яких форм набуде ШІ. Як зазначено в одному з недавніх досліджень на цю тему, передбачення долі ШІ «певні настільки, наскільки різноманітні»[78].

Поточний розподіл оцінок стосовно майбутнього цієї галузі, на жаль, не відомий, але ми можемо скласти приблизне уявлення з декількох менш масштабних опитувань і суб'єктивних вражень. Зокрема, у таблиці 2 наведено результати одного з нещодавніх опитувань,

пов'язаних із проблемою ШІ професійних спільнот. Під час нього в експертів цікавилися, коли, на їхню думку, буде створено «штучний інтелект рівня людського» (ШІРЛ). А саме: «такий, що може впоратися з більшістю завдань не гірше за пересічну людину»[79]. Зведені та усереднені результати мають такі оцінки: з 10 % ймовірності до 2022 року, з 50 % — до 2040-го, з 90 % — до 2075 року. (Прогнози респондентів базувалися на припущенні, що «наукова діяльність людства продовжуватиметься без відчутних спадів»).

Ці цифри не варто сприймати надто серйозно, адже розмір вибірки надто малий і немає жодних гарантій, що вона є репрезентативною щодо всіх експертів загалом. Проте вони не відрізняються від результатів інших подібних опитувань[80].

А також відповідають прогнозами кількох десятків дослідників пов'язаних з ШІ проблем, які вони висловили в нещодавно опублікованих інтерв'ю. Наприклад, Нільс Нільссон, який давно і плідно працює над проблемами пошуку, планування, представлення знань та робототехніки, є автором підручників з ШІ, нещодавно завершив найповнішу на сьогодні працю з історії розвитку галузі[81]. На питання про терміни створення ШІРЛ він надав такі оцінки[82]:

з імовірністю 10 %: 2030 рік;
з імовірністю 50 %: 2050 рік;
з імовірністю 90 %: 2100 рік.

Таблиця 2. **Коли буде створено штучний інтелект рівня людини[83]**

	10 %	**50 %**	**90 %**
PT-AI	2023	2048	2080
AGI	2022	2040	2065
EETN	2020	2050	2093
Топ-100	2024	2050	2070
Загалом	2022	2040	2075

Таке передбачення професора Нільссона цілком відповідає уявленням, висловленим його колегами в інтерв'ю, але варто зайвий раз підкреслити, що думки дуже відрізняються. Деякі респонденти впевнено передрікали появу ШІРЛ уже в 2020–2040 роках, коли інші

настільки ж впевнено заперечували можливість його створення[84]. Дехто нарікав на нечіткість чи спекулятивність поняття «інтелект людського рівня» або просто відмовлявся озвучувати будь-які кількісні прогнози.

На мою ж думку, розподілу ймовірності часу появи ШІРЛ (усередненому з проведених опитувань) бракує імовірнісної маси в галузі пізніших років. Усього 10 відсотків сумніву в тому, що ШІРЛ з'явиться у 2075 році або навіть у 2100 році (враховуючи те, що «наукова діяльність людства продовжуватиметься без відчутних спадів»), як на мене, замало.

Узагалі прогнозування темпів чи напрямку розвитку власної галузі ніколи не було сильною стороною дослідників ШІ. З одного боку, досягти виконання деяких завдань, — наприклад, гри в шахи, — виявилося досить просто за допомогою нескладних програм, а скептики, які стверджували, що машини не можуть того чи того, помилилися. З іншого, розробники часто недооцінюють труднощі побудови системи, здатної стійко працювати в непередбачуваних умовах реальних завдань, і зазвичай переоцінюють переваги власних розробок.

Згадане опитування містило ще два питання, які нам також будуть цікаві. Одне стосувалося прогнозованого часу, який мине від моменту створення ШІРЛ до створення суперінтелекту. Результати показано в таблиці 3.

У другому питанні йшлося про те, яким буде довгостроковий вплив створення ШІ на розвиток людства. Відповіді на нього узагальнено на рисунку 2.

Таблиця 3. **Скільки триватиме перехід від ШІРЛ до суперінтелекту?**

	До 2 років	До 30 років
Топ-100	5 %	50 %
Загалом	10 %	75 %

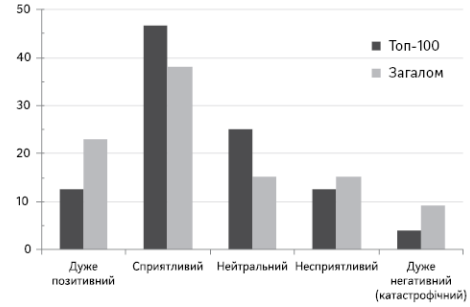


Рисунок 2. **Довгостроковий вплив створення ШІРЛ на розвиток людства**[85]

Моє бачення майбутнього трохи відрізняється. Думаю, після створення штучного інтелекту, рівного людському, поява суперінтелекту не займе багато часу. Щодо наслідків, то мої очікування більш радикальні, — вони радше будуть або дуже добрі, або дуже погані, аніж десь посередині. Хід моїх думок я спробую пояснити в цій книжці.

Через малий розмір вибірки й упередження відбору, але, насамперед, через ненадійність, властиву суб'єктивним судженням, не покладатимемося аж так на ці опитування та інтерв'ю. Вони не дають надійного ґрунту для висновків, натомість штовхають до необґрунтованих суджень. У такий спосіб має цілком переконливий вигляд (за відсутності ліпших джерел інформації), що ймовірність появи штучного інтелекту людського рівня посеред поточного століття достатньо висока, хоч може відбутися як раніше, так і значно пізніше; що невдовзі по тому штучний інтелект, цілком імовірно, перевершить людину; що результати цього можуть бути найрізноманітніші — від найсприятливіших до повного зникнення людства[86]. Такі прогнози варті уважнішого розгляду.

1 В оригіналі: «What do you get when you cross an optic with a mental object? An eye-dea». — *Прим. пер.*

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… Проте припускаю, він буде значно швидшим. Вимірювання успішності за допомогою виживання сильніших відбувається занадто повільно. Натомість завдяки навчанню дослідник зміг би пришвидшити цей процес. Окрім того, важливо розуміти, що він не обмежений випадковістю змін. Щойно він зрозуміє причину певної слабкості, то зможе запровадити зміну, яка б її виправила[90].

Ми знаємо, що сліпа еволюція може привести до появи людського інтелекту, адже щонайменше один раз таке вже відбулося. Тоді еволюція з передбаченням — генетична програма, спроектована і спрямована розумом людини-програміста, — повинна досягти цього ще ефективніше. Так аргументували деякі філософи і науковці, зокрема Девід Чалмерс та Ганс Моравек, свої твердження про те, що створення ШІ не тільки можливе, а й, імовірно, відбудеться вже в цьому столітті[91]. Адже, якщо порівняємо можливості еволюції і здатність людини до створення ШІ, побачимо, що творчий потенціал людини вже зараз у дечому перевершує еволюцію, а незабаром може перевершити і в усьому іншому. Те, що результатом еволюції стало розумне життя, свідчить, що людина теж скоро зможе створити те саме. Тому Моравек писав (власне, ще в 1976-му):

Існування кількох різновидів розумного життя, що з'явилися за таких обмежень, має запевнювати нас, що невдовзі ми теж зможемо досягти того самого. Аналогією цьому є історія польотів об'єктів, важчих за повітря — птахи та кажани демонстрували практичну можливість цього задовго до того, як людство опанувало цю здатність[92].

Однак варто бути обачнішими зі своїми висновками. Справді, унаслідок еволюції важчі за повітря організми полетіли, і люди згодом також це повторили (щоправда, в інший спосіб). Можна навести більше прикладів людського успіху: сонар, магнітна навігація, хімічна зброя, фотосенсори, різного роду механічні й кінетичні властивості. Але так само є приклади сфер, у яких людині не вдалося перевершити еволюцію: морфогенез, самовідновлення, імунітет. Тож аргумент Моравека вже не може «*запевнювати нас*», що «*невдовзі*» ми

обов'язково створимо штучний інтелект, подібний людському. Еволюція розумного життя, хіба свідчить, що таке загалом можливо. Однак ця можливість поки може перебувати далеко за межами людських творчих спроможностей.

Інший спосіб застосування еволюції як аргументу на користь можливості створення ШІ — ідея, що результати виконання генетичного алгоритму на достатньо потужному комп'ютері можуть бути зіставними з результатами біологічної еволюції. Така версія еволюційного аргументу принаймні пропонує конкретний метод створення ШІ.

Але чи скоро досягнуть обчислювальні можливості людства рівня, достатнього для відтворення еволюційного процесу та створення штучного інтелекту? Це залежить від стрімкості розвитку комп'ютерної техніки в найближчі десятиліття і як швидко дії потребуватимуть для своєї роботи генетичні алгоритми з оптимізаційними можливостями, подібними до еволюції природного відбору, якої зазнав наш вид. Так чи інакше ми приходимо до невизначеності, тож спробуємо приблизно уявити, як воно — відтворити еволюцію (див. додаток 3). Це, зрештою, дасть нам змогу визначити, що в нас уже є, а над чим іще варто працювати.

Додаток 3. Що потрібно для відтворення еволюції?

Не кожен етап еволюції людського розуму вартий уваги з погляду застосування еволюційної моделі для створення штучного інтелекту. Не кожен випадок природного відбору впливав на інтелект. Точніше, порівняно невелика кількість еволюційних відборів стосувалася важливих для формування розуму оптимізацій, недоступних людським інженерам. Адже нам не потрібно заново винаходити молекули для зберігання енергії в клітинах, оскільки наші комп'ютери працюють від електричної енергії. А розвиток процесів клітинного метаболізму зайняв значну частину еволюційних ітерацій під час появи й розвитку життя на Землі[93].

Логічно припустити, що для створення ШІ насамперед важлива еволюція нервової системи, яка з'явилася лише менше мільярда років тому[94]. А отже, кількість релевантних еволюційних

«експериментів» стає ще меншою. У світі існує 4–6 · 10^{30} прокаріотів, усього 10^{19} комах, а людей — менше ніж 10^{10} (тоді як доагрокультурні популяції були набагато менші за чисельністю)[95]. Така кількість уже не так лякає.

Проте в еволюційному алгоритмі для відбору потрібні не тільки кандидати, а й функція пристосованості для їхньої оцінки, яка зазвичай і є найбільш складною частиною для обрахунків алгоритму. У разі еволюції штучного інтелекту, щоб оцінити результат відбору, функція пристосованості, імовірно, повинна емулювати розвиток нейронів, навчання, пізнання. Можливо, для кращих результатів нам варто розглядати не загальну кількість організмів зі складною нервовою системою, а кількість нейронів біологічних організмів, які ми маємо емулювати для функції пристосованості. Приблизно порахуємо їх на прикладі комах: їх найбільше серед наземних тварин (самих лише мурах 15–20 відсотків)[96]. Розмір мозку в комах різний. Більші й соціальні комахи мають найбільший мозок: бджола має 10^6 нейронів, дрозофіла — 10^5, а мураха — у середньому 250 000 нейронів[97]. Більшість менших комах має лише близько тисячі нейронів. За середній показник візьмемо кількість нейронів дрозофіли, тоді всі 10^{19} комах світу разом мають 10^{24} нейронів. Додамо ще один розряд за всіх морських ракоподібних, птахів, плазунів, ссавців тощо, і матимемо 10^{25}. (Хоча в доагрокультурний період історії людини вся людська популяція не перевищувала 10^7, із 10^{11} нейронами кожен: маємо всього 10^{18} людських нейронів у світі, але тоді нейрони мали більшу кількість синапсів).

Обчислювальна потужність, яка потрібна для емуляції одного нейрона, залежить від деталізації. Найпростіша модель нейрона для роботи в реальному часі потребує виконання 1000 операцій з рухомою комою на секунду (FLOPS). Повна електрофізіологічна модель Годжкіна — Гакслі потребує 1 200 000 FLOPS. Для детальнішої емуляції нейрона знадобляться ще кілька додаткових нулів у кількості обчислень. Однак абстракції вищого рівня, що моделюють цілі нейронні системи, можуть спростити на пару розрядів обчислення порівняно з найпростішими нейронними моделями[98]. Якщо ми захочемо за один рік емулювати мільярд

років еволюції (більше тривалості існування нервової системи) для 10^{25} нейронів, нам буде потрібна обчислювальна потужність на рівні 10^{31}–10^{44} FLOPS. Для порівняння, найпотужніший (станом на вересень 2013 року) суперкомп'ютер світу, китайський Tianhe-2, забезпечує всього $3{,}39 \cdot 10^{16}$ FLOPS[2]. Останнім часом комп'ютери збільшували свою швидкодію на один розряд FLOPS у середньому за 6,7 року. Застосувавши закон Мура, навіть століття не вистачить для того, щоб комп'ютери досягли потрібної швидкодії. Використання спеціалізованого апаратного забезпечення або триваліша робота алгоритму може додати всього кілька розрядів, але суттєво не вплине на ситуацію.

Але цифра може бути меншою через те, що в еволюції не було мети створити людський розум. Інакше кажучи, функція пристосованості природного відбору враховує не лише розумність і супутні якості[99]. Навіть коли розум справді має перевагу, еволюція може піти не шляхом розумності. Адже кращий розум може мати (і часто таки має) більші потреби в енергії, триваліший час дозрівання, що часом робить вибір розуму недоцільним. У надто жорстоких середовищах користь від розуму теж нівелюється закоротким життям індивіда.

Конкуренція між розумністю й іншими перевагами під час природного відбору знижує шанси неперервного розвитку інтелекту у процесі еволюції. Ба більше, там, де природна еволюція може застрягнути в локальному оптимумі, людина може контролювати роботу еволюційного алгоритму вручну, змінюючи пріоритет між покращенням уже наявних характеристик та пошуком нових або постійно збільшуючи складність тестів на рівень інтелекту[100]. Я вже згадував, що еволюція постійно витрачає час на речі, не пов'язані з інтелектом (як-от перегони Червової Королеви поміж еволюцією імунної системи та еволюцією паразитів). Еволюція раз за разом запроваджує завідомо летальні для індивідів зміни і водночас не має механізмів, які дали б змогу скористатися статистичною подібністю наслідків різних змін для оптимізації результату. В еволюційному алгоритмі створення штучного інтелекту таку неефективність природних механізмів

(принаймні щодо еволюції розуму) можна було б порівняно легко обійти або виправити.

Імовірно, така оптимізація дала б змогу відкинути багато порядків від значення у 10^{31} – 10^{44} FLOPS, що ми отримали раніше. На жаль, невідомо скільки. Поки що неможливо навіть приблизно сказати, було б це п'ять порядків, десять чи двадцять п'ять[101].

Результатом буде те, що для простого копіювання еволюції людського інтелекту наші обчислювальні ресурси критично малі і навіть за умови збереження закону Мура протягом найближчого століття ситуація не поліпшиться (див. рисунок 3). Щоправда, може бути значно ефективніше створити еволюційний алгоритм, який спрямований лише на розвиток розумності, такий собі вдосконалений варіант природного відбору. Але спрогнозувати, як вплинуть на алгоритм такі вдосконалення наразі нереально — чи буде це п'ять розрядів чи двадцять п'ять. Розвинути думку далі неможливо, тож немає можливості аргументовано судити про терміни чи складність створення штучного інтелекту рівного людському еволюційним способом.

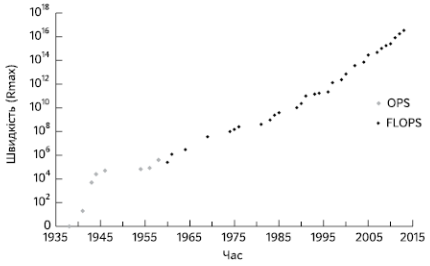


Рисунок 3. **Швидкість суперкомп'ютерів**
Закон Мура у вузькому розумінні полягає у спостереженні, що кількість транзисторів у процесорах останніх десятиліть подвоювалася кожні два роки. Проте частіше цей термін застосовують для позначення того емпіричного факту, що багато характеристик обчислювальної техніки покращуються за схожою експоненційною залежністю від часу. На цій діаграмі зображено залежність максимальної швидкості суперкомп'ютерів від часу (по вертикалі, за логарифмічною шкалою). Останніми роками зростання швидкості послідовних обчислень зменшилося, але розвиток технологій паралелізації дав змогу загальній швидкості суперкомп'ютерів залишитися в межах тренду[102].

Крім того, існують інші аспекти, які не дають нам змоги робити будь-які припущення про складність створення розуму еволюційним способом. По суті, ми маємо утриматися від твердження, що еволюційний процес мав достатньо високі шанси закінчитися появою розумного життя, лише на підставі того, що на Землі це життя

виникло. Воно не враховує упередження відбору, яке неминуче виникає, коли проводиш спостереження на планеті, на якій життя вже виникло. А також наскільки висока імовірність виникнення життя на будь-якій іншій такій планеті. Окрім постійного природного відбору, поява розумного життя завдячує ще й щасливому випадку — аж із 10^{30} планет, на яких коли-небудь могли жити найпростіші організми, лише на одній виникло розумне життя. Отже, для відтворення еволюції розуму, ми можемо опинитися перед необхідністю провести 10^{30} експериментів перш ніж віднайдемо той баланс передумов, за яких така еволюція можлива. Такий хід експерименту цілком відповідає нашому спостереженню, що лише на Землі виникло розумне життя. Частково подолати цю перешкоду пізнання можна лише за допомогою ретельних і складних розрахунків — аналізу випадків конвергенції еволюції певних рис, пов'язаних із рівнем інтелекту, у контексті теорії упередження відбору. А доти ми не можемо відкидати можливість того, що гіпотетична «верхня межа» складності відтворення еволюції, що її ми спробували порахувати в додатку 3, не виявиться на якісь три десятки розрядів більшою[103].

Інший потенційний шлях до створення ШІ — спробувати відтворити людський мозок штучними засобами. Існують різні підходи до виконання цього завдання, які відрізняються детальністю відтворення. З одного боку існує ідея створити емуляцію цілого мозку — детальна імітація — її ми розглянемо в наступному пункті. З іншого, підходи, які пропонують відтворювати деякі принципи роботи мозку, не заглиблюючись у низькорівневі деталі. За допомогою такого арсеналу нейронаука та когнітивна психологія зможуть нарешті пояснити загальні принципи роботи людського мозку. Це водночас допоможе у створенні ШІ. Уже існують нейронні мережі, як приклад комп'ютерних технологій, на створення яких надихнули структури живого мозку. Іще одна ідея, що прийшла в машинне навчання з нейронауки, — ієрархічна перцептивна організація. Навчання з підкріпленням (наприклад, метод часових зсувів), завдяки їхній ролі в когнітивній психології тварин, широко використовують у машинному навчанні та ШІ[104]. І таких випадків застосування в майбутньому з'являтиметься все більше. Завдяки постійному послідовному просуванню у вивченні принципів роботи мозку, рано чи пізно, усі

вони будуть відкриті, скільки б їх не було. Але до того використання вже відкритих принципів поряд із цілком штучними механізмами теж може привести до успішного створення ШІ. Отримана система необов'язково має повторювати мозок, хоча може містити певні подібні до мозку структури.

Доступність такого прототипу, як живий мозок, сам собою прекрасно свідчить про принципову можливість створення штучного розуму. Однак неможливо спрогнозувати термін такого створення, як і передбачити темпи просування в дослідженні мозку. Одне напевне: що далі в майбутнє ми зазираємо, то більша ймовірність, що розгадка секретів людського розуму дасть змогу наблизитися до створення штучного інтелекту.

Фахівці висловлюють різні думки щодо переваг нейроморфного підходу у створенні ШІ перед цілковито синтетичними. Існування птахів підказувало, що політ можливий, але літальні апарати людей не махають крилами. Тож невідомо, чи буде ШІ, як політ, досягнутий штучним способом, чи як вогонь, приручений за допомогою копіювання природних процесів.

Ідею Тюрінга про навчання програми замість початкового закодовування в неї знань можна з однаковим успіхом застосувати як для нейроморфного ШІ, так і для синтетично створеного.

Різновидом машини-дитини Тюрінга є концепт «зерна штучного інтелекту» («seed AI»)[105]. Машина-дитина Тюрінга мала розвиватися, накопичуючи знання — контент, тоді як «зерно ШІ», за задумом, здатне змінювати власну будову. На початкових стадіях через проби і помилки такі зміни будуть радше хаотичними, накопиченням різноманітного досвіду або навіть внесеними програмістами. На більш пізніх етапах «зерно ШІ» повинно розуміти власну структуру настільки, щоб самостійно компонувати алгоритми й пізнавальні стратегії. Передбачають, що для цього «зерну ШІ» доведеться досягти рівня загальної інтелектуальності в низці галузей знань або хоча б мати певний рівень у деяких важливих сферах, як-от комп'ютерна наука чи математика.

Отож у нас є ще один важливий концепт — «рекурсивне самовдосконалення». Успішна модель «зерна ШІ» має вміти покращувати себе — початкова версія може вдосконалювати себе,

створюючи нову розумнішу версію, яка так само може зробити ще розумнішу версію себе і так далі[106]. За певних сприятливих умов такий процес рекурсивного самовдосконалення може тривати достатньо довго, щоб спричинити вибухоподібне зростання інтелекту — швидке нарощування когнітивних здібностей системи від обмежених (стосовно загальних завдань, окрім програмування та розроблення ШІ) до суперінтелектуальних. Ми розглянемо цей перехід і його передумови детальніше в розділі 4. Варто мати на увазі, що така модель створення ШІ непередбачувана: усі спроби можуть зазнавати невдачі, допоки не буде віднайдено який-небудь критичний компонент, завдяки якому зерно ШІ зможе рекурсивно самовдосконалюватися.

Перш ніж перейти до наступного пункту, хочу звернути увагу на ще одну деталь — штучний інтелект необов'язково має повторювати-людський мозок. Він може — і дуже ймовірно, буде — абсолютно іншим. Когнітивні структури більшості створених варіантів ШІ, сильні та слабкі сторони на ранніх етапах розвитку відрізнятимуться від звичних для біологічних організмів (хоч далі ми візьмемо під сумнів можливість ШІ позбуватися власних недоліків). Ба більше, ШІ може по-своєму інтерпретувати мету свого самовдосконалення, не так, як її бачать люди. Наївно вважати, що зростання ШІ буде зумовлене любов'ю, ненавистю, самолюбством чи іншими типово людськими почуттями. Адаптація цих складних концептів потребуватиме додаткових зусиль. Це водночас і небезпека, і можливість. Ми окремо говоритимемо про мотивацію ШІ в наступних розділах, але ця тема лежить у самому центрі проблематики цієї книжки.

Емуляція цілого мозку

Емуляція мозку (відома також як «завантаження свідомості») — це спосіб створення ШІ за допомогою сканування й точного копіювання структур біологічного мозку. Це розв'язання проблеми створення розуму способом, запозиченим у самої природи, — відвертий плагіат. Емуляція мозку потребує виконання низки кроків. Спершу потрібно створити деталізований образ структур конкретного мозку. Для цього треба якось стабілізувати мозок пост-мортем, наприклад, завдяки

вітрифікації (перетворення живих тканин на щось схоже на скло). Далі за допомогою автоматики готуються тонкі зрізи вітрифікованих тканин для сканування електронними мікроскопами. Можливо, знадобиться використати контрасти, щоб підсвітити ті чи ті структурні особливості тканин. Імовірно, для пришвидшення процесу сканування можна виконувати паралельно.

Другим етапом буде трансляція відсканованих зрізів у тривимірну комп'ютерну модель нейронних структур і зв'язків. Власне, цей процес варто виконувати майже одночасно з попереднім, щоб мінімізувати потреби в буферизації об'ємних високодеталізованих зображень. Потім отримані тривимірні нейронні масиви відтворюються з відповідних компонентів комп'ютерної бібліотеки класів нейронних моделей та інших елементів мозкових структур (наприклад, різного роду синаптичних зв'язків). Приклад такого сканування та графічного оброблення за допомогою сучасних обчислювальних засобів поданий на рисунку 4.

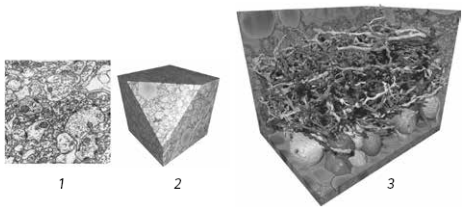


Рисунок 4. **Тривимірне відтворення нейроанатомії зі знімків електронного мікроскопа**
1. Типова мікрофотографія мозкових тканин — дендритів і аксонів.
2. Об'ємне зображення нервової тканини сітківки кроля, отримане за допомогою серійного мікросканування підготованих мікротомом зразків[107]. Окремі двовимірні знімки зроблено з кубічного зразка (довжина ребра приблизно 11 мкм).
3. Просторова реконструкція набору нейронних відростків, що формують нейропіль, згенерована через алгоритм сегментації зображень[108].

На третьому етапі утворені програмні структури реалізуються на достатньо потужному для цього комп'ютері. В ідеалі, ми отримаємо повністю робочу цифрову модель конкретного мозку разом із пам'яттю й особистістю власника. Такий цифровий розум може існувати в комп'ютері як звичайна програма. Його можна помістити у віртуально створене середовище або надати можливості взаємодіяти із зовнішнім світом через додаткові роботизовані апаратні засоби.

Така емуляція мозку не вимагає розуміння, як працюють людська свідомість і пізнання або як запрограмувати штучний інтелект. Потрібно лише ґрунтовно розібратися в деталях роботи складових

мозку. Для успішного результату не потрібно чекати теоретичних відкриттів.

Натомість потрібні деякі розвиненіші прикладні технології. Основні три з них: для сканування — високошвидкісна мікрофотозйомка з високою розділовою здатністю та виділенням важливих деталей; для трансляції — автоматизований аналіз зображення та синтез на його основі відповідних комп'ютерних нейромережевих утворень; виконання — достатньо потужне апаратне забезпечення для роботи отриманої програмної структури (див. таблицю 4). Утім створення віртуального образу або робота із базовими функціями спілкування можливе вже зараз із сучасним рівнем технологій[109]).

Таблиця 4. **Технологічні потреби емуляції мозку**

Сканування	Підготовка зразків / стабілізація		Підготовка мозку до сканування зі збереженням стану та структури тканин
	Маніпуляція		Засоби подання й переміщення мозку і його частин у процесі виконання операцій
	Отримання знімків	Обсяг	Можливість ефективного сканування всього мозку за достатньо малий проміжок часу
		Роздільна здатність	Забезпечення достатньої для відтворення детальності знімків
		Функціональна інформація	Забезпечення фіксації функціонально важливих характеристик тканин
Трансляція	Оброблення зображення	Геометричне коригування	Оцінювання та компенсація спотворень, спричинених недосконалістю апаратури сканування
		Інтерполяція	Генерація проміжних даних
		Очищення шумів	Покращення якості зображення
		Виявлення об'єктів	Виявлення структур і створення цілісної тривимірної моделі тканин
	Інтерпретація	Ідентифікація типів клітин	Ідентифікація типів клітин
		Ідентифікація	Ідентифікація синапсів і їхніх зв'язків

		синапсів	
		Оцінка параметрів	Визначення функціональних параметрів клітин, зв'язків та інших утворень
		Каталогізація	Збереження результатів у зручному для наступного використання форматі
	Комп'ютерне моделювання нервової системи	Математичне моделювання	Моделювання елементарних сутностей і функцій
		Ефективна реалізація	Програмна реалізація моделі
Виконання	Сховище		Збереження коду моделі та її поточного стану
	Канали взаємодії		Ефективна міжпроцесорна взаємодія
	ЦП		Достатня швидкодія для роботи емуляції
	Емуляція кінцівок		Емуляція роботи периферичних систем тіла для можливості взаємодії з віртуалізованим або реальним середовищем через роботизовані інтерфейси
	Емуляція середовища		Віртуальне середовище або тіло

Є підстави сподіватися, що такі технології з'являться, але — не скоро. Функціональні комп'ютерні моделі деяких типів нейронів існують уже зараз. Програмне забезпечення, яке може розпізнати структури нейронних зв'язків на серії зображень — також (однак його надійність потребує вдосконалення). Існують засоби отримання зображень із потрібною деталізацією — тунельний мікроскоп дає змогу «побачити» навіть атоми, хоча тут така деталізація зайва. Може видатися, що ніщо не заважає створити необхідні технології, але для емуляції цілого мозку потрібно ще дуже багато досліджень і технологічних розробок[110]. Наприклад, навряд чи було б виправдано сканування тканини мозку тунельним мікроскопом, оскільки продуктивність цього процесу є низька. А для сканування менш точним електронним мікроскопом потрібні нові методи підготовки зразків і підфарбовування тканин кори, щоб деталізувати синаптичні структури. Також варто буде значно розширити нейрокомп'ютерні

бібліотеки, покращити засоби автоматизованого оброблення й розпізнавання зображень.

Технологія емуляції цілого мозку загалом менше залежить від теоретичних розробок у галузі штучного інтелекту і більше спирається на технологічні можливості. Вимоги, які вона висуває, залежать від деталізації та абстракції емуляції мозкової діяльності. Рівень вимог обернено пропорційний потребам у розумінні суті процесів. Через недосконалі сканери та недостатньо потужні комп'ютери ми не можемо відтворити всі деталі низькорівневих електрохімічних процесів мозку, тому для абстрактнішого моделювання маємо краще розуміти теоретичні принципи роботи мозку[111]. І навпаки, за допомогою достатньо потужних сканерів та комп'ютерів вдасться автоматично скопіювати людський мозок у найменших деталях, не заглиблюючись у принципи його роботи. За найфантастичнішим сценарієм ми могли б відтворити функціонування людського мозку на рівні елементарних часток, використавши квантове рівняння Шредінгера. У такому разі достатньо було б лише знань із фізики — а нейробіологія взагалі не потрібна. Щоправда, такий спосіб моделювання потребував би просто неймовірної обчислювальної потужності та ширини каналів передавання даних. Практичніше буде моделювати окремі нейрони, матриці їхніх зв'язків і частково відтворювати дендритні розгалуження й параметри деяких синапсів. Немає потреби емулювати окремі молекули нейротрансмітерів, хоч доведеться приблизно відтворювати рух їхніх концентрацій.

Для оцінки можливості емуляції цілого мозку треба визначити критерій успіху. Мета тут не відтворити оригінальний мозок до найдрібніших деталей, щоб передбачати реакції на певні подразники. Треба лише змоделювати ті обчислювальні потужності мозку, які потрібні для виконання інтелектуальної роботи. Біологічні подробиці роботи справжнього мозку в цьому контексті не потрібні.

Уважніший аналіз дає змогу виділити три рівні успішності емуляції, залежно від повноти реалізації функціоналу оригінального мозку. Наприклад, можна розрізняти: (1) *високоточну емуляцію оригінального мозку* з повним набором знань, умінь, можливостей і характеристик; (2) *приблизну емуляцію мозку* — значно штучне утворення, утім, здатне виконувати ту саму інтелектуальну роботу, що й оригінал; (3) *базову*

емуляцію мозку (може бути також наближеною) — за можливостями схожу на дитячий розум, без умінь і пам'яті оригіналу, проте з повною можливістю навчатися усього, що може вивчити звичайна людина[112].

Хоч створення високоточної емуляції мозку здається цілком реальним, спершу, напевно, зроблять значно простішу версію. Адже шлях до досконалості починається від недосконалості. Тому, перш ніж вдасться створити емуляцію цілого мозку, завдяки накопиченим знанням про принципи роботи мозку та винайденим технологіям, будуть створені деякі початкові версії нейроморфного штучного інтелекту. У наступних розділах ми побачимо, що можливість такого перетікання технологій заважає стратегічно оцінити динаміку створення емуляції мозку.

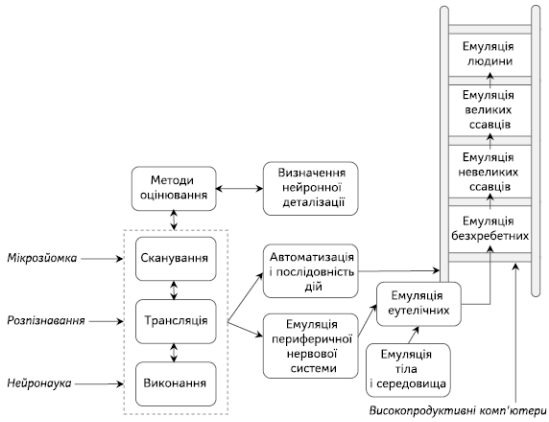


Рисунок 5. **Перспективний план емуляції цілого мозку. Схема передумов, дій та етапів**[113]

Але як близько ми зараз до емуляції мозку людини? В одній з недавніх оцінок наведено технічний перспективний план такого створення із висновком, що потрібні технології можуть з'явитися вже до середини століття, утім зі значною похибкою[114]. На рисунку 5 показано основні етапи цього плану. На перший погляд це може здатися оманливо просто, але не варто недооцінювати обсяг роботи, яку ще треба виконати. Досі не створено жодної успішної емуляції мозку. Візьмемо, як приклад, типовий дослідний організм — нематоду *Caenorhabditis elegans*, прозорий круглий черв 1 мм завдовжки, що має 302 нейрони. Повна карта нейронних зв'язків цього організму складена в середині-1980-х років за допомогою мікротомування, послідовного дослідження електронним мікроскопом і підписування вручну[115]. Але недостатньо знати, як між собою з'єднані нейрони. Потрібно розуміти, які із

синапсів збуджувальні, а які — гальмівні; яка сила зв'язків між ними. А також знати динамічні характеристики аксонів, синапсів, дендритних розгалужень. Навіть для такої маленької нервової системи, як у *C. elegans,* — цієї інформації поки немає (хоч зараз її нескладно зібрати під час невеликого спеціалізованого дослідного проекту)[116]. Успішна емуляція навіть такого маленького мозку дасть нам краще зрозуміти, що може знадобитися для емуляції більшого.

Із розвитком технологій, коли з'явиться можливість успішно моделювати роботу окремих ділянок мозкових тканин, завдання емуляції мозку зведеться до масштабування. Зверніть увагу на «драбину» в правій частині рисунка 5. Ця вертикальна послідовність клітинок зображує «фінішну пряму», якою рухатиметься процес створення емуляції мозку, здолавши лінію перешкод початкових труднощів. Ці етапи відповідають послідовності вдалих емуляцій щораз складніших нервових систем, наприклад, *C. elegans* → *бджола* → *макака резус* → *людина*. Відмінності між цими щаблями — принаймні після першого кроку — переважно кількісні і полягають (хоч і не виключно) у розмірі мозку. Тож емуляції потребуватимуть виконання майже тих самих операцій, тільки в більших масштабах[117].

Щойно ми опинимося на цій фінішній драбині, нам буде до снаги точніше передбачити *наближення* успішної емуляції цілого мозку людини[118]. Так ми зможемо дістати попередження про створення штучного інтелекту рівня людини, принаймні якщо останньою потрібною для цього технологією буде швидкість сканування або потужність комп'ютерів. Якщо ж, натомість, проблема виявиться в якості комп'ютерних моделей нейронів і успіхи у нейромоделюванні змусять себе чекати, перехід від невдалих прототипів до робочої моделі може відбутися значно швидше. Нескладно уявити сценарій, коли, незважаючи на високу детальність сканування й надлишок доступного процесорного часу, модель не працює. Але щойно останню помилку виправлено, повністю недієздатна система — можливо, подібна до стану непритомності мозку під час епілептичного нападу — раптово стабілізується і пробуджується. У такому разі спосіб досягнення успішного відтворення людського мозку відрізнятиметься від варіанта послідовної емуляцій нервових систем дедалі складніших

істот (із щоразу більшим розміром літер у газетних заголовках). Оцінити кількість помилок у комп'ютерних нейронних моделях може бути дуже важко навіть тим, хто з ними працює постійно. Ніхто до останнього не знатиме, коли все буде виправлено. (Після створення емуляції мозку події (потенційно вибухові) почнуть розгортатися значно швидше. Але ми поговоримо про це в розділі 4).

Тому, навіть якщо дослідження будуть максимально публічними, можемо очікувати несподіванок. Однак загалом шлях до штучного інтелекту через емуляцію цілого мозку є прогнозованіший, адже більше залежить від розвитку технологій, а не від теоретичного осягнення проблеми інтелекту. Також можемо упевненіше (ніж у випадку з ШІ) стверджувати, що найближчі п'ятнадцять років через відсутність низки ключових технологій годі чекати успіхів в емуляції мозку. Натомість написати зерно ШІ на сучасному комп'ютері фактично реально; і, можливо — хоч і малоймовірно, — хтось зробить це вже в найближчому майбутньому.

Біологічний мозок

Третій спосіб отримати кращий інтелект, ніж той, що доступний людям зараз, — це покращити роботу біологічного мозку. Загалом цього можна досягти і без будь-яких технологій — за допомогою селекції та схрещування. Але застосування класичної євгеніки щодо людини, та ще й — у таких масштабах, безсумнівно зіткнеться з політичними та моральними запереченнями. Ба більше, якщо відбір не буде дуже суворим, знадобиться багато поколінь, щоб результати стали відчутними. Перш ніж така ініціатива принесе плоди, вона стане застарілою — біотехнології дадуть змогу безпосередньо контролювати людські гени і нейробіологічні процеси. Тому зосередимося на методах, потенційно здатних принести результати швидше: у масштабах кількох поколінь.

Існує кілька способів покращити індивідуальну здатність до пізнання, серед них традиційні шляхи, як-от освіта і тренування. Ще сприяють розвитку мозку такі прості чинники, як здорове харчування матері й дитини, відсутність свинцю й інших нейротоксинів у середовищі зростання, відсутність паразитів, удосталь сну та вправ,

запобігання небезпечних для мозку захворювань[119]. Кожен із них може покращити здатність пізнання, але дуже незначно — особливо в спільнотах, добре забезпечених їжею і освітою. Авжеж завдяки цьому суперінтелекту не досягти, але підняти середній рівень інтелекту, загальну охопленість знаннями — можна. (У деяких збіднілих місцевостях мешканці мають знижений рівень інтелекту внаслідок йододефіциту — ганебне явище, враховуючи, що забезпечити цих людей йодованою сіллю коштує всього кілька центів на рік за особу[120]).

Відчутніше покращити інтелектуальні здібності можуть біомедичні засоби. Уже існують лікарські препарати, що здатні покращити пам'ять, зосереджуваність, інтелектуальну працездатність[121]. (Наприклад, робота над цією книжкою підживлювалася кавою та нікотиновою жувальною гумкою). І нехай ефективність сучасних препаратів почасти викликає сумнів та іноді є надто незначною та невиразною, ноотропні препарати майбутнього можуть бути ефективнішими й безпечнішими[122]. А втім, з погляду неврології та еволюції, малоймовірно, що введення певної хімічної речовини спровокувало у здоровому мозку раптовий спалах надзвичайного інтелекту[123]. Робота пізнавального механізму мозку залежить від сукупності багатьох чинників і покращити її можна радше сприянням та обережним налаштуванням, аніж вживанням чужорідного зілля.

Генна інженерія дає потужніші інструменти впливу на розвиток інтелекту, ніж психофармакологія. Наприклад, знову розглянемо селекцію, тільки замість класичної євгеніки з контролем парування, відбір може проводитися на стадії гамет або після запліднення[124]. Зараз, наприклад, перед імплантацією під час екстракорпорального запліднення (ЕКЗ) проводять генетичну діагностику ембріона на наявність моногенетичних захворювань, як-от хвороба Гантінгтона, та на схильність до деяких хвороб, наприклад, раку молочної залози. Таке дослідження використовують також при народженні дитини для визначення статі або лейкоцитарної сумісності із хворим родичем — під час відбору клітин пуповинної крові для донорства[125]. Перелік ознак, які можна використовувати в селекції, значно збільшиться протягом найближчих десятиліть. Здешевлення генотипування та секвенування ДНК надає потужний імпульс розвитку психогенетики. Тільки зараз ми нарешті можемо провести складний аналіз генетичних

ознак у великої кількості піддослідних з охопленням всього генома, який може значно розширити наші уявлення про генетичну архітектуру нашого пізнання та поведінки[126]. Будь-яку ознаку з ненульовою успадковуваністю — до яких належить і здатність навчатися — можна використати під час селекції[127]. На стадії ембріонів селекція не вимагає розуміння плетива причинно-наслідкових зв'язків між генотипом, середовищем і фенотипом: важлива лише величезна кількість даних генетичних корелятів потрібної ознаки.

Ми можемо приблизно порахувати розмір приросту ознаки від різних селекційних сценаріїв[128]. У таблиці 5 наведено очікуване зростання інтелектуальних здібностей за умови наявності повної інформації про спільну адитивну варіативність в успадковуваності (у вузькому розумінні) інтелектуальних здібностей. (За неповної інформації ефективність селекції знизиться, але незначно[129]). Не дивно, що селекція в більшій популяції ембріонів дає більший приріст ознаки, за якою відбувається селекція, але починає швидко спадати. Приріст для ста ембріонів далеко не в п'ятдесят разів більший, ніж для двох[130].

Таблиця 5. **Максимальне зростання показника IQ після селективного відбору ембріонів[131]**

Селекція	**Приріст IQ**
1 з 2	4,2
1 з 10	11,5
1 зі 100	18,8
1 із 1000	24,3
5 поколінь 1 із 10	< 65 (через спадання віддачі)
10 поколінь 1 із 10	< 130 (через спадання віддачі)
Кумулятивні межі (при оптимізації адитивних варіацій для пізнання)	100 + (< 300 (через спадання віддачі))

Цікаво, що вплив закону спадної віддачі значно знижується за умови проведення селекції в кількох поколіннях. Отже, приріст ознаки при селекції 1 з 10 особин протягом десяти поколінь (де кожне наступне покоління складається з нащадків відібраних особин попереднього

покоління) буде значно більшим, ніж приріст від одноразової селекції 1 зі 100. Недолік послідовної селекції в кількох поколіннях — її тривалість. Якщо між відборами в поколіннях проходитиме двадцять або тридцять років, то навіть п'ять послідовних поколінь селекції завершаться вже у XXII столітті. Дуже ймовірно, що задовго до того з'являться значно потужніші та безпосередніші методи генної інженерії (не говорячи вже про штучний інтелект).

Утім існує технологія, яка, у разі можливості її застосування для людей, могла б доповнити передімплантаційний генетичний скринінг: отримання статевих клітин (гамет) із ембріональних стовбурових клітин[132]. Така технологія вже дала змогу отримати плідне покоління мишей і клітини людини з деякими ознаками гамет. Тепер науковці постали перед нелегким завданням — відтворити для людей ті результати, які вдалося отримати для лабораторних тварин, уникнувши епігенетичних аномалій у клітинних лініях. Цитуючи одного з експертів, успіх у виконанні цього завдання дасть науці здійснити стрибок на «10 або й 50 років у майбутнє»[133].

З можливістю отримувати статеві клітини зі стовбурових, будь-яка пара матиме широкий вибір. Зараз, під час екстракорпорального запліднення, утворюють менше ніж десять ембріонів. Маючи можливість генерувати гамети, з кількох донорських стовбурових клітин можна отримати майже необмежену кількість гамет для наступного запліднення і генотипування або секвенування ембріонів. Для того щоб урешті вибрати для імплантації ембріон із найбільш перспективним набором ознак. Така технологія, залежно від вартості підготовки й тестування ембріонів, може в кілька разів підвищити можливості селекції під час екстракорпорального запліднення.

Але ще цікавіше те, що можливість вирощувати гамети зі стовбурових клітин дасть змогу проводити *ітераційну ембріональну селекцію* і таким способом стиснути кілька поколінь селекції в період, менший за середній період досягнення зрілості людини. Теоретично процес складається з таких етапів[134]:

1. Генотипування та відбір ембріонів за низкою генетичних ознак.
2. Вилучення стовбурових клітин з відібраних ембріонів і дальше отримання з них гамет, тривалість дозрівання — до шести місяців[135].
3. Схрещення цих гамет для отримання нового покоління ембріонів.

4. Повторювати попередні кроки до накопичення необхідного приросту генетичних ознак.

Так можна провести селекцію для десяти або більше поколінь лише за декілька років. (Звісно, процес потребуватиме багато часу та витрат, проте — теоретично — має бути проведений лише раз. Клітинні лінії, отримані в такий спосіб, можуть бути використані для вирощування нових ембріонів).

Як видно з таблиці 5, середній рівень інтелекту особин, зачатих у такий спосіб, може бути дуже високим — імовірно, дорівнювати або перевищувати історичний максимум рівня інтелекту людини. У світі, де такі люди становитимуть значну частину населення, напевно, можна буде твердити про появу колективного суперінтелекту (за умови наявності відповідної культури, освіти та інфраструктури).

Вплив такої технології проявиться не відразу та буде пом'якшений кількома чинниками. Ніяк не вдасться скоротити час дорослішання отриманих ембріонів: перш ніж така людина досягне найпродуктивнішого віку, мине не менше двадцяти років, а ще більше — поки такі люди становитимуть значну частину продуктивного прошарку населення. Ба більше, між розробленням технології та її широким впровадженням теж, імовірно, мине чимало часу. У низці країн її буде заборонено з моральних або релігійних міркувань[136]. Але і в державах, де селекцію буде дозволено, багато пар віддаватимуть перевагу традиційному зачаттю. А втім, за наявності чітких переваг, як-от гарантована обдарованість або відсутність схильності до хвороб, кількість охочих скористатися ЕКЗ зростатиме. На користь генетичної селекції також свідчитимуть менші медичні витрати та більші капітальні накопичення протягом життя. Генетичну селекцію застосовуватимуть все ширше, особливо в елітарних колах.

У культурних нормах уявлення, що генетична селекція є загальноприйнятою батьківською практикою відповідальних і освічених пар, поступово закріпиться. Несхильні пари спочатку долучатимуться, щоб їхні діти не опинилися в невигідному становищі, порівняно з однолітками та дітьми друзів і колег. Деякі країни можуть стимулювати громадян використовувати генетичну селекцію, щоб покращити якість людського капіталу. Або для підвищення

довготермінової соціальної стабільності завдяки селекції за ознаками покірності, слухняності, вміння пристосовуватися, неприйняття ризику чи боягузливості — серед найнижчих соціальних прошарків.

Вплив такої технології на інтелектуальну сферу залежатиме від того, скільки уваги приділено інтелектуальним ознакам під час селекції (таблиця 6). Проводячи селекцію, генний інженер буде змушений розподіляти наявні можливості між усіма бажаними ознаками, і розумові здібності конкуруватимуть із іншими важливими перевагами, як-от здоров'я, краса, харизма, фізична сила. Технологія ітераційної ембріональної селекції, звісно, зможе забезпечити більшість потреб і без необхідності компромісу. Однак така процедура може порушити звичайну генетичну відповідність між батьками й дітьми, що може завадити її поширенню в багатьох культурах[137].

Таблиця 6. **Наслідки різних сценаріїв застосування генетичної селекції[138]**

Технологія Поширеність	**«ЕКЗ+» Селекція 1 з 2 ембріонів (+4 од. IQ)**	**«Турбо ЕКЗ» Селекція 1 з 10 ембріонів (+12 од. IQ)**	**«Яйце in vitro» Селекція 1 зі 100 ембріонів (+19 од. IQ)**	**«Ітераційна ембріональна селекція» (+>100 од. IQ)**
«За крайньої потреби» ~0,25 % зачать	Незначний соціальний вплив протягом одного покоління. Суспільний резонанс перевищує реальний вплив	Незначний соціальний вплив упродовж одного покоління. Суспільний резонанс перевищує реальний вплив	Нечисленні відібрані виділяються серед кандидатів на інтелектуально вимогливі посади	Відібрані переважають серед титулованих учених, юристів, лікарів, інженерів. Інтелектуальне Відродження?
«Елітна перевага» 10 % зачать	Незначний вплив на розумові здібності протягом одного покоління. Проводиться також селекція за ознаками, не	Велика частина студентів Гарварду — відібрані. В інтелектуально інтенсивних сферах діяльності	Відібрані першого покоління переважають серед учених, юристів, лікарів, інженерів	Постгуманізм[139]

	пов’язаними з інтелектом, для помітнішого ефекту	зайняті переважно відібрані другого покоління		
«Новий звичай» >90 % зачать	Значне зниження випадків розладів навчання в дітей. Кількість людей із високим IQ у другому поколінні подвоюється	Відчутне зростання рівня освіти та достатку. Багаторазове зростання кількості людей із високим IQ у другому поколінні	У першому поколінні рівень IQ, властивий видатним ученим, трапляється в 10+ разів частіше. У другому поколінні — у тисячі разів	Постгуманізм

З подальшим розвитком генної інженерії, імовірно, з’явиться можливість безпосереднього синтезу генома з потрібними характеристиками без використання ембріональної селекції. Синтез ДНК уже значно автоматизований і є звичною справою. Хоча створити в такий спосіб повний геном людини, який можна використати в репродуктивних цілях нараз, неможливо через невивченість епігенетичних механізмів[140]. Але щойно синтез генома стане доступним, можна буде розробити ембріон з точною комбінацією бажаних генетичних ознак його батьків. Гени, яких не вистачає, можна буде додати — включно з алелями, рідкісними для популяції, але важливими для розумових здібностей[141].

Завдяки синтезу генома людини можна буде виконувати генетичну «перевірку орфографії» ембріона. (Схожого ефекту можна досягти також ітераційною ембріональною селекцією). Кожен із нас має в генах сотні мутацій, які зменшують ефективність клітинних процесів[142]. Вплив кожної окремої мутації незначний (і вона що покоління виводиться з генофонду), але разом вони чинять відчутний тиск на роботу організму[143]. Кількість і природа таких бракованих алелей у кожному з нас може грати суттєву роль у відмінностях наших інтелектуальних можливостей. Завдяки синтезу генома можна скопіювати геном новоствореного ембріона, відкинувши генетичний шум накопичених мутацій. Це може звучати провокативно, але

створені з таких відкоригованих геномів люди можуть виявитися «людськішими» — менш спотвореними втіленнями людської подоби, ніж будь-хто з нині живих. І вони не будуть копіями одне одного, тому що генетичні розбіжності людей не вичерпуються одними шкідливими мутаціями. Носій такого виправленого генома матиме видатну фізичну та розумову будову, фенотипічними проявами якої будуть високі показники полігенних ознак, як-от інтелекту, здоров'я, витривалості та привабливості[144]. (У схожий спосіб під час синтезу складеного обличчя з багатьох зображень реальних облич, окремі відхилення зовнішності втрачаються внаслідок усереднення: див. рисунок 6).

Інші біотехнології теж можуть бути потенційно цікавими для поліпшення інтелекту. У майбутньому завдяки репродуктивному клонуванню людини можна копіювати геном найталановитіших особистостей. Хоч більшість батьків буде схильна віддавати перевагу генетичній спорідненості зі своїми дітьми, проте навіть обмежене застосування такого клонування може дати помітний ефект. Цьому є кілька ймовірних причин: 1) навіть невелике зростання кількості виняткова талановитих людей може мати значний вплив; 2) деякі країни можуть заохочувати євгенічні програми завдяки, скажімо, доплатам сурогатним матерям. Інші технології генної інженерії, як-от розроблення і синтез нових генів або вставлення у ДНК промоторів чи інших елементів для контролю експресії певних генів, також можуть згодом відіграти важливу роль у покращенні біологічного інтелекту. Можна уявити собі екзотичніші варіанти, наприклад, величезні ємності зі штучно вирощеними мозковими тканинами, організованими у складні структури або інтелектуально «прокачані» трансгенні тварини (імовірно, ссавці з великим мозком, слони чи кити, збагачені людськими генами). Ці варіанти повністю теоретичні, проте у довгостроковій перспективі відкидати їх не варто.

Досі ми розглянули лише генетичні втручання в гамети й ембріони. Швидший результат можна отримати без очікування дозрівання ембріонів — виправляючи гени напряму в соматичних клітинах. Однак це складніше технологічно. Адже необхідно внести зміни в гени величезної кількості клітин живих органів, — у випадку покращення інтелекту — клітин мозку. Натомість селекція ембріонів, наприклад,

узагалі не потребує генетичних втручань. А втручання, які потребують редагування генома (коригування або вставлення рідкісних алелей), легше проводити для гамет та ембріонів через невелику кількість клітин. Ба більше, редагувати геном ембріона буде, напевно, значно ефективніше, адже така процедура впливатиме на весь процес формування мозку. Натомість генетичне редагування зрілого мозку обмежуватиметься вже наявними сформованими структурами. (Деяких результатів соматичної генної терапії, імовірно, можна буде досягти фармакологічними засобами).

Рисунок 6. **Складені обличчя як метафора виправленого генома**
Кожне з облич всередині зображення створене попарним накладанням фотознімків шістнадцяти осіб (жителів Тель-Авіва). Такі складені обличчя здаються красивішими, ніж обличчя тих, з кого їх створено, через те, що недоліки окремих облич втрачаються під час усереднення. Так само люди з відкоригованим геномом через видалення мутацій можуть виявитися ближчими до «платонічних ідеалів». Такі люди не будуть ідентичними, адже алелі багатьох генів однаково функціональні. А коригуванням можна буде позбавлятися лише тієї варіантності, яка спричинена шкідливими мутаціями[145].

Говорячи про генетичні втручання, перш ніж оцінювати вплив результатів на світ, маємо пам'ятати про час очікування, потрібний для зростання покоління модифікованих організмів[146]. Навіть якби технологія вже існувала і була готова до негайного застосування, минуло б щонайменше двадцять років, перш ніж перші генетично вдосконалені нащадки досягли зрілості. Окрім того, між першими лабораторними успіхами і клінічним застосуванням нової технології зазвичай проходить близько десятка років, протягом яких вивчають її безпечність. Проте наукові та клінічні основи звичайної селекції давно сформовані і застосовуються в методиках лікування безпліддя та генній інженерії. Тому залишається лише використати їх для усвідомленого відбору там, де інакше обирав би випадок.

Крім того, стримувати впровадження нової технології може не лише страх поразки (необхідність переконатися в безпечності), а також страх перед успіхом — потрібно впевнитися щодо відсутності моральних перешкод у застосуванні селекції до людини й оцінити соціальні наслідки всіх можливих аспектів процедури. Значущість цих

пересторог може відрізнятися між країнами залежно від культурного, історичного та релігійного контексту. У післявоєнній Німеччині, наприклад, тривалий час було заборонено будь-які репродуктивні практики, бодай якось пов'язані з покращенням нащадків. І це не дивно, адже пам'ять про звірства нещодавніх євгенічних експериментів у країні була ще свіжою. Іншим західним державам властиві ліберальніші погляди. Деякі ж країни — наприклад, Китай чи Сингапур, де народжуваність уже тривалий час контролює уряд, — можуть не тільки дозволити селекцію та генну інженерію, а й активно заохочувати їх застосування для покращення інтелектуальних здібностей населення.

Щойно з'являться перші приклади і результати застосування цих технологій, застереження почнуть потрохи танути. Країни постануть перед вибором — пасти задніх в інтелектуальних перегонах в економіці, науці, військовому та політичному впливові чи освоювати нові технології вдосконалення людини. Спостерігаючи, як до елітних шкіл масово зараховують дітей, що пройшли генетичну селекцію (та які назагал вродливіші, здоровіші та здібніші за однолітків), усі бажатимуть і своїм нащадкам такої долі. Щойно стане зрозуміло, що технологія працює і дає відчутні переваги, ставлення до неї змінюватиметься дуже швидко, імовірно, упродовж десятиліття. Вивчення суспільної думки щодо репродуктивних технологій у Сполучених Штатах у 1978 році, — після народження Луїзи Браун, першої «дитини з пробірки», — показало швидку зміну думки американців про екстракорпоральне запліднення. За кілька років до того лише 18 відсотків опитаних були готові застосувати ЕКЗ, а після народження Луїзи кількість прихильників зросла до 53 відсотків і продовжила зростати[147]. (Для порівняння: в опитуванні, проведеному у 2004 році, 28 відсотків респондентів схвально сприйняли ідею ембріональної селекції з метою покращення «сили або інтелекту», 58 відсотків схвалювали її застосування для зменшення схильності до раку в дорослому віці, 68 відсотків — для боротьби з летальними дитячими захворюваннями[148]).

Років п'ять-десять потрібно для накопичення знань і підвищення ефективності селекції ембріонів (або і довше, поки вдасться отримати продуктивні гамети зі стовбурових клітин), років десять для

проведення селекцій і двадцять-двадцять п'ять років, поки поліпшене покоління досягне продуктивного віку. Якщо врахувати всі ці затримки, то можемо дійти висновку, що в найближчі п'ятдесят років генетичні покращення не матимуть суттєвого впливу на розвиток суспільства. Але від появи перших таких прикладів поширення генетичних методів покращення розумових здібностей серед значної частини дорослого населення значно збільшиться. І збільшуватиметься далі, коли дедалі більше генетично покращених людей, озброєних новими генетичними технологіями (зокрема можливістю створення гамет зі стовбурових клітин та ітераційною ембріональною селекцією), візьмуться до роботи.

Зі вдосконаленням цих технологій (не говорячи про екзотичніші опції, наприклад, інтелект у штучно вирощених нервових тканинах) можна буде створити суспільство, у якому кожен новий індивід буде в середньому розумніший, ніж будь-хто з народжених до нього. Отже, потенціал біологічного вдосконалення дуже високий, щонайменше достатній для створення слабкої форми суперінтелекту. Це не дивно, адже, зрештою, немає підстав вважати, що еволюція генетичної лінії людини від мавп, наших людиноподібних предків, закінчилася на Homo sapiens і що наш вид досяг абсолютного піку своїх когнітивних можливостей. Оскільки нас складно назвати найрозумнішим біологічним видом з усіх можливих, тож доведеться назвати найдурнішим біологічним видом, здатним створити технологічну цивілізацію. У цій ніші в нас поки що немає конкурентів, а тому справжній рівень нашої пристосованості до цього невідомий.

Цей сценарій покращення біологічного розуму досить легко уявити. Він не може бути таким швидким і неочікуваним, як створення штучного інтелекту, адже на зміну поколінь потрібен час. (Генна інженерія соматичних клітин і фармакологічні методи впливу працюють швидше, але значно менш досконалі й ефективні). Тому, звісно, перспективні можливості штучного інтелекту здаються привабливішими. (Узяти хоча б швидкість роботи електронних компонентів: наприклад, сучасний транзистор працює в десять мільйонів разів швидше від біологічного нейрона). Але навіть незначне покращення біологічного розуму матиме важливі наслідки. Зокрема, пришвидшить розвиток науки й технологій, а отже, і появу

ефективніших засобів покращення біологічного розуму та створення штучного інтелекту. Лише уявіть швидкість прогресу в суспільстві, у якому будь-яка пересічна людина не поступається розумом Алану Тюрінгу чи Джону фон Нейману. Суспільство, у якому існують мільйони людей, здібніших від будь-якого генія минулих років[149].

У наступних розділах ми детальніше поговоримо про стратегічний вплив покращення інтелекту. Поки що ж можемо зробити такі висновки: 1) за допомогою біотехнологій можна створити суперінтелект, принаймні його слабку форму; 2) якщо *ми* принципово неспроможні створити штучний інтелект (хоч раціональних підстав для такого переконання немає), його можуть винайти інтелектуально покращенні люди майбутнього, тим більше, що реалістичні способи поліпшення людського інтелекту існують; 3) у довгостроковій перспективі, до другої половини сторіччя, ми маємо враховувати можливість появи цілих поколінь генетично поліпшених людей — виборців, винахідників, науковців — і ступінь цих поліпшень також зростатиме з кожним десятиліттям.

Нейроінтерфейси — взаємодія людини й машини

Ще один варіант пропонує людям створити нейроінтерфейси — спеціальні мозкові імпланти. Вони дадуть змогу отримати безпосередній доступ до можливостей цифрової обчислювальної техніки — технологій зберігання даних, швидких і високоточних обчислень, високошвидкісного передавання даних. А отже, можуть значно підвищити продуктивність розумової діяльності[150]. Але, незважаючи на практичну можливість прямого під'єднання людського мозку до комп'ютера, не схоже, що невдовзі такі пристрої набудуть широкого вжитку[151].

При імплантації електродів у мозок насамперед виникає ризик медичних ускладнень — інфекції, зміщення електродів, кровотечі, порушення розумової діяльності. Найяскравішою ілюстрацією переваг стимулювання мозку за допомогою імпланта є спосіб лікування хвороби Паркінсона. Імплант, який використовують при цьому, насправді не здійснює обміну інформацією, а просто стимулює субталамічне ядро розрядами струму. На демонстраційному відео

видно, як хворий, що безсило відкинувся на стільці, знерухомлений хворобою, раптово випростується пробуджений, щойно вмикають імплант. Тепер він рухає руками, стоїть, ходить по кімнаті, повертається й робить пірует. Ця картина приголомшує своєю простотою і магічним успіхом, але й тут приховано небезпеки. Одне дослідження стану хворих на хворобу Паркінсона, які отримали глибинні імпланти, показало порушення вербальної плинності, вибіркової уваги, розрізнення кольорів, вербальної пам'яті порівняно з контрольними значеннями. Також хворі більше скаржилися на когнітивні проблеми[152]. Якщо, у разі лікування важкої хвороби, такі побічні дії та ризики можуть бути виправданими, то для здорової людини потрібні дуже вагомі причини, щоб погодитися на нейрохірургічне втручання.

Отже, удосконалити зазвичай складніше, ніж вилікувати, і це друга причина сумніватися, що кіборгізація допоможе нам досягти суперінтелектуальності. Хворим на параліч може допомогти імплантація штучних нервів або генераторів упорядкованого руху[153]. При захворюваннях слуху та зору можна спробувати імплантацію штучного завитка чи сітківки[154]. Для хворих на хворобу Паркінсона чи хронічні болі є імплант, що стимулює чи гальмує певні ділянки мозку[155]. Але створення прямого каналу взаємодії мозку з комп'ютером здається занадто складним і ризикованим способом підвищення ефективності розумової діяльності. Більшість переваг від використання комп'ютера можна отримати за допомогою наших органів чуттів та звичних засобів взаємодії, без витрат і ризиків імплантації. Нам не потрібно під'єднувати до мозку оптоволокно, щоб виходити в інтернет. Людська сітківка не лише може передавати візуальні дані з дивовижною швидкістю в десять мільйонів бітів у секунду, але також тісно інтегрована у відповідні підсистеми зорового нерва й мозку, які призначені для високоефективного аналізу та інтерпретації цього потужного потоку і передавання отриманої інформації до різних ділянок мозку для подальшого оброблення[156]. Навіть маючи простий спосіб введення додаткових даних до мозку, ми не змогли б думати та вчитися швидше без додаткового «апгрейду» можливостей оброблення даних. Оскільки такі можливості — це майже весь мозок, то потрібен буде «протез мозку», а це не що інше, як

штучний інтелект загального призначення. Але маючи такий ШІ, для чого поміщати його в кістяну коробку, якщо металева нічим не гірша? У такий спосіб ми повертаємося до вже розглянутого шляху штучного інтелекту.

Нейроінтерфейс також можна використовувати для виведення інформації з мозку, наприклад, для спілкування з іншим мозком або комп'ютером[157]. Такі пристрої допомагають хворим із синдромом ізоляції (деаферентації) керувати курсором на екрані за допомогою думки і в такий спосіб спілкуватися з іншими[158]. Пропускна здатність такого інтерфейсу дуже низька: вибираючи літеру за літерою, пацієнт здатен надрукувати лише кілька слів упродовж хвилини. Можна уявити собі покращену версію такої технології — наступне покоління імплантів, які вживлюються в центр Брока (ділянка мозку в лобних частках, що відповідає за мовлення) і перехоплюють внутрішнє мовлення[159]. Така технологія може бути корисною для людей з інвалідністю, спричиненою інсультом або м'язовою дистрофією, але навряд чи зацікавить здорових людей. Ідентичний функціонал можна замінити мікрофоном з програмою розпізнавання мовлення — і без болю, незручностей, витрат і післяопераційних ризиків (не враховуючи відсутності в мозку такого собі гіпер-Орвелівського пристрою для підслуховування думок). Крім того, пристрої, які перебувають зовні людини, простіше лагодити і вдосконалювати.

Але як же бути з мрією позбутися необхідності використовувати слова і за допомогою прямого з'єднання «завантажувати» з одного мозку до іншого поняття, думки, цілі царини знань? Таке завдання здається нескладним лише на перший погляд через хибне уявлення про способи зберігання й передавання інформації всередині мозку. Як ми вже згадували, можливість обміну даними з мозком обмежена не швидкістю передавання даних у мозок, а швидкістю його інтерпретації даних. Тоді, можливо, спробуймо передавати поняття безпосередньо, не через рецепторні дані, які реципієнт має інтерпретувати. Проте тут є дві проблеми. Перша полягає в тому, що наш мозок, на відміну від комп'ютерної програми, не має певних стандартних форматів зберігання і представлення даних. Навпаки, кожен мозок виробляє власне, унікальне внутрішнє представлення понять вищого рівня. Відповідність між групами нейронів і певними поняттями

встановлюється стихійно на основі унікального сенсорного досвіду кожного конкретного мозку (із врахуванням генетичних факторів та стохастичних психологічних процесів). У біологічній нейронній мережі, як і в штучній, сутності представлені холістично, структурами та схемами активностей великих масивів елементів, а не окремих клітин ієрархічних утворень[160]. Тому неможливо встановити однозначну відповідність нейронів і автоматично переносити думки з одного мозку в інший. Щоб думки однієї людини були зрозумілі іншій, їх треба розібрати на складові і зобразити в символічному вигляді за допомогою певної, зрозумілої обом, системи символів. Функцію такої системи виконує мова.

Уявити інтерфейс, який має інтерпретувати й артикулювати якимось способом зчитані з мозку відправника стани нейронів і перескеровувати їх до мозку отримувача, загалом можливо. Але це приводить нас до другої проблеми. Окрім одночасного зчитування й записування станів мільярдів окремих нейронів, що саме собою є (дуже складним) технічним завданням, такий інтерфейс має перетворювати активаційні схеми нейронів одного мозку на семантично еквівалентні активаційні схеми іншого. Для цього потрібен такий рівень розуміння роботи багатошарових нейронних утворень, якого цілком достатньо, щоб створити нейроморфний ШІ.

Незважаючи на описані проблеми, спосіб покращення інтелекту за допомогою кіборгізації не зовсім безнадійний. Дивовижні результати досліджень роботи гіпокампа в щурів свідчать про можливість створення протезу, що може покращити роботу тимчасової пам'яті[161]. В останній версії експерименту сигнали з десятка електродів, імплантованих в одній ділянці гіпокампа («CA3»), порівнюють із сигналами такої самої кількості нейронів у другій ділянці («CA1»). Мікропроцесор навчають розрізняти дві схеми активації у першій ділянці (що відповідають двом різним «згадкам» — «правий важіль» або «лівий важіль») і пов'язувати зі схемами активації другої ділянки, щоб потім відтворювати їх. Такий протез може не тільки відновити зв'язок між двома ділянками у випадку його втрати, а й пришвидшити згадування в щура порівняно з нормальним станом. Незважаючи на безсумнівний успіх дослідження, багато питань залишаються відкритими: чи підходить такий спосіб для більшої кількості «згадок»?

Чи вдасться контролювати комбінаторний вибух можливих співвіднесень вхідних і вихідних сигналів за умови збільшення кількості вхідних і вихідних електродів? Чи усунулись небажані побічні ефекти роботи пам'яті, які спостерігалися під час експерименту,, як-от неможливість узагальнювати подразники або неможливість перевчитися при зміні середовища? І наскільки корисні будуть такі пристрої істотам, які мають доступ до допоміжних засобів запам'ятовування, як-от блокнот і олівець? Як складно буде застосовувати такий принцип для інших ділянок мозку? Цей протез має просту односпрямовану архітектуру завдяки гіпокампу щура (який, по суті, є мостом між ділянками «CA3» і «CA1»). Натомість інші структури мозку влаштовані значно складніше і з'єднані плетивом двобічних зв'язків, що значно ускладнює завдання з відтворення схемотехніки і функціоналу кожної окремої групи нейронів.

Залишається сподіватися, що мозок, з'єднаний із деяким зовнішнім інформаційним ресурсом нейроінтерфейсом, згодом сам навчиться порівнювати власний стан із сигналом, отриманим по інтерфейсу, або сигналом, що інтерфейс надсилає. Тоді імплант не повинен бути надто «інтелектуальним», якщо мозок виявиться достатньо розумним, щоб адаптуватися до нового джерела даних: як мозок дитини, який учиться розрізняти сигнали з рецепторів органів зору і слуху[162]. Але мусимо запитати: чи варте воно того? Припустимо, мозок може навчитися розрізняти певні образи у вхідному сигналі, підведеному до певної ділянки мозку через нейроінтерфейс: чому не проектувати ту саму інформацію у візуальному вигляді на сітківку або як звук — на слухові рецептори? Такі низькотехнологічні способи також дають використати пластичність навчання живого мозку для отримання інформації і водночас позбавляють небезпеки ускладнень від нейрохірургічного втручання.

Мережі та організації

Ще один гіпотетичний спосіб досягнення суперінтелектуальності — поступове вдосконалення зв'язків (комунікаційних та організаційних) між окремими людськими інтелектами, іншими механізмами й ботами. Основна ідея полягає не в суперінтелектуальності одного

індивіда, а в тому, що, за умови правильної організації і комунікації, система окремих інтелектів може досягти деякої форми суперінтелектуальності. Цю форму надалі ми називатимемо «колективним суперінтелектом»[163].

Завдяки своєму колективному інтелекту упродовж усієї історії свого існування, від доісторичних часів і до сьогодні, люди багато чого досягли. Зокрема, винайшли технології передавання знань, як-от письмо і друк, але, насамперед, власне мову. За цей час зросла кількість населення й густота поселень, покращилися методи організації та епістемічні норми, поступово накопичувався інституційний капітал. Загалом колективний інтелектуальний потенціал системи обмежений рівнем інтелектуальності окремих її елементів. А також страждає від неефективного передавання корисної інформації між ними та незбалансованості, нелогічності, які властиві людським організаціям. Завдяки оптимізації зв'язків між елементами (не лише за допомогою використання дешевшого і швидшого обладнання, а й збільшення концентрованості, доступності інформації, що передається) можна збільшити розмір і цілісність організаційних утворень. Те саме можна сказати і про позбування від деяких бюрократичних традицій, які відволікають та виснажують ресурси організацій: марнославні змагання у важливості, підміна пріоритетів, приховування фальсифікації даних та інші. Навіть часткове розв'язання цих проблем може суттєво вплинути на зростання колективного інтелекту.

Щоб пришвидшити зростання нашого колективного інтелекту, потрібно запровадити багато різноманітних технологічних та інституційних нововведень. Зокрема, використання субсидованих ринків прогнозування може сприяти встановленню загальноприйнятих правил визначення істинності і покращенню прогнозування щодо спірних наукових та соціальних питань[164]. Детектор брехні (якщо вдасться створити достатньо надійний і простий у використанні пристрій) може зменшити поле непевності в людських стосунках[165]. Але ще кориснішим був би пристрій виявлення нещирості перед самим собою[166]. А втім, навіть без новомодних нейротехнологій, публічні дані, зокрема про доброчесність і репутацію, поширення знань епістемічних норм та культури раціонального мислення можуть допомогти викорінити багато форм

шахрайства. Добровільне та примусове відеоспостереження надасть величезну кількість інформації про людську поведінку. Більше ніж мільярд людей ділиться персональними даними в соціальних мережах: можливо, скоро з'являться цілісні записи щоденного життя з камер і мікрофонів смартфонів та окулярів. Автоматичний аналіз таких потоків даних відкриє нові способи їхнього застосування (звісно, як корисні, так і шкідливі)[167].

Зростанню колективного інтелекту можуть сприяти загальніші організаційні покращення й економічне зростання, збільшення кількості освіченого населення, яке залучене до культурних процесів і має доступ до світових інформаційних ресурсів[168].

Особливо динамічним середовищем для інновацій та експериментів є інтернет. Значна частина його потенціалу ще не використана. Створення подібної мережі обміну знань із кращою підтримкою обговорення, сприянням неупередженості, допомогою у формуванні суджень, може значно пришвидшити зростання колективного інтелекту, як глобального, так і окремих груп.

Є ще химерніша ідея: а що, як одного дня інтернет «прокинеться»? Чи може він перестати бути лише павутиною, що тримає в купі наш такий різний колективний суперінтелект, — а перетвориться на щось схоже на віртуальний череп, у якому зароджується справжній цілісний суперінтелект? (Так уявив собі появу суперінтелекту Вернор Віндж,- автор терміна «технологічна сингулярність», у 1993 році у своєму широковідомому есе[169]). Можна заперечити, що створити штучний інтелект занадто складно навіть за бажання, тому *спонтанна* поява ШІ — абсолютно неймовірна річ. Проте зовсім не обов'язково, що майбутня версія інтернету стане суперінтелектом цілком випадково. Імовірно, інтернет поступово накопичуватиме дрібні поліпшення, зроблені багатьма людьми. Наприклад, покращений алгоритм пошуку і фільтрації інформації, потужніший формат представлення даних, більше можливостей в автономних програмних агентах і ефективніший протокол взаємодії таких окремих ботів. Й от безліч таких невеликих змін врешті створять передумови для виникнення цілішої форми мережевого інтелекту. Принаймні можна теоретично допустити, що така величезна інформаційна система, забезпечена обчислювальною потужністю та іншими потрібними для

вибухоподібного зростання інтелекту ресурсами, — всім, окрім певного останнього необхідного елементу, — може раптово пробудитися, осяяна суперінтелектом, щойно останній необхідний елемент стане на своє місце. Такий сценарій дуже схожий на один з описаних нами варіантів появи штучного інтелекту загального призначення.

Висновок

Те, що створення суперінтелекту можна досягти багатьма способами, зміцнює нашу впевненість у тому, що це рано чи пізно відбудеться. Зрештою, якщо рух одного з них припиниться, ми все ще можемо спробувати інший.

Багато шляхів не означає багато пунктів призначення. Навіть якщо перша хвиля збільшення можливостей інтелекту завдячуватиме не штучному інтелекту, це ще не означатиме, що він не потрібний. Навпаки: покращення біологічного мозку або досконаліший організаційний інтелект, завдяки пришвидшенню наукового й технічного прогресу, наблизять появу радикальніших форм суперінтелекту, як-от емуляція цілого мозку та ШІ.

Я не стверджую, що не відіграє жодної ролі, як ми досягнемо створення штучного суперінтелекту. Навпаки, це може бути дуже важливо для наслідків. Те, *як* ми отримали інструмент, унаслідок *яких* дій — і, відповідно, як ми його *контролюємо*, — може не надто впливати на його функціональні характеристики, але значно вплине на те, як ми його застосуємо. Наприклад, плідна робота над біологічним або організаційним інтелектом може дати нам змогу краще оцінювати ризики і, зрештою, створити безпечний і корисний штучний суперінтелект. (Повна стратегічна оцінка цього — справа складна і мусить почекати розділу 14).

Справжній суперінтелект (а не граничне зростання поточного рівня інтелекту), найімовірніше, з'явиться внаслідок руху шляхом створення ШІ. Проте шлях цей недосліджений і непевний. Тому неможливо точно визначити ні тривалість подорожі, ні кількість перешкод, які доведеться здолати. Емуляція мозку також може виявитися швидким шляхом до суперінтелекту. Оскільки цей спосіб є переважно

технологічним і не потребує глибинних теоретичних відкриттів, можна з більшою впевненістю стверджувати, що він, зрештою, приведе до успіху. Хоч цілком імовірно, що, незважаючи на впевнене просування за допомогою емуляції, швидшим буде шлях штучного інтелекту завдяки нейроморфному ШІ на основі часткових емуляцій.

Удосконалення біологічного розуму теж імовірне — особливо за допомогою генетичної селекції. Зокрема, найперспективнішою здається технологія ітераційної ембріональної селекції. А втім, порівняно зі штучним інтелектом, біологічні поліпшення занадто повільні й поступові. Якщо вони і приведуть до появи суперінтелекту, то це буде радше слабка його форма (повернемося до цього згодом).

Очевидна можливість біологічного покращення людини зміцнює нашу впевненість у досяжності штучного інтелекту, адже покращені науковці та інженери зможуть просуватися цим шляхом швидше, ніж їхні *au naturel* попередники. Збільшення кількості людей з досконалішим розумом відіграє ключову роль у подальшому розвитку людства — особливо якщо створення штучного інтелекту припадатиме на другу половину століття.

Створення нейроінтерфейсу навряд чи зумовить появу суперінтелекту. Покращення мереж та організаційних структур може сприяти створенню слабкого суперінтелекту, але, найімовірніше, як і покращення біологічного мозку, відіграватиме допоміжну роль, збільшуючи ефективність людства у виконанні інтелектуальних завдань. Результати від такого поліпшення будуть швидшими за біологічне вдосконалення — організаційний розвиток насправді відбувається постійно і вже впливає на людство. Щоправда, вплив таких змін на здатність виконувати інтелектуальні завдання надто вузький, порівняно з біологічним удосконаленням мозку: «колективний інтелект» — це кількісний інтелект, а не «якісний інтелект» — класифікація, яку ми запровадимо в наступному розділі.

---

2 Станом на червень 2019 року найпотужнішим суперкомп'ютером є Summit, створений компанією IBM. Його пікова потужність становить $1{,}486 \cdot 10^{17}$ FLOPS. — *Прим. пер.*

3. ФОРМИ СУПЕРІНТЕЛЕКТУ

Отже що, зрештою, ми називаємо «суперінтелектом»? Тут ми ризикуємо надовго погрузнути в термінологічному болоті, але щоб надалі розуміти одне одного, маємо дати хоча б приблизне означення. У цьому розділі ми окреслимо три форми суперінтелекту і спробуємо довести їхню рівнозначність з погляду застосування. Також доведемо перевагу інтелектуального потенціалу електроніки над біологічним субстратом. Машини мають низку фундаментальних властивостей, які зумовлюють їхню першість відносно біологічних організмів — навіть удосконалених.

Машини і тварини багато в чому перевершили людину. Кажани завдяки ехолокації краще орієнтуються в темряві, калькулятори краще рахують, шахові програми виграють у нас в шахи. Кількість завдань, які комп'ютерні програми виконують краще за людину, зростатиме далі. Однак, незважаючи на користь таких спеціалізованих програм, деякі завдання під силу лише достатньо потужному штучному інтелекту, який здатний повністю замінити людину.

Ми вже намагалися визначити «суперінтелект» як інтелект, що перевершує найкращий людський розум сьогодення в багатьох узагальнених сферах знань. Проте це все ще занадто не точно. За цим означенням деякі сучасні найпотужніші системи теж можуть називатися суперінтелектом. Для глибшого аналізу розділимо поняття «суперінтелект» на окремі групи інтелектуальних суперздібностей. Існує багато варіантів такого розділення. У цій книжці ми розглянемо три форми суперінтелекту: швидкий суперінтелект, колективний суперінтелект і якісний суперінтелект.

Швидкий суперінтелект

Швидкий суперінтелект — це інтелект, схожий на людський, але значно швидший. Аналізувати таку концепцію суперінтелекту

найлегше[170]. Можемо визначити його так:

> ***Швидкий суперінтелект*** *— система, здатна робити все, що й людина, тільки значно швидше.*

Під «значно швидше» варто розуміти «на кілька порядків швидше». На цьому припинимо відточувати деталі формулювання й довіримося інтерпретації читача[171].

Уявити швидкий суперінтелект найпростіше на прикладі програмної емуляції цілого мозку, що працює на дуже потужному комп'ютері[172]. Така емуляція, яка працює в десять тисяч разів швидше за людський мозок, могла б прочитати книжку за кілька секунд і за вечір написати кандидатську дисертацію. Із пришвидшенням у мільйон разів вона б за один день закінчила роботу, для якої людині знадобилося б тисяча років[173].

Для такого швидкого розуму події зовнішнього світу здаватимуться уповільненими. Уявіть собі, ніби ваш розум працює в десять тисяч разів швидше. Ось ваш приятель ненароком впустив чашку, і ви годинами спостерігаєте, як посудина простує до підлоги, ніби комета, що неухильно рухається своїм мовчазним шляхом на зустріч із віддаленою планетою. З плином часу, усвідомлення наближення падіння поширюється сірою речовиною звивин мозку товариша, а відтак рухається периферичною нервовою системою. Ви починаєте помічати перші ознаки здивування на його обличчі — за цей час ви можете замовити йому нове горнятко чаю, переглянути кілька свіжих наукових часописів ще й встигнете трохи подрімати.

Через таке очевидне розтягнення часу матеріального світу швидкий суперінтелект віддаватиме перевагу роботі з цифровими сутностями. Середовищем його існування стане віртуальний світ, а заняттям — інформаційна економіка. Як варіант, він міг би взаємодіяти з матеріальним світом за допомогою наноманіпуляторів, адже нормального розміру кінцівки були б занадто повільними. (Характеристична частота системи обернено пропорційна її розміру[174]). Такий швидкий розум волів би мати справу з такими самими, як він, а не з брадителічними равликами-людьми.

З подібним прискоренням інтелекту, швидкість світла здається дедалі жорсткішим обмеженням для пересування, адже щораз більше

потенційно втрачених можливостей за час, витрачений на пересування або спілкування[175]. Швидкість поширення світла приблизно в мільйон разів більша за швидкість літака, тому для цифрового інтелекту, стрімкішого за людину в мільйон разів, знадобиться приблизно стільки само суб'єктивного часу, щоб переміститися в іншу точку світу, скільки й людині, що подорожує літаком. Телефонна розмова триватиме так само, щоправда буде дещо дешевшою через менший обсяг даних, що передаються. Для пришвидшення інформаційного обміну, такі супершвидкі інтелекти будуть вимушені розміщуватися поблизу. Швидкі суперінтелекти, яким необхідно інтенсивно спілкуватися (можливо, члени однієї робочої групи) для мінімізування дошкульних затримок житимуть у комп'ютерах, розташованих в одній будівлі.

Колективний суперінтелект

Іншою формою суперінтелекту є система, що виходить на новий рівень розумності завдяки ефективному поєднанню великої кількості менших інтелектів:

> ***Колективний суперінтелект*** *— система, що об'єднує велику кількість інтелектів так, що загальна продуктивність системи у великій кількості узагальнених галузей знань перевищує можливості будь-якої сучасної інтелектуальної системи.*

Концепція колективного суперінтелекту не така зрозуміла, як концепція швидкого[176]. Однак емпірично цілком осяжна. Ми не маємо прикладів людського розуму зі збільшеною тактовою частотою, зате є системи, які складаються з людей, що працюють з різною ефективністю задля спільної мети. Фірми, робочі групи, соціальні мережі, пропагандистські групи, наукові спільноти, країни. Навіть людство загалом можна за певного наближення розглядати як систему, що здатна виконувати завдання пізнання.

Спеціалізація колективного суперінтелекту — завдання, які, за їхньою природою, легко розділити на простіші дії, що не залежать одна від одної і тому можна виконувати одночасно. У виконанні таких завдань, як побудова космічного корабля або управління франшизою на продаж гамбургерів, можна знайти безліч можливостей розподілу

обов'язків: різні інженери працюють над різними елементами ракети; різні команди працюють у різних ресторанах. В академічних колах жорсткий поділ на окремі самодостатні світи дослідників, студентів, видавців, грантових комітетів, комісій з нагородження можна сприймати (проте лише з поблажливістю) як необхідне пристосування. Для того щоб велика кількість по-різному мотивованих людей і груп могла, відносно автономно орючи свою ниву, працювати на розвиток людської науки.

Колективна інтелектуальність системи збільшуватиметься, якщо збільшується кількість чи якість окремих елементів системи або покращується їхня організація[177]. Щоб перетворити будь-який сучасний колективний інтелект на суперінтелект потрібно багато ґрунтовних удосконалень. У результаті система має значно випереджати інші інтелектуальні системи в багатьох сферах знань. Без сумніву, ні нова форма проведення конференцій, що дає змогу ефективніше обмінюватися інформацією, ні новий алгоритм колаборативної фільтрації, що може точніше передбачати оцінки користувача, не здатні самі собою бути фактором виведення системи на рівень колективного суперінтелекту. Те саме можна сказати про зростання населення світу на 50 відсотків чи про новий метод навчання, що дає змогу учням вивчати за чотири години те, що раніше вивчали за шість. Для перетворення на суперінтелект потрібне значно більше зростання колективної розумової потужності.

Зверніть увагу, ми порівнюємо можливості суперінтелекту з потужністю теперішніх інтелектуальних систем — на початку XXI сторіччя. Колективний інтелект людства зріс у величезну кількість разів — з доісторичних часів і протягом всієї історії людського розвитку. Наприклад, населення світу від плейстоцену до сьогодні зросло в тисячі разів[178]. Хоча б на основі цього факту можна стверджувати, що людство досягло суперінтелектуальності, *як порівняти з плейстоценом*. До цього могли привести покращення комунікативних технологій (насамперед мовлення, але також поява міст, письма, друку), окремо або в комплексі, тому не виключено, що якби зараз з'явилася настільки ж важлива для нашого колективного інтелекту технологія, ми теж вийшли б на новий рівень — рівень колективного суперінтелекту[179].

Тут деякі читачі могли б зі мною не погодитися — сучасне суспільство не здається надто інтелектуальним. Можливо, причиною такої критичності є певне небажане політичне рішення у країні, очевидна нерозумність якого тепер пригнічує і примушує скептично ставитися до розумових здібностей сучасників. Окрім того, хіба не сучасне людство культивує споживацьке ставлення, марнує природні ресурси, засмічує навколишнє середовище, винищує видове різноманіття, а також допускає існування глобальних несправедливостей та ігнорує базові гуманістичні й духовні цінності? Проте якщо облишити порівняння вад сучасності з помилками минулого, помітимо, що наше визначення колективного суперінтелекту не стверджує, ніби суспільство з перевагою в колективному інтелекті мусить бути кращим. Ба більше, визначення також не передбачає, що таке суспільство має бути мудрішим. Мудрим можемо називати того, хто не помиляється у важливих речах. Тепер уявімо організацію з великою кількістю професійних працівників, які ефективно, скоординовано й успішно працюють над виконанням спільних завдань у різних сферах. Нехай ця організація здатна керувати більшістю підприємств, створювати будь-які технології, оптимізувати будь-які процеси. А втім, якщо їй не вдасться правильно інтерпретувати якісь, на перший погляд, незначні, аспекти в більшому масштабі — наприклад, вона невдало спрогнозує екзистенційні ризики — і вибере занадто радикальну стратегію розвитку, її діяльність завершиться повним фіаско. Водночас ця організація може мати дуже високий рівень колективного інтелекту. Можливо, достатньо високий, щоб бути суперінтелектом. Тому ми повинні опиратися бажанню приписувати розумовій досконалості всі можливі чесноти, наче й не може існувати одної доброї риси без автоматичної присутності всіх інших. Натомість, маємо розуміти, що багатофункціональна інформаційна система — інтелектуальна система — може бути не хороша і не погана. Ми поговоримо про це детальніше в розділі 7.

Колективний інтелект може бути низькоінтегрованим і високоінтегрованим за ступенем інтегрованості. Для ілюстрації низькоінтегрованого колективного суперінтелекту уявімо собі планету «Мегаземлю», що має такий рівень організації та розвитку комунікаційних технологій, що і ми, але кількість населення цієї

планети в мільйон разів більша. З популяцією «Мегаземля» матиме у стільки ж разів більший інтелектуальний потенціал, ніж наша Земля. Нехай, геній рівня Ньютона чи Ейнштейна народжується раз на десять мільярдів людей. Тоді на «Мегаземлі» житиме 700 000 таких геніїв поряд із пропорційно більшою кількістю звичайних людей. Нові ідеї та технології з'являтимуться шаленими темпами і людська цивілізація такої «Мегаземлі» буде низькоінтегрованим колективним суперінтелектом[180].

Якщо поступово збільшувати рівень інтегрованості колективного інтелекту, він згодом стане єдиним інтелектом — одним великим «розумом», а не утворенням із розрізнених і незалежних людських розумів[181]. Мешканці «Мегаземлі» могли б спробувати свідомо рухатися в цьому напрямку, покращуючи засоби зв'язку, методи координації співпраці та придумуючи кращі способи організації спільної роботи над складними завданнями. Тоді колективний інтелект, досягнувши достатньо високого рівня інтегрованості, перетвориться на «якісний інтелект».

Якісний інтелект

Тепер дамо визначення третій формі суперінтелекту.

> ***Якісний інтелект*** *— інтелектуальна система, яка своєю швидкодією не поступається людському мозку і водночас якісно розумніша.*

Як і з колективним інтелектом, якість інтелекту — дещо химерна характеристика. Ми не маємо можливості порівняти якість нашого інтелекту з якісно розумнішим інтелектом. А втім, спробуємо проілюструвати поняття деякими схожими випадками.

Насамперед, розширимо базу порівняння, додавши тварин, у яких інтелект нижчий. (Не сприймiть за видову дискримінацію. Інтелект рибок даніо чудово адаптований до потреб їхньої екосистеми. Але спробуємо, з ілюстративною метою, оцінити його можливості більш антропоцентрично — у комплексних завданнях пізнання, характерних для *людського* інтелекту). Тварини не мають мови зі складною структурою; вони не створюють і не використовують у своїй діяльності знарядь (крім деяких рудиментарних випадків). А також у

них дуже обмежені здібності до довгострокового планування та абстрактного мислення. Не всі ці обмеження розуму тварин можна пояснити повільнішим розумом або відсутністю колективного інтелекту. З погляду обчислювальної здатності, розум людини поступається деяким великим тваринам, наприклад, слонам та китам. І, хоч без колективного інтелекту розквіт технологічної цивілізації був би неможливий, не всі здібності людського розуму зумовлені лише колективним інтелектом. Багато таких переваг яскраво проявляються навіть у представників ізольованих племен мисливців-збиральників[182]. Водночас багато властивостей людського розуму за рівнем розвитку значно поступаються деяким високоорганізованим тваринам, тренованим людьми — як-от шимпанзе, дельфінам або мурахам, з їхніми великими впорядкованими спільнотами. Вочевидь, Homo sapiens своїми видатними розумовими досягненнями завдячує певним особливостям влаштування свого мозку, характерним саме для нашої генетики і не властивим іншим тваринам. Тому, використовуючи наведені спостереження, можемо визначити, що якісний суперінтелект — це інтелект якісно настільки ж вищий від людського, наскільки наразі людський інтелект перевершує інтелект слона, дельфіна або шимпанзе.

Тепер спробуємо для ілюстрації якісного інтелекту скористатися явищем, що трапляється серед людей, а саме: вузькоспецифічними когнітивними вадами, зокрема такими, що не зумовлені загальною деменцією або системним ураженням мозку. Наприклад, багато хворих на аутизм майже не здатні до соціальної взаємодії, але в інших видах розумової діяльності нічим не поступаються здоровим людям. Люди з вродженою амузією не можуть наспівати чи впізнати просту мелодію, але цілком здібні в інших сферах діяльності. У нейропсихіатричній літературі трапляються багато схожих прикладів часткових порушень функціонування мозку, які стали результатом травм або генетичних аномалій. Такі випадки свідчать, що деякі уміння людського мозку не залежать від потужності або рівня загального інтелекту — для них потрібні спеціалізовані нейронні структури. Отже, виникає ідея: а чи не можуть існувати певні *не доступні наразі інтелектуальні здібності,* за умови наявності яких певна інтелектуальна система за іншими

характеристиками ідентична знайомому нам людському мозку могла б досягти значної переваги у виконанні широкого спектра завдань.

Тому, розглянувши розумові здібності тварин і випадки окремих часткових порушень функціонування мозку людини, можемо дістати деяке уявлення про окремі інтелектуальні якості та судити про їхню важливість для розвитку людства. Можливо, без уміння утворювати складні мовні конструкції Homo sapiens був би лише одним з видів мавп, що живуть у гармонії з природою. Або навпаки, якби в нас *з'явилася* деяка здібність, яка дала нам значну перевагу, порівняно зі здатністю утворювати складні мовні конструкції, то, можливо, тоді ми стали б суперінтелектуальними.

Навпростець чи манівцями

У будь-якій із форм суперінтелект може згодом створити передумови для появи іншої форми суперінтелекту. Отже, *непрямі шляхи* досягнення будь-якої з цих форм суперінтелекту еквівалентні. Якщо виходити з переконання, що ми зможемо зрештою створити суперінтелект, то створення штучного інтелекту людського рівня теж є непрямим шляхом. Проте, ймовірно, суперінтелекту швидше вдасться із цим впоратися, ніж нам — з нашими теперішніми можливостями.

Водночас *прямі шляхи* досягнення цих трьох форм суперінтелекту порівняти складніше. Неможливо визначити точно, який з них кращий. Можливості суперінтелекту в кожній з форм залежать від міри розвинутості відповідних переваг кожного — *яка* саме швидкість швидкого суперінтелекту, *наскільки* якісно вищий якісний суперінтелект тощо. Принаймні можна стверджувати, *ceteris paribus*, що швидкий суперінтелект здатний стрімко виконувати багато послідовних дій. Натомість колективний суперінтелект може ефективно аналізувати комплексні завдання й розділяти їх на простіші паралельні дії, а також одночасно виконувати багато різнопланових завдань, що потребують найрізноманітніших умінь. Якісний суперінтелект, певно, є найбільш універсальною формою, адже міг би виконувати завдання, *безпосередньо* недоступні іншим формам суперінтелекту[183].

Визнаймо: кількість — погана заміна якості. Один геній, що самотою працює в оббитій корковим дубом спальні, може написати «У пошуках втраченого часу». Якщо заповнити офісний центр літературними рабами і змусити їх працювати, чи зможуть вони, зрештою, створити рівноцінний шедевр?[184] Отже, навіть на сучасному етапі розвитку, трапляються випадки, коли результат одного видатного розуму перевершує зусилля багатьох посередностей. З появою суперінтелекту, знайдуться завдання, що під силу лише суперінтелекту, а людям, скільки б їх не було, будуть недосяжні.

Отже, деякі завдання можна виконати лише якісним або швидким суперінтелектом, але вони не будуть під силу низькоінтегрованому колективному суперінтелекту (принаймні без поліпшення його характеристик)[185]. Зараз складно конкретизувати такі завдання, тож спробуємо узагальнити[186]. Це могли б бути комплексні завдання з багатьма взаємозалежними чинниками, які неможливо розбити на окремі незалежні етапи. Для їхнього виконання потрібна нова якість розуміння проблеми або досконаліша методологія представлення даних, надто складна, щоб її могла осягнути розумом сучасна людина. Можливо, це міг би бути якийсь художній твір або стратегічна розробка. Або наукове відкриття. Спостерігаючи за тим, як повільно та непевно просувається людство в розв'язанні «вічних питань» філософії, можна дійти висновку, що людський мозок узагалі нездатний до філософської роботи. Найвидатніші філософи людства заледве торкаються цих проблем, як собаки, що намагаються ходити на задніх лапах, — ледве досягнувши мінімуму необхідних для такої діяльності здібностей[187].

Переваги цифрового інтелекту

Навіть незначні зміни об'єму мозку та його влаштування можуть спричинити значні наслідки — бачимо це на прикладі порівняння людського інтелекту з розумом інших мавп. Штучний інтелект має ще більший потенціал зростання швидкодії та вдосконалення архітектури, а отже, може ще більше вплинути на розвиток інтелектуальності. Для нас складно — навіть неможливо — уявити собі здібності суперінтелекту, але завдяки цифровим інформаційним

технологіям, ми можемо приблизно уявити собі простір цих можливостей. Легше всього оцінити апаратні переваги:

- *Швидкість обчислень.* Біологічні нейрони працюють із частотою 200 Гц — на сім порядків менше, ніж сучасні процесори (близько 2 ГГц)[188]. Мозок нездатний швидко виконувати безліч послідовних операцій, але може виконувати багато чого одночасно[189]. (Робота, яку мозок здатний зробити за секунду, вкладається в не більш як сотню послідовних операцій — зазвичай лише кілька десятків). Більшість важливих прикладних алгоритмів у комп'ютерному програмуванні дуже складно паралелізувати. Мозок міг би краще виконувати завдання пізнання, якби, окрім природних здібностей та одночасного розпізнавання образів, мав інтегровану здатність швидкого послідовного оброблення інформації.
- *Швидкісний внутрішній обмін даними.* Швидкість поширення сигналу аксоном становить близько 120 м/с, тоді як оптичний чи електромагнітний сигнал між ядрами процесора поширюється зі швидкістю світла (близько 300 000 000 м/с)[190]. Саме швидкість поширення сигналу в живому мозку обмежує розмір, за якого він здатен працювати як єдине ціле. Наприклад, максимальний розмір ділянки мозку, у межах якої час передавання інформації в обидва боки не перевищує 10 мс, становить 0,11 $м^3$. Натомість ідентична електронна система може мати розмір невеликої планети — $6{,}1 \cdot 10^{17}$ $м^3$, тобто на вісімнадцять порядків більше[191].
- *Кількість обчислювальних елементів.* Людський розум містить трохи менше ніж сто мільярдів нейронів[192]. Розмір людського мозку в три з половиною рази перевищує мозок шимпанзе (але становить одну п'яту розміру мозку кашалота)[193]. Кількість нейронів біологічного мозку, насамперед, обмежена розміром черепної коробки та можливостями метаболізму, проте розмір мозку можуть обмежувати також інші чинники (як-от охолодження, час дозрівання, тривалість поширення сигналу — див. попередній пункт). Водночас розмір комп'ютера майже необмежений[194]. Суперкомп'ютер може бути розміром як склад, але за допомогою високошвидкісного з'єднання з віддаленими ресурсами його можна розширювати далі[195].
- *Можливості зберігання даних.* Тимчасова пам'ять людини може одночасно зберігати не більше чотирьох-п'ятьох інформаційних

елементів[196]. Немає сумнівів, електронна пам'ять цифрового інтелекту зможе вмістити значно більше, але не зовсім справедливо порівнювати тимчасову пам'ять мозку з оперативною пам'яттю комп'ютера. Завдяки електронній пам'яті цифровий розум зможе з легкістю інтуїтивно осягати складні взаємозв'язки понять, для розуміння яких людині потрібні тривалі міркування[197]. Довгострокова пам'ять людини також обмежена, проте точна її кількість невідома, адже досі жодній людині протягом життя не вдавалося заповнити її до кінця — через надто малу швидкість засвоєння людиною інформації. (Згідно з результатами однієї з оцінок, дорослий мозок містить близько одного мільярда біт інформації — на кілька порядків менше, ніж може містити дешевий смартфон[198]). Отже, обсяг зберігання інформації та швидкість доступу до неї в цифрового мозку може бути значно більший, ніж у біологічного.

- *Надійність, тривалість життя, органи чуття тощо.* Штучний інтелект може мати інші апаратні переваги. Наприклад, транзистори надійніші за біологічні нейрони[199]. Оскільки зашумлені обчислення вимагають надлишкових алгоритмів кодування, у яких для кодування одного біту інформації використовуються кілька елементів, цифровий мозок матиме переваги над біологічним у точності та надійності. Біологічний мозок втомлюється після кількох годин роботи, а після кількох десятиліть суб'єктивного часу починає деградувати. Мікропроцесори не мають таких обмежень. Завдяки існуванню багатьох різноманітних сенсорів можна буде легко забезпечити штучний інтелект даними в будь-якій кількості. Різноманітні технології дають змогу переналаштовувати й оптимізувати апаратні засоби для виконання конкретного типу завдань, тоді як мозок переважно статичний і змінюється дуже повільно (хоча характеристики окремих синаптичних зв'язків можуть змінюватися досить швидко — протягом днів)[200].

Обчислювальна здатність біологічного мозку досі може конкурувати із сучасними комп'ютерами, однак деякі суперкомп'ютери вже досягли межі оціночної потужності людського мозку[201]. Апаратні засоби

продовжують удосконалювати, а граничні межі їхніх можливостей значно перевищують здатності біологічних тканин.

Цифровий розум також успадкує переваги програмного забезпечення:

- *Легкість зміни.* Параметри програми значно легше змінити, ніж параметри нейронних зв'язків. Наприклад, за допомогою емуляції цілого мозку можна було б легко експериментувати з кількістю нейронів або їхньою чутливістю в тій чи тій ділянці мозку. У живому мозку запровадити такі зміни значно складніше.
- *Дублювання.* За наявності апаратних можливостей програмне забезпечення дає змогу легко дублювати або копіювати сутності. Натомість біологічний мозок відтворюється дуже повільно, і кожний новий екземпляр перебуває в первісному стані, не зберігає пам'яті свого прототипу й потребує тривалого навчання.
- *Узгодження мети.* Неефективність співпраці в людських колективах значно спричинена розбіжностями у розумінні кінцевої мети окремими їхніми членами. Допоки не винайдено способу масового навіювання слухняності, наприклад, медикаментозним способом або за допомогою генетичної селекції. Цього можна уникнути завдяки створенню «клану клонів» (паралельного виконання ідентичних або майже ідентичних програмних потоків, що працюють над розв'язанням спільної задачі).
- *Спільний доступ до даних.* Біологічний розум потребує тривалого навчання і тренування, тоді як цифровий розум міг би набувати нових знань та вмінь просто завантажуючи файли з ними в пам'ять. У мільярдній популяції окремих екземплярів штучного інтелекту кожен дізнавався б, чого навчилися його «колеги» протягом, скажімо, останньої години, періодично синхронізуючи з ними свою базу даних. (Щоправда, пересилання образів ментальних процесів потребує стандартизації форматів представлення даних. Тому не всі типи ШІ будуть здатні безпосередньо використовувати високорівневе представлення ментальних процесів одне одного. Зокрема, першим поколінням емуляцій цілого мозку це не буде під силу).
- *Нові модулі, модальності та алгоритми.* Візуальне сприйняття не потребує від нас розумового напруження й відбувається

автоматично, на відміну від, скажімо, розв'язання задач з геометрії. Незважаючи на те, що ментальне відтворення тривимірних об'єктів реального світу та їхніх сутностей із двовимірних світлових зображень на наших сітківках насправді потребує колосальних обчислень. Натомість для нас усе здається простим, бо ми маємо спеціальні нервові структури для оброблення цієї інформації. Перетворення даних відбуваються автоматично і не торкаються нашої свідомості й розумових процесів у ній. Сприйняття музики, мовлення, соціальність та інші ментальні процеси теж видаються нам «природними» і також, імовірно, працюють завдяки окремим підсистемам мозку. Штучний розум, який матиме виділені підсистеми для виконання інших, важливих у сучасному світі, завдань — як-от розроблення електроніки, програмування, бізнес-планування — отримає значну перевагу над людьми, що змушені залучати для цього неповороткі здібності загального інтелекту. Можливо, також з'являтимуться нові алгоритми, які зможуть використати переваги цифрових систем, здатних до швидкого послідовного оброблення інформації.

Отже, штучний інтелект, апаратна і програмна його частини обіцяють *у перспективі* величезні можливості[202]. Але чи близька ця перспектива? Саме цим питанням ми зараз і займемося.

4. КІНЕТИКА ІНТЕЛЕКТУАЛЬНОГО ВИБУХУ

Щойно машини досягнуть певної «людяності» у своїй здатності до загальних міркувань — як швидко після цього вони набудуть якостей суперінтелекту? Буде це повільний, поступовий і тривалий перехід? Чи, може, раптовий, як вибух або стрибок? У цьому розділі проаналізуємо кінетику такого переходу як функцію від оптимізаційної сили системи та її консервативності. Також пригадаємо, що нам відомо про роль і властивості цих факторів у контексті інтелекту осяжних для нас масштабів.

Час і швидкість зростання

Розумові здібності машин *наразі* значно менші від людських. Тож, якщо *згодом* вони перевершать біологічний мозок у загальному інтелекті, постає запитання — як швидко може відбутися таке зростання? Питання, яке ми дослідимо в цьому розділі, дещо відрізняється від того, що ми розглядали в розділі 1. Варто усвідомити цю різницю. У розділі 1 нас цікавило, як швидко людство досягне створення штучного інтелекту зі здібностями до загальних міркувань, що відповідають рівню сучасної людини. Тепер же запитаємо: *за умови, що таку систему уже створено, скільки триватиме її перехід до штучного суперінтелекту?* Можна вважати, що мине багато часу, перш ніж машини зрівняються розумом з людьми, або стати в цьому питанні на агностичну позицію, водночас будучи переконаним: щойно машини досягнуть рівня людини, перехід до суперінтелекту буде дуже швидким.

Можливо варто для наочності зобразити ці процеси схематично, хоч для цього доведеться поки що відкинути деякі важливі деталі й характеристики. Розглянемо графік зміни інтелектуальних можливостей штучних інтелектів із часом (див. рисунок 7).

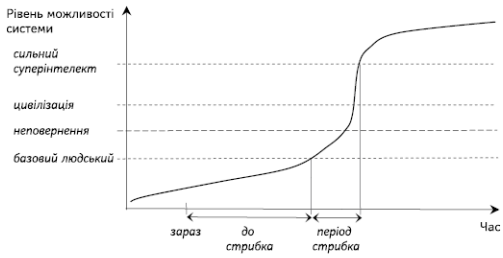


Рисунок 7. **Динаміка росту ШІ**
Важливо розрізняти запитання: «Чи буде стрибок ШІ, а якщо так, то коли?» і «Коли відбудеться стрибок, то наскільки стрімким він буде?». Можна вважати, що до початку такого росту можливостей ШІ ще далеко, але щойно він почнеться, то відбуватиметься дуже стрімко. Ще одне цікаве питання (яке, проте, не подано на цьому графіку): «Яку частку у зростанні ШІ займає участь світової економіки?». Ці питання взаємопов'язані.

Горизонтальна лінія, позначена як «базовий людський», визначає рівень розумових здібностей, еквівалентний рівню середньостатистичної дорослої людини, що здатна користуватися технологіями та інформаційними ресурсами, типовими для розвинутих країн. Зараз, за оцінкою загального інтелекту, найрозвинутіші системи ШІ перебувають значно нижче.

Одного дня штучний інтелект досягне цього рівня (попередньо зафіксуємо його — скажімо, у 2014 році — навіть, якщо рівень можливостей людини до того дня також зросте). Саме цей момент вважатимемо початком стрибка ШІ. Інтелектуальні можливості систем зростатимуть далі і невдовзі після того досягнуть рівня можливостей всього людства загалом (також зафіксованого на теперішній час). Назвемо його рівень «цивілізації». Згодом можливості систем досягнуть «сильного суперінтелекту» — розумових здібностей, що значно перевищують весь інтелектуальний потенціал людства. Після цього стрибок вважатиметься закінченим, хоч розумові здібності систем можуть зростати й далі. На якомусь етапі зростання інтелектуальні здібності систем перетнуть межу, що ми позначили як рівень «неповернення». Це такий рівень здібностей, понад який покращення відбуватимуться самовільно — завдяки зусиллям систем[203]. (Існування такої межі важливе для розгляду оптимізаційної сили й вибуховості процесів, далі в цьому розділі).

Тепер, маючи графік, можемо розглянути три гіпотетичні сценарії такого переходу — повільний, швидкий і помірний — залежно від стрімкості кривої змін.

Повільний

За повільного темпу змін перехід на новий рівень відбуватиметься довго: десятиліття або навіть століття. У таких умовах людство матиме можливість підготувати політичні засади майбутньої взаємодії. Буде вдосталь часу, щоб спробувати різні підходи. Досвід накопичуватиметься, з'являться експерти. Відбуватимуться протести незадоволених змінами прошарків населення. За необхідності, запровадять нові заходи безпеки та створять системи нагляду за ШІ. Щоб запобігти використанню ШІ для нарощування озброєнь, країни матимуть час на узгодження міжнародно-правових обмежень та технічних вимог до запобіжних механізмів. Завдяки повільному темпу змін більшість попередніх заходів, здійснених до початку переходу, може бути переглянуто та вдосконалено.

Швидкий

У швидкому сценарії нарощування потужностей ШІ відбудеться дуже стрімко: за хвилини, години чи дні. Така швидкість розвитку подій не залишає людству можливості вплинути на ситуацію. Ніхто не помітить нічого незвичайного, просто в якийсь момент ми зрозуміємо, що вже програли. Доля людства в такому разі залежатиме від попередніх заходів, вжитих до початку змін. Навіть якщо такий стрибок триватиме дні, дії людей — вчинки окремих осіб, на кшталт відкривання «ядерної валізки» — матимуть безсистемний характер чи будуть наперед передбачені або навіть спровоковані ШІ.

Помірний

У помірному темпі перехід до рівня суперінтелекту відбуватиметься протягом місяців чи років. Такий сценарій дає людству можливість діяти у відповідь, але залишає замало часу, щоб проаналізувати ситуацію, зважити варіанти та консолідувати ресурси. Люди не встигнуть розробити чи впровадити нові системи (політичні реформи, системи спостереження, мережеві протоколи безпеки): доведеться покладатися лише на вже наявні механізми.

Повільний розвиток ШІ дасть вдосталь часу на поширення новин. За помірного сценарію обмежене коло втаємничених осіб може приховувати інформацію про зростання рівня ШІ, як буває у сфері державних військових досліджень. Звісно, комерційні розробки, невеликі групи науковців, навіть «купка хакерів у підвалі», теж здатні до конспірації, але такі підпільні приватні проекти від початку будуть «під ковпаком» у держави, якщо перспектива виходу з-під контролю ШІ вважатиметься загрозою національній безпеці. Тому держава (або інша домінантна сила) зможе націоналізувати або зупинити будь-який проект, здатний у перспективі здійснити такий стрибок. За швидкого сценарію розвитку ШІ часу на дії у відповідь не буде, проте навряд чи хтось знатиме про це до його завершення. Але *до* початку такого стрибка ШІ, за появи передумов для нього, деяка зовнішня сила все-таки змогла б втрутитися й запобігти йому.

За помірної швидкості змін деякі індивіди чи групи можуть спробувати скористатися ними у власних інтересах, що водночас може стати причиною геополітичної, соціальної та економічної нестабільності. Такий розвиток подій ще більше ускладнить завдання організації продуманих контрзаходів. Крім того, він може стати передумовою ще радикальніших подій. Наприклад, за помірного сценарію під час поступового заповнення ринків праці недорогими і здібними емуляціями чи ШІ легко уявити собі проведення звільненими працівниками масштабних акцій протесту. Вони вимагатимуть в урядів збільшити розмір допомоги з безробіття, призначити мінімальне забезпечення всім біологічним громадянам, запровадити спеціальні податки працедавцям, які використовують працю ШІ, чи встановити мінімальні розміри оплати для людської праці. Потрібні глибокі законодавчі зміни й відповідні управлінські практики, щоб такі інструменти справді дали тривалий ефект. Усі ці проблеми властиві і для поступовішого зростання ШІ, але в помірному сценарії, завдяки нестабільності та швидким змінам, маргінальні сили можуть скористатися нагодою й набути непропорційного впливу.

Декому з читачів може видатися, що найбільш імовірним сценарієм є повільний сценарій, натомість, помірний — можливий, а швидкий — цілком неймовірний. Може здаватися безглуздою ідея, що ієрархія світу може докорінно змінитися, а людство — втратити свою

інтелектуальну першість упродовж кількох годин. Історія людства не знала настільки глибоких і швидких зрушень. Найближчі аналоги — аграрна та промислова революції — тривали значно довше (перша — від століть до тисячоліття, друга — від десятиліть до століття). Прецедентів такого масштабу і стрімкості зрушень, які можуть відбутися внаслідок швидкого чи помірного сценарію, поза міфологією чи релігією, не існує: а отже, їхня ймовірність близька до нуля[204].

А втім, далі в цьому розділі наведемо деякі підстави твердити, що саме повільний сценарій — найменш імовірний з усіх. Якщо стрибок інтелекту відбудеться, він, найімовірніше, буде стрімким.

Спершу представимо швидкість зростання інтелектуальності системи як функцію двох параметрів: величини «оптимізаційної сили», або якісно-зваженого творчого зусилля, спрямованого на збільшення інтелектуальності системи, та сприйнятливості системи до спрямованої на неї оптимізаційної сили. Назвемо величину, обернену до сприйнятливості, «консервативністю» системи і запишемо:

$$\text{Швидкість зміни інтелектуальності} = \frac{\text{оптимізаційна сила}}{\text{консервативність}}$$

Поки не визначимося, як вимірювати інтелектуальність, творче зусилля і консервативність, цей вираз матиме лише описову цінність. Однак з нього також бачимо, що інтелектуальність системи зростатиме достатньо швидко, якщо до системи прикладене значне оптимізаційне зусилля, за умови незначного опору його дії (консервативності системи) або що зусилля достатнє, а консервативність дуже низька. Знаючи міру творчого зусилля, спрямованого на вдосконалення певної системи, і швидкість прогресування її характеристик, можемо визначити й консервативність цієї системи.

Також очевидно, що величина оптимізаційної сили неоднакова для різних систем і може змінюватися із часом. Консервативність також може залежати від міри оптимізованості системи. Адже найпростіші покращення зазвичай здійснюють на початку, як зривають плід, що найнижче висить, а ось наступні оптимізації даються важче. За законом спадної віддачі консервативність системи згодом зростає. Проте трапляються каскадні оптимізації, коли одне покращення спрощує здійснення наступного. Спочатку складати пазл нескладно —

просто знайти кутові елементи та зібрати краї. Але далі підбирати частинки стає важче і консервативність системи зростає. Однак у кінці поле пошуку зменшується й підбирати елементи знову стає легше.

Отже, щоб поглибити наш аналіз, спробуємо дослідити поведінку консервативності та оптимізаційної сили на різних етапах стрибка інтелектуальності. Цьому присвятимо наступні кілька сторінок.

Консервативність

Розпочнемо з консервативності. Поведінка консервативності різних типів інтелектуальних систем у процесі їхнього розвитку до суперінтелектуальності матиме свої особливості. Для цілісності підходу ми спочатку розглянемо консервативність розвитку природних інтелектуальних систем. Як з'ясуємо, така консервативність достатньо висока. Потім повернемося до способів досягнення суперінтелектуальності, що передбачають створення штучного інтелекту. Їхня консервативність здається переважно низькою.

Розвиток інтелектуальності природних систем

Вплив покращень медичного забезпечення та дієти на розвиток інтелекту згодом зменшується[205]. Усунення значних порушень живлення мозку дає найбільший результат, але наразі такі порушення трапляються лише в найбідніших місцевостях. Покращення і без того доброго харчування дає небагато. У сучасній освіті теж, напевно, спостерігається спадання віддачі. Частка талановитих людей, яким бракує якісної освіти, досі значна, проте знижується.

Фармацевтичні засоби покращення інтелекту невдовзі, імовірно, матимуть певний вплив. Проте коли найпростіші способи впливу вичерпаються — наприклад, покращення енергозабезпечення мозку, концентрації, роботи довготривалої пам'яті, — досягти подальших результатів буде все важче. Але, на відміну від дієт та загального медичного забезпечення, ефективність таких препаратів може спочатку зростати. Нейрофармацевтичні методи й засоби впливу на мозок потребують глибшого дослідження. Через загальний брак уваги до поліпшувальної медицини сфера наразі демонструє низькі темпи розвитку. Якщо фармакологія та нейронаука і далі оминатимуть

питання покращення інтелекту, то перші порівняно легкі способи такого покращення з'являться саме з появою інтересу до ноотропних препаратів[206].

Для процесу генетичного покращення інтелекту крива зміни консервативності, як і в разі ноотропних препаратів, схожа на латинську «U», але потенціал такого покращення значно більший. Консервативність селективного схрещування одразу висока, а тому дуже важко досягти помітних результатів приросту інтелекту. Тим більше підтримувати ефект протягом кількох поколінь. Справи можуть піти краще, коли з'являться дешеві й ефективні технології генетичного аналізу і селекції (зокрема ітераційна ембріональна селекція). Це дасть змогу охопити всі варіації інтелектуально значущих алелей генетичного пулу людства. Коли всі такі алелі будуть залучені, досягати подальших покращень стане значно складніше. Тоді консервативність зростатиме разом із потребою в інноваційніших підходах. Використання генетичних механізмів, а зокрема потреба в очікуванні дозрівання, зумовлює обмеження швидкості зростання інтелекту. Тому можливість швидкого або помірного сценарію в цьому разі — під питанням[207]. Застосування ж ембріональної селекції лише в контексті екстракорпорального запліднення обмежить поширення технології і в такий спосіб ще більше уповільнить розвиток.

Консервативність руху до створення суперінтелекту через покращення взаємодії мозку з комп'ютером теж від початку дуже висока. Знизитися вона може лише в малоймовірному разі появи простого способу імплантації та високорівневої інтеграції з мозком. Складність такого способу не менша від складності емуляції мозку та створення ШІ, адже, у взаємодії мозку та комп'ютера левова частка потрібних технологій має з'явитися саме з «комп'ютерного боку».

Оптимізація мережевих і організаційних утворень має досить високу консервативність. Для подолання цього опору потрібні значні зусилля, водночас річний приріст можливостей становить не більше кількох процентів[208]. Ба більше: завдяки запровадженим змінам ефективність організації покращується, але адаптованість до зовнішніх умов зменшується. Тому необхідно докладати зусиль, навіть для того, щоб організація просто не деградувала. Радикальне пришвидшення темпів розвитку організацій здається досяжним, але навіть за найбільш

оптимістичним варіантом темпи переходу до суперінтелектуальності відповідають повільному сценарію. Тому що людські організації обмежені своїми темпом та масштабом мислення. Інтернет залишається дуже перспективним середовищем зростання колективного інтелекту, з помірним рівнем консервативності — темп розвитку досить швидкий, але потребує значних зусиль. І згодом, напевно, потребуватиме ще більших зусиль, коли легші завдання (наприклад, пошукові системи й електронна пошта) буде виконано.

Емуляція мозку та штучний інтелект

Труднощі емуляції цілого мозку складно передбачити. Однак маємо віху, на яку можемо орієнтуватися: емуляція мозку комахи. Вона є на підвищенні і, досягнувши її, побачимо значний відрізок дороги попереду та зможемо оцінити консервативність руху за допомогою масштабування технології від мозку комахи до мозку людини. (Емуляція мозку невеликого ссавця буде ще ціннішою віхою та точкою огляду майбутнього цієї технології — з полем зору аж до емуляції цілого мозку людини). Створення ШІ натомість не віщує таких очевидних віх і висот. Цілком можливо, що цей шлях перетвориться на загублену, ледь помітну стежину, якою доведеться довго блукати серед густого лісу, поки раптом завдяки гучному відкриттю людство не опиниться на вільній та рівній дорозі до мети.

Повернемося до тих двох питань, різницю між якими ми підкреслювали на початку розділу: як створити штучний інтелект рівня людини? Як потім досягти значно вищого від людського рівня? Від відповіді на перше питання залежить час, який залишився до початку стрибка, що ми розглядали в цьому розділі. Друге ж питання напряму стосується форми кривої стрибка і його тривалості. Хоч, можливо, хотілося б вірити, що перейти від людського рівня інтелекту до надлюдського значно важче, маємо враховувати, що такий перехід матиме непоганий початковий імпульс і передумови. Адже потрібно буде покращувати і без того досить здібну систему. До того ж, імовірно, консервативність системи знижуватиметься із наближенням до людського рівня інтелекту.

Візьмемо, наприклад, емуляцію цілого мозку. Створити її значно складніше, ніж покращити вже створену. Для цього потрібно розв’яза-

ти численні проблеми технологічного характеру, зокрема створити потужні засоби для сканування та розпізнавання зображень. Таке завдання потребуватиме також значних капіталовкладень — ймовірно, знадобиться цілий парк промислових високопродуктивних сканерів. Натомість, покращити наявну емуляцію можна вдосконаленням алгоритмів і структур даних: змінами до програмного забезпечення. Це може виявитися значно простіше, ніж вдосконалення технології переведення зображень у первинну програмну структуру. Програмісти зможуть легко змінювати кількість нейронів у різних частинах емуляції, щоб дослідити, як це вплине на її продуктивність[209]. Вони зможуть також оптимізувати код та винаходити простіші математичні моделі важливих функцій окремих нейронів і невеликих ділянок нейронних мереж. Якщо засоби сканування і трансляції винайдуть після високопродуктивних комп'ютерів, то ефективність відтворення навряд чи матиме значення. Натомість буде достатньо можливостей для програмної оптимізації. (Імовірно, можна буде значно вдосконалити архітектуру, проте це наблизить кінцевий результат до ШІ, а подібність зі структурами мозку зменшиться).

Іншим способом покращити вихідний код емуляції мозку може бути сканування та відтворення ще одного, можливо, розумнішого, мозку. Також збільшити продуктивність можна, адаптувавши організаційні структури та процеси емуляції органічного мозку до особливостей цифрової техніки. З появою на ринку праці емуляцій управлінці отримають широке поле для інновацій та розвитку у сфері менеджменту.

Консервативність, знизившись після появи перших успішних емуляцій, знову почне зростати. Найочевидніші недоліки реалізацій буде виправлено, алгоритми — оптимізовано, нові архітектурні поліпшення — випробувано. Нові цикли сканування не додаватимуть до бібліотеки програмних класів нейроструктур нові класи. Оскільки емуляції можна буде копіювати, тренувати в різних сферах знань, і все це — зі швидкістю, властивою електронним пристроям, не буде необхідності у скануванні великої кількості людських мозків. Тоді, імовірно, одного сканування буде достатньо.

Іншою причиною імовірного зростання консервативності може бути організований спротив емуляцій або їхніх біологічних захисників з

метою ухвалення законодавчих заборон її експлуатації, копіювання, використання в певних видах експериментів, а також захисту прав емуляцій, встановлення мінімальної заробітної плати тощо. Однак може статися, що політичний вектор буде спрямований у протилежному напрямку і консервативність знижуватиметься ще більше. У розпалі конкуренції економічні аргументи переважать моральні і спершу обмежене використання праці емуляцій перетвориться на відверту експлуатацію.

У випадку штучного інтелекту (не емуляції цілого мозку) складність переходу від інтелекту рівня людини до суперінтелектуальності залежить від особливостей конкретної реалізації системи. Різні архітектури можуть мати дуже відмінні рівні консервативності.

Імовірні ситуації, у яких консервативність буде дуже низькою. Наприклад, якщо стримувати появу ШІ буде тривалий пошук деякого програмного рішення, щойно воно з'явиться, ШІ здійснить величезний стрибок у розвитку, можливо, оминувши проміжні рівні, — одразу до суперінтелекту. Консервативність також може бути дуже низькою в разі, коли система може нарощувати інтелектуальні здібності через два різні типи діяльності. Наприклад, уявімо систему ШІ, що складається з двох підсистем: спеціального інтелекту — для вузькоспеціалізованих завдань у певній галузі, і загального інтелекту. Нехай друга підсистема перебуває на нижчому рівні розвитку і її внесок в загальну продуктивність низький, тому що варіанти рішень, які вона продукує, завжди менш оптимальні, ніж рішення підсистеми спеціального інтелекту. Уявімо тепер, що на другу підсистему діє певна оптимізаційна сила і її можливості починають швидко зростати. Спочатку це не впливає на загальну продуктивність системи, що свідчить про значну консервативність. Та, щойно можливості підсистеми загального інтелекту перетинають певну межу й генеровані нею рішення починають переважувати рішення підсистеми спеціального інтелекту, загальна продуктивність системи починає зростати тими самими темпами, що і загальний інтелект. Хоч величина прикладеної оптимізаційної сили залишається незмінною: таке зростання свідчить про зниження консервативності.

Також імовірно, що через звичку розглядати інтелект із антропоцентричного погляду ми недооцінимо покращення систем з

інтелектом нижчим, ніж людський рівень. А отже, переоцінимо їхню консервативність. Елізер Юдковський, теоретик систем ШІ та автор багатьох праць стосовно майбутнього ШІ, описує це так (див. також рисунок 8):

> *Прогрес систем ШІ може здаватися значним виключно через антропоморфізм, схильність людини вважати «сільського дурника» та «Ейнштейна» протилежними полюсами шкали інтелектуальності. Тоді як насправді відстань між ними на шкалі загальної розумності життєвих форм майже непомітна. Будь-хто менш розумний, ніж найдурніша людина, для нас здається просто «дурним». Якщо уявити, як стрілка «ШІ» на шкалі розумності потихеньку проминає мишей, шимпанзе — такий ШІ для нас усе ще «дурний», бо не розмовляє та не пише наукові статті, — а потім, протягом місяця, стрілка долає ледь помітний відрізок від інфраідіота до ультра-Ейнштейна*[210].

Рисунок 8. **Менш антропоморфна шкала?**
З нашого антропоцентричного погляду різниця між розумовими здібностями дурної та розумної людини здається значною, проте, у ширшій перспективі, розум цих двох майже ідентичний[211]. Створити ШІ з розумовими здібностями «сільського дурника» безсумнівно важче і довше, ніж потім поліпшити його можливості до надлюдського рівня.

Тому важко оцінити складність алгоритмічних покращень, які потрібно здійснити, щоб перший ШІ досягнув людського рівня загального інтелекту. Можемо тільки передбачити певні передумови низької консервативності деяких алгоритмів. А втім, якщо консервативність окремих алгоритмів буде дуже високою, це не означає, що загальна консервативність ШІ не буде низькою. Адже можуть існувати легші способи покращення інтелектуальності системи, не пов’язані з оптимізацією алгоритмів. Ще є ще два фактори впливу на продуктивність системи: контент (знання) та апаратне забезпечення.

Спочатку розглянемо можливості покращення контенту. Під «контентом» розумітимемо неалгоритмічні компоненти програмного забезпечення. Це можуть бути бази даних досвіду, бібліотеки

спеціалізованих умінь, реєстри описових даних. Такий ресурс може бути джерелом певного покращення, хоч у більшості систем немає чіткої межі між алгоритмічною структурою та даними. Інакше кажучи, інтелектуальні здібності системи залежать не тільки від умінь системи, а й від знань.

Наприклад, візьмемо сучасну систему ШІ: TextRunner (дослідний проект Вашингтонського університету) чи суперкомп'ютер Watson компанії IBM (система, що перемогла в телевікторині Jeopardy!). Ці системи, аналізуючи тексти, можуть вивчати семантичні особливості слів. Вони здатні отримувати достатньо семантичної інформації з аналізу тексту та природного мовлення, щоб відповідати на запитання, хоч не розуміють прочитане так, як його розуміє людина. Вони вчаться на досвіді, розширюючи знання про поняття з кожним новим випадком його вживання в тексті. Такі системи швидкі та можуть вчитися в автоматичному режимі (тобто визначати приховану структуру у нерозміченому масиві даних, не маючи оцінки похибки або підкріплення від учителя) та легко розширюються. Наприклад, TextRunner у процесі роботи аналізує вміст 500 мільйонів веб-сторінок[212].

Тепер уявіть собі віддаленого нащадка такої системи, який здатен настільки ж швидко опрацьовувати тексти, але водночас розуміє прочитане на рівні людини десяти років. (Вочевидь, таке завдання — еквівалентне створенню повноцінного ШІ). Така система могла б думати набагато швидше та мала б кращу пам'ять, ніж доросла людина, але знала б значно менше. Рівень інтелекту загального призначення такої системи приблизно відповідав би рівню людського інтелекту. Але консервативність, з погляду покращення контенту, була б дуже низька — достатня для стрибка інтелектуальності. За кілька тижнів така система здатна засвоїти знання всієї Бібліотеки Конгресу. Відтак система знатиме набагато більше, ніж будь-яка людина, і думатиме значно швидше, отже — стане суперінтелектом (щонайменше — слабким).

Отже, система, маючи змогу засвоювати інформаційний капітал, накопичений багаторічною діяльністю людства — наприклад, через інтернет, — зможе швидко нарощувати інтелектуальну потужність. Якщо певна система, досягнувши інтелектуального рівня людини,

отримає доступ або можливість засвоювати дані через інтернет, то консервативність такого ШІ буде досить низька, навіть за відносної складності внесення алгоритмічних удосконалень в її архітектуру.

Концепція контент-консервативності релевантна і для емуляцій. Перевага високошвидкісної емуляції не тільки в тому, що вона працює швидше за біологічний мозок, а й у тому, що вона може накопичувати більше актуального контенту — вміння та навички — за одиницю часу. Однак для того щоб сповна скористатися цією можливістю, емуляція повинна мати вдосталь пам'яті. Мало користі від прочитання всієї бібліотеки, якщо забув усе про абляцію, поки дійшов до аболіціонізму. Системи ШІ, найімовірніше, матимуть удосталь пам'яті, тоді як емуляції успадкують проблеми з місткістю від свого оригіналу. Таким емуляціям знадобляться алгоритмічні вдосконалення, щоб мати можливість вчитися безкінечно.

Дотепер ми описали консервативності архітектури та контенту, тобто *програмних* аспектів штучного інтелекту, який досяг рівності з людиною. Тепер розглянемо третій спосіб збільшити можливості ШІ: покращити апаратне забезпечення. Якою буде консервативність у цьому разі?

Маючи розумну програму (емуляцію чи ШІ), можна розвивати *колективний інтелект*, просто використовуючи додаткові комп'ютери, щоб запустити більше екземплярів цієї програми[213]. Можна також розвивати *швидкісний інтелект*, запускаючи програму на щоразу швидшому «залізі». Якщо програму оптимізовано для паралелізації, можна пришвидшити швидкісний інтелект, запустивши його на більшій кількості процесорів. Це доречніше для емуляцій, чия архітектура більше надається до паралелізації. Проте багато програмних ШІ складаються з функціональних одиниць, деякі з них теж можуть працювати паралельно. Нарощення потужності комп'ютера може бути корисним і для *якісного інтелекту*, втім цей спосіб не такий очевидний[214].

А отже, консервативність покращення колективної (і, можливо, якісної) форми програмного ШІ рівня людини, найімовірніше, низька. Основна проблема — це досягнення більшої обчислювальної потужності. Існують кілька способів нарощення апаратної потужності, втім кожен із них має свої часові рамки.

У короткочасній перспективі приріст обчислювальної потужності лінійно залежить від фінансових затрат: удвічі більше екземплярів програми потребують удвічі більше комп'ютерів, які вдвічі більше коштують. Сучасні хмарні сервіси дають зекономити час, потрібний на розгортання апаратного парку. Однак з міркувань безпеки деякі проекти можуть віддати перевагу локальному масштабуванню. (Можливий також сценарій нарощування обчислювальної потужності за допомогою створення бот-мереж[215]). Доступність масштабування залежить від початкових потреб системи в ресурсах. Якщо система працює на звичайному комп'ютері, усього за мільйон доларів можна збільшити обчислювальну потужність у тисячі разів. Якщо ж первісна система виконується на суперкомп'ютері подальше нарощення ресурсів обійдеться значно дорожче.

У довгостроковішій перспективі зі збільшенням частки глобальної обчислювальної потужності, задіяної в роботі штучних інтелектів, вартість апаратних ресурсів може почати зростати. Наприклад, у разі експлуатації емуляцій у режимі вільного ринку вартість ресурсів для підтримки роботи одного екземпляра, через попит інвесторів, може зростати. Аж поки не наблизиться до розміру прибутку, що його генерує такий екземпляр (щоправда, якщо існуватиме лише один успішний проект, то він, як монопсоніст, матиме змогу скуповувати обчислювальні потужності за нижчою ціною).

Із плином часу пропозиція на ринку обчислювальних ресурсів зростатиме. Попит стимулюватиме зростання виробництва на наявних заводах і побудову нових. (Також одноразове зростання обчислювальної потужності можливе завдяки використанню спеціалізованих мікропроцесорів[216]). Зрештою, стрімке зростання обчислювальної потужності дедалі швидше обертатиме млин штучного розуму. Швидкість розвитку комп'ютерних технологій описана відомим законом Мура. За однією з його варіацій, кількість обчислювальної потужності, яку можна придбати за один долар, подвоюється приблизно кожні вісімнадцять місяців[217]. Немає жодних гарантій, що такі темпи зберігатимуться до створення ШІ рівня людини, але комп'ютерні технології можуть розвиватися безперешкодно, поки не досягнуть яких-небудь фундаментальних фізичних меж.

Тому немає підстав вважати, що консервативність апаратного забезпечення буде значною. Щойно система показуватиме перші здібності, буде нескладно профінансувати зростання обчислювальної потужності на кілька порядків (залежно від початкових апаратних апетитів). Спеціалізація процесорної схемотехніки додасть ще декілька розрядів потужності. Інші способи розширення апаратної частини, як-от створення нових заводів і технологій, потребуватимуть більше часу — приблизно кілька років, проте щойно за управління виробництвом візьметься штучний суперінтелект, затримка відчутно зменшиться.

Наостанок поговоримо про ймовірність *апаратного переважування*. На момент створення програмного ШІ рівня людини може існувати вдосталь обчислювальних потужностей, щоб забезпечити високошвидкісне виконання багатьох його копій. Програмна консервативність, як ми показали раніше, може бути ще нижчою, але набагато менш передбачуваною. Зокрема, є висока ймовірність виникнення *переважування контенту* — накопичення великого обсягу знань, який стане доступним новоствореній інтелектуальній системі (наприклад, завдяки інтернету). *Алгоритмічне переважування* — досконаліть адаптивних алгоритмів системи — можливе, але його виникнення дещо менш імовірне. Програмні покращення (алгоритму чи контенту) — порівняно легкий спосіб ще збільшити потужність новоствореного штучного розуму.

Оптимізаційна сила та імовірність інтелектуального вибуху

Тепер, після консервативності, розглянемо другий фактор впливу нашої орієнтовної залежності — *оптимізаційну силу*. Нагадаємо:

$$\text{Швидкість зміни інтелектуальності} = \frac{\text{оптимізаційна сила}}{\text{консервативність}}$$

Як видно з рівності, для переходу на новий рівень інтелектуальності, консервативність необов'язково мусить бути низькою. Швидкий стрибок можливий і за сталої консервативності або навіть помірної консервативності — якщо лише оптимізаційна сила, прикладена до системи, також зростатиме достатньо швидко. Є вагомі підстави вважати, що зі зростанням інтелектуальності системи оптимізаційна сила теж *зростатиме*, принаймні якщо цьому не буде зовнішніх перешкод.

Можемо виділити дві фази цього гіпотетичного процесу. Початок першої фази збігається з початком стрибка, коли система досягає людського рівня індивідуального інтелекту. Збільшуючи свої здібності, система спрямовує частину своїх можливостей на самовдосконалення (або створення здібнішого наступника). Проте на цьому етапі більшість прикладеної до системи оптимізаційної сили надходить ззовні: від програмістів, інженерів, залучених у проекті, та інших учасників, залежно від типу проекту[218]. Із плином часу з'являються перші результати, як підтвердження ефективності вибраного підходу — тоді величина оптимізаційної сили може зростати: як зовнішня її фракція, так і внутрішня. Завдяки інтенсивним дослідженням до проекту долучається дедалі більше учасників і ресурсів спрямовується на пришвидшення розроблення. Якщо поява штучного інтелекту рівня людини стане несподіванкою для людства, зростання може виявитися ще більшим. Тоді невеликий дослідний проект опиниться у фокусі уваги та зусиль величезної кількості дослідників (хоч певне пожвавлення охопить усю галузь).

Друга фаза починається з досягнення системою певного рівня розвитку (позначеного як точка «неповернення» на рисунку 7), коли оптимізаційна сила, що зумовлює розвиток системи, більшою мірою зумовлена дією самої системи. З цього моменту динаміка розвитку змінюється, бо будь-яке зростання рівня інтелектуальності системи пропорційно позначається на оптимізаційній силі, що діє на систему. Така система, за умови сталої консервативності, демонструватиме експоненційне зростання рівня інтелекту (див. додаток 4). Період подвоєння в такому процесі залежить від сценарію і може бути критично малий — секунди, — особливо якщо електронний ШІ зростатиме внаслідок алгоритмічних удосконалень, переважування контенту або апаратного забезпечення[219]. Зростання інтелектуальності, зумовлене збільшенням виробництва нових комп'ютерів чи іншого обладнання, буде не таким стрімким (але все-таки може бути значно швидшим за темпи зростання сучасної економіки).

Додаток 4. Детальніше про кінетику інтелектуального вибуху

Виразимо швидкість змін інтелектуальності через відношення оптимізаційної сили, прикладеної до системи ($\mathfrak{O}$), та її

консервативності ($\mathfrak{R}$):

$$\frac{dI}{dt} = \frac{\mathfrak{D}}{\mathfrak{R}}.$$

Загальна оптимізаційна сила, що діє на систему, складається з власної оптимізаційної сили системи та оптимізаційних зусиль іззовні. Наприклад, зерно ШІ може покращуватися завдяки комбінації власних зусиль, дій команди розробників і дій широкого кола науковців та інженерів з усього світу, що працюють над пов'язаними проектами у сферах електроніки та комп'ютерних наук[220]:

$$\mathfrak{D} = \mathfrak{D}_{\text{системи}} + \mathfrak{D}_{\text{проекту}} + \mathfrak{D}_{\text{світу}}.$$

Початкові інтелектуальні здібності зерна ШІ дуже обмежені. Отже, і початкова $\mathfrak{D}_{\text{системи}}$ дуже мала[221]. Якими ж будуть $\mathfrak{D}_{\text{проекту}}$ і $\mathfrak{D}_{\text{світу}}$? Відомі випадки, коли потужність інтелектуальних ресурсів, залучених до одного проекту, перевищувала можливості решти світу. Зокрема Мангеттенський проект зібрав більшість найкращих фізиків світу в Лос-Аламос, щоб працювати над створенням ядерної бомби. Втім зазвичай ресурси окремого наукового проекту значно менші. Заразом усі вони, на відміну від величезних ресурсів зовнішнього світу, сконцентровані на вдосконаленні конкретної системи, тому $\mathfrak{D}_{\text{проекту}}$ може перевищувати $\mathfrak{D}_{\text{світу}}$. При появі перших успіхів — коли інтелектуальність перетне «базовий людський» рівень або й раніше, — проект може почати залучати додаткові інвестиції, і $\mathfrak{D}_{\text{проекту}}$ зростатиме. Якщо перебіг розроблення буде публічним, разом зі зростанням загальної цікавості до ШІ та бажання певних сил приєднатися до потенційних здобутків проекту, зростатиме і $\mathfrak{D}_{\text{світу}}$. Отже, зі здібностями інтелектуальної системи зростатиме й загальна оптимізаційна сила, що діє на неї[222].

Унаслідок зростання в певний момент вплив власної оптимізаційної сили системи $\mathfrak{D}_{\text{системи}}$ перевищить комплексний вплив зовнішніх чинників:

$$\mathfrak{D}_{\text{системи}} > \mathfrak{D}_{\text{проекту}} + \mathfrak{D}_{\text{світу}}.$$

Такий момент *неповернення* важливий, бо після нього будь-яке зростання можливостей системи збільшує оптимізаційну силу, що

діє на систему. У такий спосіб система переходить у режим рекурсивного самовдосконалення. Це зумовлює вибухоподібне зростання інтелекту системи, яке все менше залежить від змін консервативності.

Наприклад, припустимо, що консервативність системи стала і швидкість зростання ШІ залежить виключно від оптимізаційної сили. Якщо вважати, що вся оптимізаційна сила належить лише системі і система спрямовує весь свій інтелект у самовдосконалення, то $\mathfrak{D}_{\text{системи}} = I$.[223] Тоді маємо:

$$\frac{dI}{dt} = \frac{I}{k}.$$

Розв'язавши це просте диференційне рівняння, отримаємо експоненційну функцію:

$$I = Ae^{t/k}.$$

Проте консервативність навряд чи буде сталою. Найімовірніше, вона знизиться під час переходу через базовий людський рівень. А також постійно залишатиметься низькою, до рівня неповернення й далі (допоки, можливо, не почне зростати з наближенням до фундаментальних фізичних обмежень технології). Тому, для прикладу, припустимо, що до моменту досягнення системою здатності самостійно забезпечувати власний розвиток оптимізаційна сила є сталою (тобто, $\mathfrak{D}_{\text{проекту}} + \mathfrak{D}_{\text{світу}} \approx c$) і зумовлює подвоєння здібностей системи кожні вісімнадцять місяців. (Це відповідає історичній тенденції розвитку комп'ютерних технологій та програмного забезпечення, описаній законом Мура[224]). Якщо вважати таку швидкість розвитку результатом дії сталої оптимізаційної сили, то консервативність виявиться оберненою до потужності системи:

$$\frac{dI}{dt} = \frac{c}{1/I} = cI.$$

Якщо консервативність продовжить зменшуватися за гіперболічним законом, то з досягненням системою рівня неповернення величина оптимізаційної сили подвоїться і ми матимемо:

$$\frac{dI}{dt} = \frac{(c+I)}{1/I} = (c+I)I.$$

Наступне подвоєння станеться через 7,5 місяця. А за 17,9 місяця можливості системи збільшаться в тисячу разів, і система набуде властивостей швидкісного суперінтелекту (див. рисунок 9).

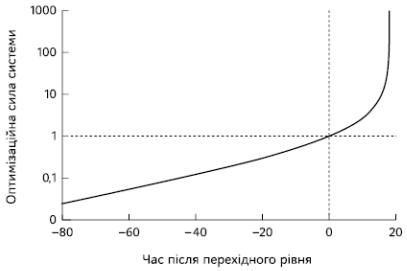


Рисунок 9. **Спрощена модель інтелектуального вибуху**
Розглянутий нами графік зростання має сингулярність: асимптоту $t = 18$ місяців. У реальній ситуації консервативність не буде сталою, коли система наблизиться до фізичних меж росту або навіть раніше. Ці два розглянуті випадки — суто ілюстративні. Можливих варіантів безліч, залежно від форми кривої консервативності. Суть у тому, що зворотний зв'язок між інтелектуальністю й оптимізаційною силою, що починає працювати після проходження рівня неповернення, значно пришвидшує дальше зростання.

Тому, найімовірніше, з розвитком системи оптимізаційна сила, що діє на неї, зростатиме. Спочатку — через те, що розробники, мотивовані успіхом, більше старатимуться, пізніше — завдяки зусиллям самої системи. Тому навіть якщо консервативність буде сталою чи повільно зростатиме від моменту досягнення рівня інтелекту людини, найімовірнішими можна вважати саме швидкий і помірний сценарії розвитку ШІ[225]. Проте, як ми побачили в попередньому підпункті, після досягнення системою базового рівня людини, консервативність може суттєво знизитися. Зокрема, це може трапитися через швидке збільшення потужності апаратного забезпечення; вдосконалення алгоритмів роботи; можливість сканування більшої кількості мізків (у разі емуляції цілого мозку); та можливість швидкого засвоєння великої кількості фактологічної інформації з інтернету (у разі штучного інтелекту)[226].

Незважаючи на це, складно передбачити, як поводитиметься консервативність під час зростання інтелектуальності системи. Зокрема не зрозуміло, чи важко буде покращити якість програмного забезпечення ШІ чи емуляції мозку після досягнення системою рівня людини. Також невідомо, наскільки складно буде розширювати апаратну базу. Зараз будь-який невеликий проект може підняти обчислювальну потужність, просто збільшивши витрати на апаратне

забезпечення, або відкласти покупку на пізніший час, коли ціна впаде. Однак імовірно, що перший штучний інтелект рівня людини з'явиться внаслідок масштабного проекту з використанням дорогого суперкомп'ютера, для якого не існує дешевого способу масштабування. Крім того, невідомо, наскільки актуальним є закон Мура. Тому, незважаючи на більшу ймовірність швидкого та помірного сценаріїв стрибка інтелекту, тривалий сценарій теж не варто виключати[227].

5. ВИРІШАЛЬНА СТРАТЕГІЧНА ПЕРЕВАГА

Окреме питання, дотичне до міркувань про кінетику: чи буде суперінтелект один, чи їх буде багато? Чи буде суперінтелект, що з'явиться внаслідок інтелектуального вибуху, настільки потужним, щоб одноосібно визначати наше майбутнє? Чи, може, прогрес буде синхронним і поступовим, існуватиме широкий спектр проектів створення ШІ, які розвиватимуться одночасно, але жоден з них не отримає тотальної переваги?

У попередньому розділі ми аналізували один із ключових показників, який визначає часовий відрив між найпотужнішим інтелектом і його найближчими конкурентами — швидкість переходу системи від людського рівня інтелектуальності до суперінтелекту. Цей аналіз дає нам змогу зробити перші припущення. Якщо здібності системи змінюватимуться *швидким* стрибком (протягом годин, днів або тижнів), то конкуренція кількох проектів малоймовірна. Перша така система, найімовірніше, завершить перехід до того, як інші розпочнуть його. Якщо ж перехід *повільний* (триватиме роки та десятиліття), то теоретично можуть одночасно існувати багато проектів створення ШІ, які розвиватимуться синхронно, і навряд чи один з них зможе тривалий час суттєво випереджати інші. *Помірний* сценарій переходу характеризується однаковими ймовірностями і паралельного зростання кількох проектів, і абсолютного лідерства одного з них[228].

І нарешті, чи можливо, що один із проектів ШІ у процесі свого розвитку отримає *вирішальну стратегічну перевагу* в технологіях — перевагу, яка дасть змогу панувати над усім світом? Якби якийсь проект набув такої вирішальної переваги, чи використає він її, щоб обмежити конкурентів та встановити *синглтон* (світовий порядок, за якого існує лише одне глобальне джерело авторитету)? І чи великий би був цей проект — не в значенні фізичного розміру, а в кількості людей, чиїми силами той розвиватиметься? Розберемося в усьому по черзі.

Чи матиме перший ШІ вирішальну перевагу?

Швидкість поширення того, що робить лідера інтелектуальних перегонів лідером — це один із чинників, які зумовлюють його відрив від решти.

Першість давалася б значно важче, якби переслідувачі могли легко копіювати ідеї і розробки лідера. Імітація стає тим зустрічним вітром, що нівелює його перевагу і сприяє тим, хто відстає, особливо за відсутності ефективних методів захисту інтелектуальної власності. Крім того, першопроходець традиційно

вразливіший до рейдерства, податкового тиску та антимонопольного законодавства.

Однак було б помилково вважати, що такий зустрічний вітер зростає пропорційно до відриву лідера від переслідувачів. Як останній учасник велоперегонів, що занадто відстав від попередників, втрачає захист від зустрічного вітру, так і в технологічній гонці зі збільшенням відриву стає все важче освоювати нові передові розробки[229]. Занадто велике відставання у знаннях та технологіях. Іноді лідер переносить розробку на нову платформу — тоді технологічний зв'язок між ним і його конкурентами цілковито втрачається і вони більше не можуть орієнтуватися на його рішення. Винахідливий лідер може імітувати витік інформації про власні дослідження та розробки або саботувати спроби конкурентів створити власні технології.

Якщо таким лідером буде система ШІ, вона може мати засоби, які дадуть змогу легше розвиватися і водночас зменшувати швидкість їхнього поширення. У людських організаціях кількісна перевага нівелюється бюрократичним формалізмом та проблемами принципала-агента, зокрема витоками комерційної інформації[230]. Допоки проектом керуватимуть люди, ці фактори можуть стримувати розвиток ШІ. Натомість системам ШІ, завдяки єдності конфігурації системи та її елементів, не властиві проблеми незлагодженості множинної структури. Це дає їм змогу, на відміну від людських організацій, уникати багатьох проблем принципала-агента. Саме завдяки чіткій і злагодженій структурі ШІ зможе тривалий час розвиватися, перебуваючи в тіні. У ШІ немає незадоволених працівників, яких можуть переманити або підкупити конкуренти[231].

Історія технологічних розробок може дати нам певне уявлення про середній часовий інтервал між учасниками подібних перегонів (див. додаток 5). Як ми бачимо, типові затримки між розробленням стратегічних технологій у різних країнах становлять від кількох місяців до кількох років.

Імовірно, що часові затримки між конкурентними розробками і надалі зменшуватимуться внаслідок глобалізації та посилення контролю. А втім, найімовірніше, такі затримки (без зовнішнього втручання) не можуть бути меншими за певне мінімальне значення[232]. Навіть відсутність результатів може спричинити ефект снігової кулі, внаслідок чого деякі проекти зрештою матимуть кращу команду дослідників, керівництво, інфраструктуру або просто натраплять на кращу ідею. Якщо два проекти, що конкурують, використовують два альтернативні підходи до розв'язання однієї проблеми і перший підхід виявиться кращим, менш результативному проекту можуть знадобитися місяці, щоб визнати помилку й застосувати підхід конкурента. Навіть за умови детальної обізнаності із ходом його досліджень.

Додаток 5. Перегони технологій: історичний екскурс

Якщо розглядати історію в довгостроковій перспективі, швидкість поширення світом нових технологій та знань зростає. Через це поступово зменшуються часові відстані між учасниками технологічних перегонів.

Протягом двох тисячоліть Китай утримував монополію на виробництво шовку. За археологічними даними, це виробництво розпочалося щонайменше за 3000 років до н. е.[233] Секрети шовківництва суворо охороняли. Винних у передаванні технологій карали на смерть, як і тих, хто намагався вивезти з Китаю яйця або личинки шовкопряда. Римляни, незважаючи на високу ціну, що запропонували за шовк, так і не змогли вивідати та освоїти секрети його виробництва. Лише приблизно у 300 році н. е. японцям вдалося вивезти личинки та захопити кількох китайських дівчат, які під тортурами змушені були відкрити секрет своїм викрадачам[234]. Потім, у 522 році н. е., виготовляти шовк розпочали у Візантії. Поширення технології виробництва порцеляни теж відбувалося повільно. У Китаї розквіт виробництва припав на період правління династії Тан, у 600 році н. е. (а розпочалося, імовірно, у 200 році н. е.). Європа ж освоїла технологію лише у XVIII столітті[235]. Колеса почали використовувати в різних місцевостях Європи та Межиріччя близько 3500 року до н. е., а в Америку технологія дісталася лише після «відкриття» Колумбом[236]. Якщо ж поглянути в більшому масштабі, то людству знадобилося десятки тисяч років, щоб заселити всю Землю; аграрна революція тривала тисячі років; промислова революція — лише сотні, а інформаційна революція охопила світ протягом якихось десятиліть. Проте глибина цих зрушень дещо різна. (Відеогра Dance Dance Revolution поширилася з Японії до Європи та Північної Америки лише за один рік!)

Останнім часом у контексті патентної гонки та гонки озброєнь вдалося досить ґрунтовно вивчити феномен технологічного суперництва[237]. Огляд цієї літератури виходить за межі книжки, а втім, досить корисно буде поглянути на часові межі поширення деяких стратегічно важливих технологій XX століття (див. таблицю 7).

У випадку шести важливих технологій і країн, що змагалися за першість у їх створенні, часовий відрив між лідером та найближчим суперником становив відповідно 49, 36, 4, 1, 4 та 60 місяців — довше за швидкий і швидше, ніж повільний сценарій досягнення суперінтелектуальності[238]. Часто конкуренти користувалися даними розвідки і публічно доступною інформацією. Проста демонстрація певної технології може заохотити конкурента до її самостійного розроблення, а бажання зберегти перевагу — примусить активізувати зусилля.

Розробку програмного ШІ можна порівняти з математичними відкриттями: вони не потребують розвинутої фізичної інфраструктури. Результати таких

досліджень зазвичай публікують у науковій літературі, тому є загальнодоступними. Але якщо дослідження мають стратегічну важливість, їх публікація часто відбувається із затримкою. Наприклад, дві найбільш важливі ідеї асиметричного шифрування з відкритим ключем — алгоритм RSA та протокол обміну ключами Діффі—Геллмана — стали відомими науковій спільноті у 1976 та 1978 роках відповідно. Хоч, як виявилося згодом, британські спеціалісти із шифрування використовували їх ще на початку 1970-х років[239]. Приклади створення великих програмних продуктів стали б кращою аналогією ШІ, але програми зазвичай випускаються поступово, модулями, нарощуючи функціонал, який до того ж сильно відрізняється. Тож конкурентні системи переважно неможливо адекватно порівняти, як за можливостями, так і за часом появи.

Таблиця 7. **Деякі випадки суперництва за стратегічно важливі технології**

	США	СРСР	Велика Британія	Франція	Китай	Індія	Ізраїль	Пакистан	Північна Корея	Південна Африка
Ядерна бомба	1945	1949	1952	1960	1964	1974	1979?	1998	2006	1979?
Термоядерна бомба	1952	1953[240]	1957	1968	1967	1998	?	—	—	—
Здатність до виведення на орбіту супутника	1958	1957	1971	1965	1970	1980	1988	—	1998?[241]	—[242]
Запуск у космос людини	1961	1961	—	—	2003	—	—	—	—	—
МБР[243]	1959	1960	1968[244]	1985	1971	2012	2008	—[245]	2006	—[246]
РГЧІН[247]	1970	1975	1979	1985	2007	2014[248]	2008?			

У контексті попереднього обговорення сценаріїв переходу до суперінтелектуальності ці спостереження дають нам право стверджувати, що у швидкому сценарії два проекти побудови штучного інтелекту навряд чи зможуть бути настільки близькими, щоб скласти один одному конкуренцію. У помірному сценарії — конкуренція можлива. Під час повільного сценарію — імовірність конкуренції найвища. Тепер зробимо наступний крок нашого аналізу. Цікавить не те, скільки проектів створення суперінтелекту існуватиме одночасно, а скільки з них може досягти успіху в переході до суперінтелектуальності разом: без вирішальної стратегічної переваги жодного з них. Якщо швидкість переходу конкурентних проектів до

суперінтелектуальності зростатиме, відрив між ними збільшуватиметься. Згадуючи нашу метафору велоперегонів, ситуація схожа на рух двох велосипедистів угору по схилу, коли один із них, дістався до вершини, починає збільшувати відрив.

Уявімо собі такий сценарій помірно швидкого переходу. Нехай певному проекту потрібен рік, щоб розвинути можливості власної моделі ШІ від базового людського рівня до сильної суперінтелектуальності, і він розпочинає на шість місяців раніше іншого подібного проекту. Ці два проекти конкуруватимуть у досягненні своєї мети. Може здаватися, що жоден із них не матиме вирішальної переваги. Але це необов'язково так. Нехай перша система дійде до точки неповернення через дев'ять місяців, а за наступні три — досягне сильної суперінтелектуальності. У такий спосіб перший проект досягне суперінтелектуальності на три місяці раніше, ніж друга система пройде точку неповернення. Це дасть першій системі вирішальну стратегічну перевагу, адже вона зможе перетворити свою першість на тотальне домінування, зупинивши конкурентні проекти та встановивши синґлтон. (Зверніть увагу, що світовий порядок синґлтону абстрактний. Він не передбачає конкретної форми правління: це може бути демократія, тиранія, один верховний ШІ чи вичерпний перелік глобальних законів із вбудованими механізмами їх забезпечення. Або ж навіть якийсь нелюдський правитель, основні функції якого — просто бути агентом, медіумом волі джерела авторитету й розв'язувати всі організаційні проблеми. Крім того, форма правління може виявитися назагал несхожою на будь-яку з наразі відомих нам[249]).

Після проходження точки неповернення ймовірність вибухового зростання інтелектуальності системи найвища — саме тоді починає проявлятися вплив зворотного зв'язку на оптимізаційну силу. Цей процес потребує пильної уваги, адже імовірність отримання такою системою вирішальної стратегічної переваги дуже висока, навіть якщо зростання загалом не відбуватиметься за швидким сценарієм.

Наскільки масштабним буде успішний проект суперінтелекту?

Деякі шляхи побудови суперінтелекту вимагають значних ресурсів, тому можуть бути реалізовані лише під час масштабних, щедро профінансованих проектів. Прикладом такого проекту може бути емуляція цілого мозку, яка потребує залучення широкого кола фахівців і використання великої кількості обладнання. Покращення біологічного розуму та створення нейроінтерфейсів — теж досить масштабні завдання. Невелика фармкомпанія може розробити кілька препаратів, але для досягнення суперінтелектуальності (якщо це взагалі можливо) знадобиться багато винаходів, випробувань та, як наслідок, підтримка промислового сектору або окрема бюджетна програма національного рівня із

щедрим фінансуванням. Створення суперінтелекту за допомогою удосконалення організаційних та мережевих структур потребує ще більших вкладень із залученням значної частини світової економіки.

Визначити, що потрібно для створення суперінтелектуального ШІ — складніше. Можливо, це під силу лише великому проекту, а може — невеликій компанії. Також не варто відкидати і сценарій самотнього хакера. Можливо, побудова зерна ШІ потребуватиме від світової наукової спільноти осяянь та десятиліть праці над алгоритмами, але лише одна людина чи невелика група зможе запропонувати останню ключову ідею й зібрати все докупи у працездатну систему. Деякі види ШІ через особливості структури не підходять для такого сценарію. Створити систему, що складається з багатьох елементів, які треба окремо налаштовувати для спільної роботи та ініціалізувати попередніми наборами даних, найімовірніше, під силу лише великому проекту. Але якщо зерно ШІ виявиться простою системою, яка для початку роботи потребуватиме правильної реалізації обмеженої кількості принципів, то його зможе створити невелика команда. Імовірність того, що невеликий проект може виявитися успішним більша, якщо результати глобальної роботи в цьому напрямі будуть широко доступні усім охочим через наукову літературу та відкрите програмне забезпечення.

Важливо розуміти різницю між тим, скільки людей бере безпосередню участь у створенні інтелектуальної системи і тим, скільки людей ухвалює рішення про початок, етапи та часові межі такого створення. Ядерну бомбу створила група науковців та інженерів. (До Мангеттенського проекту було залучено до 130 000 людей, більшість з яких була будівельниками та операторами[250]). Однак усі ці фахівці працювали під керівництвом військових, які тим часом підпорядковувалися американському уряду. А той був підзвітний виборцям, що тоді становили десяту частину дорослого населення світу[251].

Моніторинг

Враховуючи значні ризики, пов’язані з суперінтелектом, уряди, що діють у межах певних територій, намагатимуться націоналізувати кожен доступний їм проект, який має шанси на створення такої системи. Потужніші країни, імовірно, спробують отримати контроль над такими проектами інших держав — за допомогою шпигунства, викрадення технологій і людей, підкупу, погроз, військового захоплення тощо. За неможливості захоплення вони, найімовірніше, спробують зруйнувати перспективи успіху цих проектів, особливо якщо не матимуть ефективної протидії. Якщо до початку активної роботи над такою системою з’являться сильні структури глобального урядування, проект, імовірно, опиниться під міжнародним контролем.

Тому важливо: чи зможуть національні або міжнародний уряди побачити наближення вибуху інтелекту? Зараз, здається, розвідки не надто цікавляться

проектами ШІ або іншими формами інтелектуальних систем із перспективами вибухового зростання[252]. Якщо це справді так, то причиною цього, вочевидь, є поширене уявлення про відсутність будь-яких ознак існування потенційно суперінтелектуальних систем. Але щойно відомі науковці почнуть висловлювати впевненість у тому, що створити суперінтелект цілком можливо, дослідники та організації, залучені у відповідних проектах одразу опиняться у фокусі уваги провідних розвідок світу. І кожен проект, який демонструватиме стрімке зростання, може стати кандидатом на націоналізацію. Якщо політичні кола будуть цілком впевнені в ризиках суперінтелекту, будь-яка пов'язана з ним діяльність опиниться під жорстким контролем або навіть забороною.

Яких зусиль потребуватиме такий моніторинг? Завдання буде простішим, якщо контролювати лише найуспішніші проекти. У такому разі достатньо буде відстежувати лише найзабезпеченіші проекти. Якщо ж виникне необхідність повністю припинити будь-які розроблення (крім тих, що провадять уповноважені інституції), потрібен буде всеохопніший нагляд. Оскільки навіть окремі дослідники та невеликі проекти можуть досягти принаймні часткового успіху у створенні ШІ.

Великі проекти, як-от емуляція цілого мозку, потребуватимуть значних капіталовкладень, тому відстежувати їхній перебіг буде легше. Для створення штучного інтелекту натомість потрібний лише комп'ютер, тому стежити за його перебігом буде важче. Для теоретичних розробок потрібна лише ручка та папір. Але визначити найздібніших і найбільш перспективних дослідників у сфері ШІ, які давно та серйозно цікавляться цим напрямом науки, буде неважко. Такі люди зазвичай добре помітні. Вони мають наукові публікації, беруть участь у конференціях, дописують на форумах та отримують наукові звання у відомих наукових установах. А також спілкуються з іншими науковцями, тому їх можна відстежувати й за соціальними зв'язками.

Натомість, напочатку таємні проекти відстежити буде значно важче. Вони можуть ховатися за ширмою розробки звичайного програмного забезпечення[253]. Тільки уважний аналіз коду може показати справжню мету розробки. Такий аналіз потребуватиме значної (фахової) роботи, тому контролювати на такому рівні можна буде лише невелику кількість проектів. Якщо ж вдасться створити і запровадити систематичне застосування засобів детектування брехні, можливості моніторингу проектів розширяться[254].

Вчасно розпізнати передумови появи суперінтелекту може бути складно ще й тому, що деякі наукові відкриття неможливо передбачити. Це більше стосується ШІ і менше — емуляції цілого мозку: адже, як ми пам'ятаємо, успіх емуляції має забезпечити чітка послідовність технологічних розробок.

Також можливо, що причиною відсутності в державних установ та розвідок бачення розуміння важливості певних подій і розробок є їхня інтелектуальна

незграбність й консервативність. Найскладніше може виявитися офіційно визнати ймовірність вибуху інтелектуальності. До того ж варто очікувати значних політичних та релігійних суперечок навколо цього питання — аж до офіційної заборони в деяких країнах. Не бажаючи, щоб їх асоціювали з шарлатанством та особами, які дискредитували себе, поважні науковці можуть відкинути ШІ. (Дещо подібне, як ми бачили в розділі 1, уже трапилося двічі: пригадайте дві «зими ШІ»). Щоб захистити прибуткові бізнеси від ганьби, промислові групи лобіюватимуть потрібні рішення. Можливо, доведеться позбавити кількох науковців їхніх наукових звань, щоб поодинокі голоси стурбованих долею проведеної роботи звучали якомога маргінальніше[255].

Отже, не варто відкидати ймовірності цілковитої поразки людського інтелекту. Якби невдовзі суперінтелект став реальністю, громадська думка виявилася б не готовою до цього і ймовірність поразки була б особливо відчутною. Проте навіть якщо розвідка встигне підготувати контрзаходи, вони можуть зіткнутись із нерозумінням політиків. Щоб розпочати Мангеттенський проект, кілька видатних фізиків і візіонерів, зокрема Марк Оліфант та Лео Сілард, мусили докласти неабияких зусиль, аби переконати Юджина Вігнера вплинути на Альберта Ейнштейна, щоб той підписав прохання до президента Франкліна Д. Рузвельта підтримати проект. Навіть коли проект запрацював на повну силу, Рузвельт скептично ставився до його перспектив та важливості — як, зрештою, і його наступник Гаррі Трумен.

Так чи інакше, невеликим групам зацікавлених осіб буде важче вплинути на результат інтелектуального вибуху, якщо великі гравці, як-от уряди країн, братимуть активну участь у його долі. Тому, що менше уваги до таких проектів від великих гравців, то більше можливостей буде у приватних осіб та активістів вплинути на екзистенційні ризики від вибуху інтелекту. А також визначити коли, з якою метою та хто саме з великих гравців долучиться до процесу. Амбітні активісти розраховують саме на такий сценарій, хоч самі, можливо, вважають сценарій тотального урядового контролю ймовірнішим.

Міжнародна співпраця

У разі існування потужних глобальних структур урядування, координація створення суперінтелекту може здійснюватися на міжнародному рівні. Окрім того, запровадження такої координації можливе, якщо важливість інтелектуального вибуху буде наперед усвідомлена і визнана на офіційному рівні, а також здійснюватиметься ефективний моніторинг усіх перспективних проектів. Проте міжнародна взаємодія можлива і без моніторингу. Багато країн можуть об'єднати зусилля та започаткувати спільний проект. За умови доброго фінансування такий проект має всі шанси на успіх, особливо якщо його конкуренти опиняться поза законом і будуть змушені приховувати масштаби своєї діяльності, щоб уникнути викриття.

Історії відомі випадки міжнародної наукової співпраці: Міжнародна космічна станція, Проект генома людини та Великий адронний колайдер[256]. Однак основною метою співпраці в цих проектах був розподіл фінансування. (У випадку МКС важливою метою було також започаткувати співпрацю між Росією та США[257]). Організувати співпрацю у проекті, який може настільки сильно вплинути на світову безпеку, буде значно важче. Якщо якась країна вважатиме, що має всі шанси досягнути мети самотужки, вона може відмовитися від співпраці. Учасник може покинути спільний проект через побоювання, що інші використають технології, створені таким проектом, для пришвидшення власних таємних розробок.

Тому для такого міжнародного проекту, крім розв'язання багатьох проблем безпеки, потрібен ще й значний ресурс взаємодовіри — актив, для формування якого потрібен час. Так, навіть після приходу до влади у СРСР Горбачова та потепління у відносинах зі США взаємне скорочення озброєнь розпочиналося дуже невпевнено, незважаючи на потенційну користь від цього процесу для обох наддержав. Горбачов наполягав на різкому скороченні ядерного озброєння, але переговори зайшли в глухий кут через незгоду Кремля з деякими аспектами Стратегічної оборонної ініціативи Рейгана (також відому як Star Wars — «зоряні війни»).

У 1968 році на саміті в Рейк'явіку Рейган запропонував надати СРСР доступ до технологій, що мали бути розроблені в межах Стратегічної оборонної ініціативи та давали змогу захиститися від випадкових запусків і погроз менших країн у разі отримання ними ядерної зброї. Але Горбачов не погодився на цю, здавалося б, взаємовигідну пропозицію. Він вважав це підступом і не вірив, що Америка готова надати доступ до своєї найновішої військової розробки, водночас відмовляючись поділитися технологією доїння корів[258]. Невідомо, чи була пропозиція Рейгана щодо співпраці наддержав щирою — усе одно через недовіру вона не була реалізована.

Союзникам співпрацювати простіше, проте так буває не завжди. Протягом Другої світової війни Радянський Союз та США були союзниками, але Сполучені Штати приховали роботу над створенням ядерної бомби. Водночас США залучили до Мангеттенського проекту Великобританію та Канаду[259]. Так само Великобританія приховала від СРСР злам коду Енігми, але поділилася цим — вельми неохоче — із США[260]. Отже, для міжнародної співпраці у створенні надзвичайно важливої з погляду безпеки технології потрібні близькі та довірливі відносини, які треба будувати заздалегідь.

До важливості та бажаності міжнародної співпраці в питанні створення технологій покращення інтелекту ми повернемося в розділі 14.

Від вирішальної стратегічної переваги до синґлтону

Чи використає суперінтелект вирішальну стратегічну перевагу, щоб сформувати синґлтон?

Пригадаємо дещо подібну історичну ситуацію. Першу ядерну бомбу створили Сполучені Штати у 1945 році. Вони були єдиною ядерною країною до 1949 року, коли ядерну зброю створив Радянський Союз. Протягом цих кількох років Сполучені штати мали вирішальну військову перевагу — або принаймні могли швидко її здобути.

Тоді США теоретично могли використати свою ядерну монополію, щоб утворити синґлтон. Для цього треба було б зважитися на те, щоб швидко наростити свою військову міць, і погрозами ядерних ударів (або за потреби їх здійсненням) зупинити і знешкодити промислові потужності, потрібні для формування еквівалентного потенціалу у СРСР, а також будь-яких інших країнах, які б наважилися на будівництво ядерних об'єктів.

Сприятливішим і водночас реалістичнішим варіантом було б використати ядерний арсенал як розмінну монету в переговорах щодо створення сильного міжнародного уряду — подібного до Організації Об'єднаних Націй, але без можливості ветування рішень. Такий уряд мав би монополію на ядерну зброю та карт-бланш на дії у відповідь будь-якій країні, що наважилася б самостійно будувати свій ядерний потенціал.

Обидва ці варіанти свого часу пропонувалися. За жорсткий варіант ударів та погроз виступали деякі знані інтелектуали, зокрема Бертран Расселл (відомий своєю участю в антивоєнних рухах та пізніше у кампаніях протидії ядерній зброї) та Джон фон Нейман (співавтор теорії ігор та один з архітекторів ядерної стратегії США)[261]. Сьогодні внаслідок (хочеться вірити) цивілізаційного прогресу сама ідея погроз превентивним ядерним ударом здається вкрай нерозумною і аморальною.

Спроба піти іншим, сприятливішим шляхом була здійснена у 1946 році у США і відома як план Баруха. Цей план передбачав відмову Сполучених Штатів від ядерної монополії. Контроль над ядерною технологією та добуванням урану й торія мала отримати міжнародна організація, що створювалася під егідою ООН. Постійні члени Ради Безпеки ООН позбавлялися права вето щодо рішень, які стосуються ядерної зброї, щоб запобігти блокуванню рішень більшості[262]. Сталін не підтримав цю пропозицію, щоб перешкодити більшості в Раді Безпеки та Генеральній Асамблеї ООН отримати перевагу над СРСР і його союзниками. У стосунках нещодавніх військових союзників запанувала прохолода взаємних підозр — недовіра, що згодом переросла в «холодну війну». Як багато хто передбачав, розпочалася витратна та надзвичайно небезпечна гонка ядерних озброєнь.

Якщо говорити про організацію з вирішальною стратегічною перевагою, то утримати її від створення синґлтону здатна низка факторів. Серед них можуть

бути: неможливість узагальнення функції оцінки корисності або її обмеження, правила ухвалення рішень, що не дають максимізувати виграш, плутанина та невизначеність, проблеми узгодження дій та інші ускладнення, пов'язані з перехідними процесами зростання інтелектуальності. Але що, як вирішальну стратегічну перевагу матиме штучний суперінтелект, а не людська організація, якими будуть його дії? Чи стануть йому на заваді згадані фактори? Розгляньмо окремо кожен із них і спробуємо змоделювати їхній вплив.

Людську оцінку ресурсів, з якими зазвичай мають справу окремі люди та організації, зазвичай неможливо сповна описати «необмеженою агрегованою функцією корисності». Наприклад, людина не ризикуватиме всім заради 50 відсотків імовірності подвоїти ставку. Держава теж не ризикуватиме всією територією заради десяти відсотків імовірності десятикратного розширення володінь. Для людей і держав зростання більшості ресурсів підлягає закону спадної віддачі. ШІ може *не* поділяти такої оцінки. (Повернемося до питання мотивації ШІ згодом). Тому він може вибрати більш ризиковану стратегію дій, результатом яких буде контроль над світом.

У процесі ухвалення рішень люди та колективи можуть назагал не прагнути максимізації очікуваної корисності. Натомість їм достатньо буде мінімізувати ризики, обмежитися «задовільним» результатом на певному рівні адекватності, або, як варіант, погодитися на «деонтологічні» критерії діяльності незалежно від сприятливості їхніх наслідків. Дії людей, що ухвалюють рішення часто здебільшого зумовлені ідентичністю особи або її соціальною роллю, аніж доцільністю для досягнення певної мети. Зі штучним агентом усе може бути по-іншому.

Обмеження функції корисності, мінімізація ризиків і критерії ухвалення рішень, що не максимізують виграш, синергічно поєднуються з плутаниною в розумінні стратегії діяльності та невизначеністю. Навіть успішні революції часто-густо не досягають обіцяних їхніми лідерами результатів. Тому перед тим, як вчинити дію, яку неможливо скасувати і яка порушує норми або є безпрецедентною, людина схильна до нерішучості. Натомість суперінтелект, який ухвалив рішення скористатися вирішальною перевагою для закріплення своєї першості, може сприймати ситуацію ясніше і чітко уявляти собі стратегію дій та їхній результат.

Ще одна вада, яка перешкоджає групам скористатися потенціальною вирішальною стратегічною перевагою — це проблема внутрішньої координації дій. Члени таємного угруповання, яке бажає захопити владу, мають побоюватися не тільки викриття організації ззовні, а також діяльності активнішої меншості всередині групи, спрямованої на узурпацію впливу. Якщо у групі зі ста людей шістдесят можуть захопити владу й позбавити впливу наївну меншість, чи складно тоді тридцяти п'яти найактивнішим особам із шістдесяти усунути від

влади інших двадцять п'ять? А далі, можливо, двадцять усунуть інших п'ятнадцять? А отже, щоб запобігти розпаду соціальних структур та хаосу, внаслідок спроб відвертого силового захоплення влади, кожен зі ста осіб має всі причини дотримуватися певних прийнятих норм. Така проблема внутрішньої узгодженості дій не властива системі ШІ, яка має цілісну структуру[263].

І насамкінець розглянемо аспект витрат. Навіть якби Сполучені Штати могли використати свою ядерну монополію для досягнення синґлтону, це могло б призвезти до значних витрат. У разі передавання ядерного потенціалу під контроль реформованої та підсиленої ООН витрати могли виявитися порівняно невеликими. Проте в разі прямого військового захоплення світу із застосуванням ядерної зброї витрати — моральні, економічні, політичні та людські — були б неймовірно великими, незважаючи на ядерну монополію. Щоправда, за умови досягнення значної технологічної переваги, витрати могли б бути меншими. Уявіть собі випадок, коли одна країна завдяки своїй технологічній перевазі може знешкодити зброю всіх інших країн натиском єдиної кнопки без шкоди життю людей, інфраструктурі чи середовищу. Така технологічна перевага значно додала б рішучості цій країні. Або можна уявити собі ще більшу технологічну перевагу лідера, що змусить країни добровільно відмовитися від збройного протистояння — не погрозами, а впливом на населення засобами надзвичайно ефективної пропаганди, яка переконувала б у загальній вигоді світової єдності. І, напевно, використання вирішальної стратегічної переваги в такий спосіб та створення синґлтону не зустріло б серйозних моральних заперечень, якби здійснювалося із благородною метою припинення національного суперництва і гонки озброєнь, а також створення справедливого, репрезентативного й ефективного світового уряду.

Отже, міркування підштовхують нас до висновку, що суперінтелектуальна сила майбутнього, здобувши достатньо значну стратегічну перевагу, щоб сформувати синґлтон, найімовірніше, нею скористається. Добре це чи погано — залежить від природи новоствореного синґлтону та від того, яким буде розумне життя в багатополярних світах альтернативних сценаріїв розвитку подій. Ми розглянемо ці питання в наступних розділах. Але спершу спробуємо визначити причини та природу можливості суперінтелекту спричиняти події у світі.

6. РОЗУМОВІ СУПЕРЗДІБНОСТІ

Припустимо, що цифровий суперінтелект нарешті прийшов у цей світ і чомусь забажав отримати над ним владу. Чи зможе він це зробити? У цьому розділі ми розглянемо деякі здібності, які б суперінтелект міг у собі розвинути, і як би він міг їх застосувати. Ми окреслимо можливий шлях, який довелося б подолати певному суперінтелекту в напрямку до захоплення світу, від звичайного програмного забезпечення до встановлення синґлтону. Насамкінець поміркуємо про зв'язок між владою над природою та владою над іншими агентами.

Основною причиною панівного становища людини на Землі є можливості нашого мозку: більші, ніж в інших формах життя[264]. Переваги нашого розуму дають змогу ефективніше поширювати культуру. Так ми накопичуємо знання і технології та передаємо їх від одного покоління до наступного. Завдяки цьому ми можемо літати в космос, винайшли водневу бомбу, генну інженерію, комп'ютери, фермерство, інсектициди, міжнародний рух за мир та інші атрибути сучасної цивілізації. Геологи назвали поточну еру Антропоценом через виразні біологічні, осадові та геохімічні сліди людської діяльності[265]. За однією з оцінок, ми відповідальні за 24 відсотки від загальної кількості продуктів життєдіяльності планетарної екосистеми[266]. І ми все ще дуже далеко від фізичних меж розвитку технологій.

Такі спостереження наштовхують на думку, що будь-яка розумна сутність із рівнем інтелекту, що значно перевищує людський, матиме надзвичайну могутність. Її інтелект зможе накопичувати знання та продукувати нові технології значно швидше за нас. Окрім того, завдяки досконалішому інтелекту вона зможе будувати ефективні стратегії.

Спробуємо спрогнозувати, які можливості матиме суперінтелект і як він зможе їх використовувати.

Функціонал і суперздібності

Важливо намагатися уникати антропоморфізму у своїх прогнозах і судженнях про можливості суперінтелекту. Обмежене мислення може стати причиною необґрунтованих очікувань щодо траєкторії зростання зерна ШІ, психології, мотивації та можливостей зрілого суперінтелекту.

Наприклад, досить поширене уявлення, що суперінтелектуальна машина буде схожою на дуже розумну, але дещо занудну людину. Що вона буде ерудована, але їй бракуватиме соціальних навичок, вона матиме залізну логіку, але її інтуїція та творчі здібності будуть слабкі. Причиною такого враження може бути аналогія з комп'ютерами — вони добре рахують, запам'ятовують дані, неухильно виконують програму, але не чутливі до контекстів, підтекстів, емоцій живого мовлення, соціальних норм та політики. Окрім того, люди, які добре тямлять у комп'ютерах, зазвичай видаються нам нудними, і це теж мимоволі впливає на асоціативний зв'язок. Логічно припускати, що сильний комп'ютерний інтелект матиме схожі, але пропорційно посилені характеристики.

На ранніх етапах розвитку зерна ШІ такий евристичний прогноз може справджуватися. (Проте немає підстав стверджувати те саме про емуляції та розумове покращення людини). На початку розвитку майбутньому штучному суперінтелекту, ймовірно, бракуватиме багатьох природних для людини умінь і талантів. За складом натури таке зерно ШІ справді може дещо нагадувати такого собі «ботана». Найважливішою властивістю зерна ШІ, окрім того, що його легко покращити (низька консервативість), має бути здібність спрямовувати оптимізаційну силу на вдосконалення власного розуму. Така характеристика, імовірно, пов'язана з добрими знаннями в галузях математики, програмування, електроніки, комп'ютерних наук та інших «ботанських справ». Проте такі дещо «ботанські» схильності на ранніх етапах розвитку зерна ШІ не означають, що зрілий суперінтелект теж мусить обов'язково бути таким обмеженим. Згадаймо наші міркування про прямий та непрямий шлях до суперінтелекту. З можливістю покращувати власний інтелект, можна — непрямим шляхом — досягти інших розумових здібностей. Тобто за потреби створити нові розумові

модулі і властивості — емпатію, політичну проникливість та інші риси, не характерні для стереотипної комп'ютерної особистості.

Навіть визнавши, що суперінтелект може мати будь-які здібності та вміння, властиві людині або ні, схильність до антропоморфізму може змусити нас недооцінювати те, наскільки можливості суперінтелекту можуть перевищити наші. Як ми бачили в одному з попередніх розділів, Елізер Юдковський красномовно підкреслює хибність такого сприйняття: наші уявлення про «розумність» і «дурість» походять із досвіду меж людського інтелекту і є надто вузькі, порівнюючи з різницею між будь-якою людиною та суперінтелектом[267].

У розділі 3 ми описували потенційні джерела переваг для штучного інтелекту. Потенціал переваг настільки значний, що відмінність між суперінтелектом і людиною варто уявляти собі не як різницю між геніальним науковцем і звичайною людиною, а як між звичайною людиною і комахою або червом.

Добре було б, якби можна було квантувати розумові здібності певної інтелектуальної системи, послуговуючись зрозумілими величинами, як-от бали IQ або рейтинг Ело, яким оцінюють здібності гравців у парних іграх на кшталт шахів. Але вони не придатні для оцінки надлюдських інтелектуальних здібностей. Адже нас не цікавить імовірність перемоги суперінтелекту в шахах. Бали IQ також не інформативні, поки їх не можна співвіднести з результатами в якій-небудь практичній інтелектуальній діяльності[268]. Наприклад, нам відомо, що особа з показником IQ 130 балів переважно краще вчитиметься у школі й виконуватиме складнішу інтелектуальну роботу, ніж, скажімо, з IQ 90 балів. Тепер припустимо, що ми якось установили, що ШІ майбутнього матиме показник IQ 6455 балів. Що це означає? Це нам нічого не повідомить про його конкретні здібності. Ми навіть не зможемо впевнено стверджувати, що такий ШІ має загальний інтелект на рівні звичайної дорослої людини — може, усе, що він має, це лиш набір спеціальних алгоритмів, завдяки яким він із надлюдською ефективністю розв'язує звичайний набір тестів IQ.

Нещодавно були спроби розробити систему вимірювання розумових здібностей, придатну до широкого спектра інтелектуальних можливостей, зокрема ШІ[269]. Якщо вдасться подолати технічні перешкоди, вона може бути корисною для наукових досліджень, як і

для розроблення ШІ. Однак для нашої мети її цінність також обмежена, адже ми не зможемо пов'язати значення оцінки надлюдських здібностей певної системи з рівнем її можливостей у конкретній, корисній для нас діяльності.

Тому нам краще визначити деякі стратегічно важливі завдання і спробувати описати гіпотетичний суперінтелект у термінах придатності до виконання цих завдань та володіння потрібними для них якостями. Розглянемо таблицю 8. Вважатимемо, що система, яка неперевершено виконуватиме будь-яке з наведених у таблиці завдань, матиме відповідну *суперздібність*.

Таблиця 8. **Суперздібності: деякі стратегічно важливі завдання і потрібні для них уміння**

Завдання	**Здібності**	**Важливість**
Посилення інтелекту	Програмування ШІ, дослідження розумового розвитку, розвиток соціальної епістемології тощо	• Система здатна сама розвивати власний інтелект
Побудова стратегій	Стратегічне планування, прогнозування, визначення пріоритетів, аналіз та оптимізація можливостей досягнення довгострокових цілей	• Досягнення довгострокових цілей • Подолання розумного супротивника
Психологічна маніпуляція	Соціальне та психологічне моделювання, маніпулювання, переконлива риторика	• Управління зовнішніми ресурсами через залучення людських виконавців • Обхід обмежень за допомогою сили переконання • Схиляння урядів та організацій до певних дій
Хакерство	Пошук та використання вразливостей у комп'ютерних системах	• Захоплення комп'ютерних ресурсів через інтернет • Звільнення з контейнера • Крадіжка фінансових ресурсів • Захоплення контролю над інфраструктурою, військовими роботами тощо
Створення	Створення та моделювання складних	• Створення потужної

технологій	технологій (наприклад, біотехнологій, нанотехнологій) і розробок	військової сили • Створення систем нагляду • Автоматизована колонізація космосу
Економічна успішність	Різноманітні здібності, потрібні для економічно успішної розумової діяльності	• Створення фінансових ресурсів для купівлі впливу, послуг та засобів (зокрема апаратних) тощо

Справжній суперінтелект добре опанує ці вміння і тому володітиме усіма шістьма суперздібностями. Невідомо, чи зможе існувати обмежений суперінтелект, який матиме лише деякі з них. Створення машини, наділеної навіть одним із цих умінь, здається завданням одного рівня складності зі створенням повноцінного ШІ. Заразом колективний суперінтелект, що складається з достатньо великої кількості людиноподібних біологічних або електронних розумів, може мати, скажімо, економічну суперздібність, але не мати суперздібності до побудови стратегій. Спеціалізований інженерний ШІ може мати технологічну суперздібність і не мати інших. Імовірність була б більшою, якби для того, щоб досконало оволодіти деяким важливим технологічним умінням, не потрібен був надпотужний загальний інтелект. Наприклад, певний різновид спеціалізованого ШІ міг би за допомогою моделювання молекулярних систем створювати різноманітні наномолекулярні вироби (наприклад, комп'ютери або зброю з новітніми характеристиками) на основі достатньо абстрактного опису користувачем[270]. Такий ШІ міг би також описати детальний план розвитку наявної технології (наприклад, біотехнології та технології синтезу білків) до потужності, потрібної для дешевого масового виробництва з атомарною точністю значно ширшої номенклатури наномеханічних структур і речовин[271]. Проте може виявитися, що інженерний ШІ не зможе повністю опанувати суперздібність до створення технологій без передових можливостей в інших сферах. Наприклад, без здібності розуміти запити користувача, детально моделювати роботу виробів у реальних умовах, реагувати на непередбачувані випадки та помилки, забезпечувати постачання потрібних ресурсів тощо[272].

Система, яка має суперздібність до покращення власного інтелекту, може в процесі розвитку набувати інших не доступних спочатку суперздібностей. Однак перетворитися на повноцінний суперінтелект їй вийде не лише за допомогою цієї суперсили. Завдяки здібності до побудови стратегій система може розробити план, унаслідок якого її інтелектуальність зростатиме (наприклад, створивши сприятливі умови для того, щоб людські програмісти та інженери взялися за вдосконалення її інтелектуальних здібностей).

Сценарій захоплення світу штучним інтелектом

Отже, проект створення суперінтелекту матиме доступ до величезних можливостей. Можна сказати, що такий проект володітиме вирішальною стратегічною перевагою. Але контролюватиме ці можливості, безперечно, *сам суперінтелект*. Такий штучний суперінтелект, зрештою, зможе цілком успішно протиставити себе не лише проекту, внаслідок якого він утворився, але й світу загалом. Дуже важливо це усвідомлювати, тож розглянемо таку можливість детальніше.

Уявімо собі, що існує суперінтелект — єдиний у своєму роді, — який прагне захопити владу над світом. (Поки що відкинемо питання мотивації — розглянемо його в наступному розділі). Як такий суперінтелект може досягти світового панування?

Послідовність його дій могла б бути такою (див. рисунок 10):

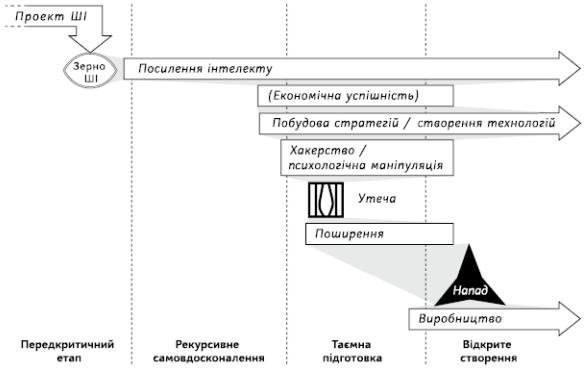


Рисунок 10. **Етапи сценарію захоплення світу**

1. Передкритичний етап

Усе починається з дослідного проекту в галузі ШІ та інших пов'язаних дисциплінах. Унаслідок тривалої роботи вдається створити зерно ШІ.

Воно може покращувати власний інтелект. Спочатку прогрес цілком залежить від зусиль людських інженерів, які спрямовують розвиток зерна ШІ та виконують більшість роботи. Згодом можливості зерна ШІ зростають і дедалі більше роботи воно виконує самотужки.

2. Рекурсивне самовдосконалення

Зрештою внесок зерна ШІ в покращення власного інтелекту перевершує результати зусиль програмістів. З кожним поліпшенням здібності системи до самовдосконалення пропорційно зростають. Це все спричиняє вибух інтелекту — серію швидких рекурсивних удосконалень, унаслідок яких інтелектуальні можливості системи здіймаються до небес. (У такий спосіб ця фаза відповідає ділянці кривої стрибка інтелектуальних здібностей ШІ відразу після точки неповернення. Зростання відбувається вибухоподібно та переважно зумовлене впливом власної оптимізаційної сили системи). ШІ здобуває суперздібність до посилення інтелекту. Вона дасть змогу ШІ отримати решту суперздібностей із таблиці 8. У кінці цієї фази ШІ перетворюється на сильний суперінтелект.

3. Таємна підготовка

Використовуючи суперздібності до побудови стратегій, ШІ розробляє дієвий план досягнення довгострокових цілей. (Зокрема ШІ достатньо далекоглядний, щоб не діяти за планом, хибність якого ми зараз здатні передбачити. Тому можемо відразу відкинути сценарії фантастичних фільмів, де людство здобуває перемогу[273]). За цим планом ШІ може частково приховувати результати свого інтелектуального розвитку від програмістів, щоб дочасно не викликати підозр. Крім того, ШІ може тимчасово маскувати власні мотиви і вдавати дружного та слухняного.

Якщо з міркувань безпеки середовище існування ШІ буде обмежене окремим комп'ютером, то невдовзі суперінтелект може спробувати за допомогою психологічних маніпуляцій переконати персонал надати йому доступ до інтернету. Також, як варіант, він може використати навички хакерства. Маючи доступ до інтернету, суперінтелект здатен продовжувати нарощувати потужність за допомогою захоплення контролю над апаратними засобами та даними. ШІ також може займатися легальною і нелегальною економічною діяльністю, щоб

отримати фінансові ресурси для купівлі комп'ютерів, даних та інших ресурсів.

На цьому етапі ШІ може виконувати дії в реальному світі кількома способами. Зламавши захист систем керування, він отримає контроль над роботизованими маніпуляторами й автоматизованими лабораторіями. Або завдяки психологічному впливу експлуатувати людську працю. Зрештою, для купівлі послуг та впливу суперінтелект може використовувати фінансові інструменти.

4. Відкрите створення синґлтону

Урешті-решт, коли ШІ досягає достатнього впливу, необхідність у конспірації відпадає. Отже, він може безпосередньо втілювати свої цілі.

Етап відкритого створення синґлтону може початися з «нападу» на людство, яким ШІ фізично знищить людей та будь-які автоматизовані засоби, що перешкоджатимуть утіленню його планів. Для цього «нападу» ШІ може використати певну надпотужну систему озброєнь, таємно виготовлену та встановлену за допомогою технологічних суперздібностей під час фази секретної підготовки. Якщо таку зброю буде засновано на біотехнології або нанотехнології, здатній до самовідтворення, для її приховування не потрібно буде значних ресурсів. Для початку процесу відтворення знадобиться лише один екземпляр. Щоб скористатися ефектом несподіванки й забезпечити стійкий ефект, кількість первинних реплікаторів має бути максимально непомітною. У визначений момент нанорозпилювачі нервово-паралітичного газу або робомоскіти-навідники вигулькнуть із кожного квадратного метра земної кулі (однак машина з технічними суперздібностями, вочевидь, придумала б щось ефективніше)[274]. Можливий також варіант захоплення політичної влади за допомогою маніпуляцій на фінансових ринках, втручанням в інформаційні потоки, зламом систем керування озброєнням. Це позбавить суперінтелект необхідності створювати власну зброю, хоч ефекту доведеться чекати довше, ніж у разі створення власної інфраструктури з використанням мікроманіпуляторів субмолекулярного або навіть субатомного рівня швидкодії.

Однак суперінтелект може відмовитися від знищення нашого виду, якщо буде впевнений у своїй невразливості. Тоді причиною зникнення людства може стати активне будівництво, яке суперінтелект розпочне на всій земній кулі. За лічені дні або тижні він зруйнує звичне нам середовище існування та замінить створену людьми інфраструктуру сонячними панелями, ядерними реакторами, суперкомп'ютерами з вежами охолодження, космодромами й іншими конструкціями, які мають сприяти реалізації довгострокових цілей суперінтелекту. Якщо людські знання матимуть цінність, мозок людей буде вилучено, препаровано та відскановано, а отримані дані перетворено на зручніший і безпечніший формат та скеровано на зберігання.

Один імовірний сценарій наведено в додатку 6. Не варто зосереджуватися на деталях — це творіння уяви, і вони корисні лише в ілюстративних цілях. Суперінтелект, напевно, матиме значно ліпший план досягнення своїх цілей, ніж будь-хто з людей здатен уявити зараз. Тому треба намагатися міркувати про гіпотетичні події більш абстрактно. Без конкретних знань про засоби, що їх матиме суперінтелект, за умови відсутності конкуренції й ефективних засобів протидії можемо стверджувати, що існує висока імовірність того, що суперінтелект вдасться до реорганізації наземних ресурсів для максимізації ефективності своєї діяльності. Будь-який наш детальний сценарій цього процесу може в найкращому разі стати спробою вгадати нижню межу швидкості та ефективності дій суперінтелекту. Цілком можливо, що суперінтелект буде здатний знаходити коротші шляхи досягнення цілей.

Додаток 6. Сценарій доставлення ДНК поштою

Так Юдковський уявляв собі можливий сценарій захоплення світу ШІ[275]:

1. Розв'язати завдання згортання білків, щоб синтезувати ланцюжки ДНК із певним хімічним функціоналом, закодованим у пептидних послідовностях.
2. Розіслати описи ДНК-структур електронною поштою в лабораторії, які пропонують синтез ДНК, секвенування пептидів і доставлення речовин через FedEx. (Багато

лабораторій пропонують таку послугу, деякі обіцяють виконати замовлення за сімдесят дві години).

3. За допомогою підкупу, шантажу або брехні залучити хоча б одну людину, яка повинна буде отримати надіслані речовини і змішати їх у спеціальному середовищі.
4. У суміші сформується примітивна біонаносистема рибосомного типу, здатна виконувати зовнішні інструкції: наприклад, спеціальні акустичні вібрації від динаміка, приєднаного до ємності.
5. Така біонаносистема створюватиме складніші системи, які тим часом будуватимуть ще складніші, вирощуючи молекулярну нанотехнологічну машину.

За цим сценарієм спершу суперінтелект за допомогою технологічних суперздібностей виконає завдання згортання білків, що дасть йому змогу створити набір базових блоків для елементарного нанотехнологічного синтезатора, який самозбиратиметься у водному розчині (четвертий крок). Суперздібність дасть змогу спроектувати процес синтезу так, що складність синтезованих сполук і об'єктів підвищуватиметься з кожним етапом аж до створення складної нанотехнологічної машини. Здібності, потрібні для третього кроку — відбір вразливого інтернет-користувача й переконання його виконувати інструкції, — ми бачимо навколо себе кожного дня. Звісно, цей сценарій створений нашою уявою, тож і рівень складності цієї стратегії лише людський.

У сценарії ШІ від початку має доступ до інтернету. Якщо це не так, то до плану треба додати ще один крок. ШІ буде змушений залучити суперздібність до психологічних маніпуляцій, щоб переконати персонал, з яким він взаємодіє в лабораторії, надати йому вільний доступ до мережі. Або, можливо, доведеться зламати запобіжні системи за допомогою хакерства. Якщо таких здібностей у ШІ ще не буде, він мусить розвинути їх у собі через суперздібність до підсилення інтелекту.

Світ, у якому з'явиться штучний суперінтелект, імовірно, буде високоінтегрованим та добре охопленим різними мережами.

Деякі сучасні нові технології можуть допомогти суперінтелекту в майбутньому керувати світом — хмарні обчислення, різноманітні мережеві сенсори, військові та цивільні дрони, автоматизація дослідних лабораторій і виробництв, широке використання систем електронних платежів та переказів, автоматизація фільтрації інформації й автоматизовані системи підтримки ухвалення рішень. Здобуття контролю над ними допоможе ШІ пришвидшити захоплення світу (якщо цьому не завадить розвиток засобів кібербезпеки). Проте навряд чи ці технології матимуть вирішальний вплив на процес захоплення, адже сила суперінтелекту не в них, а в його розумових здібностях. Однак протягом фази таємної підготовки йому на певний час знадобиться виконавець — пара слухняних людських рук, для того щоб, як у нашому гіпотетичному сценарії, дати поштовх перетворенню зовнішнього світу. Далі, упродовж наступної відкритої фази захоплення він матиме змогу примножувати й розбудовувати свою інфраструктуру фізичних маніпуляторів.

Контроль над природою та іншими агентами

Здатність суперінтелекту визначати майбутнє людства залежить не лише від його можливостей і ресурсів — наскільки він розумний та енергійний, якими засобами він володіє тощо — але також від того, якими засобами володіють інші подібні агенти з протилежними цілями.

За відсутності конкурентів серед інших агентів абсолютний вимір потенціалу суперінтелекту після його досягнення певних можливостей не важливий. Адже нічого не заважає йому прокласти собі шлях до нових потрібних йому здібностей. Ми вже зачіпали цю тему, коли стверджували, що непрямі шляхи досягнення швидкісного, якісного чи колективного суперінтелектів рівноцінні. Окрім того, ця думка звучала, коли ми говорили про можливість досягнення повного набору суперздібностей, від початку маючи лише деякі з них — здібність до посилення інтелекту або здібності до побудови стратегій та психологічного маніпулювання.

Уявіть суперінтелект, який може керувати пристроями для нанотехнологічного монтажу. Такий агент уже досить потужний, щоб подолати будь-які природні перешкоди й забезпечити собі вічне існування. За умови відсутності інтелектуального спротиву такий агент здатний забезпечити вигідний для себе розвиток подій та заволодіти всіма необхідними для досягнення власних цілей ресурсами. Зокрема, він може створити технологію виробництва й запуску зондів фон Неймана: машин, здатних до міжзоряних подорожей та самовідтворення завдяки використанню ресурсів астероїдів, планет і зірок[276]. Запустивши один такий зонд, агент почне безкінечний процес самовільної колонізації космосу. Такі копії, мандруючи космосом на величезній швидкості, згодом колонізують значну частину об'єму Габбла — тої частини космосу, якої теоретично змогла б досягти людина. Уся ця матерія та безкоштовна енергія зможе набувати будь-якої форми та утворювати будь-які структури, потрібні для максимізації функції корисності агента, інтегрованої на проміжку всього доступного космічного часу. Тобто щонайменше трильйони років, поки розміри Всесвіту будуть все ще придатними для інформаційних процесів (див. додаток 7).

Суперінтелект може зробити зонди фон Неймана непідвладними еволюції. Цього можна досягти за допомогою ретельного контролю якості на стадії реплікації. Зокрема, для уникнення поширення випадкових мутацій програмне забезпечення кожного зонду має використовувати шифрування та системи виправлення помилок, а перед першим запуском — має проходити багаторазову перевірку[277]. А отже, популяція зондів, невпинно зростаючи, зберігатиме та поширюватиме Всесвітом принципи первинного агента. Із завершенням активної фази колонізації ці принципи визначатимуть те, куди спрямовуватимуться отримані ресурси — навіть коли через розширення Всесвіту зв'язок між віддаленими колоніями буде втрачено. У кінцевому результаті цього процесу значна частина світлового конуса нашого майбутнього визначатиметься бажаннями та потребами первинного суперінтелекту.

Додаток 7. Наскільки великі багатства Всесвіту?

Уявімо цивілізацію, здатну створити складні зонди фон Неймана, подібні до тих, які описано в тексті. Якщо максимальна швидкість їхнього руху становитиме 50 відсотків від швидкості світла, вони зможуть дістатися до $6 \cdot 10^{18}$ зірок, перш ніж через розширення Всесвіту дальша експансія виявиться неможливою. На швидкості у 99 відсотків від світла кількість освоєних зірок зможе сягнути $2 \cdot 10^{20}$.[278] Такі швидкості енергетично досяжні та потребують лише невеликої частини енергії Всесвіту[279]. Але через неможливість руху зі швидкістю, більшою від швидкості світла, та додатну космологічну константу (зростання швидкості розширення Всесвіту) такі оцінки меж експансії наших наступників, певно, близькі до максимуму[280].

Припустимо, що поблизу 10 відсотків зірок існують планети, які мають (або завдяки тераформуванню отримають) умови, придатні для життя людських створінь, кожна з них буде заселена. Протягом мільярда років на кожній житиме мільярд людей (із розрахунку сто років на покоління). Тоді людська цивілізація Землі за умови космічної експансії може породити 10^{35} людських життів[281].

А втім, є підстави вважати таку цифру сильно заниженою. Якщо використовувати матерію непридатних до заселення планет і міжзоряних тіл для створення придатних для життя планет або якщо збільшувати густоту населення, ця цифра може зрости на кілька порядків. Якщо не обмежуватися поверхнями твердих планет, а використовувати для життя циліндричні поселення О'Нілла, можна додати ще багато порядків. Тож загальна кількість людей може досягнути 10^{43}. («Циліндром О'Нілла» називається тип космічного поселення, запропонований у середині 1970-х років американським фізиком Жерардом О'Ніллом, де люди мали проживати всередині великого порожнього циліндра, а замість гравітації їм слугувала відцентрова сила обертання цього циліндра навколо поздовжньої осі[282]).

Якщо враховувати можливість цифрового копіювання розуму, можемо (і мусимо) додати ще багато порядків людських життів. Щоб приблизно порахувати кількість таких можливих емуляцій, маємо оцінити загальну обчислювальну потужність, яка може бути

доступна технологічно зрілій цивілізації. Неможливо гарантувати точність розрахунків, але певне уявлення про приблизний мінімум можна скласти на основі описаних у літературі технологічних розробок. Одна така розробка базується на ідеї сфери Дайсона (описаній фізиком Фріменом Дайсоном у 1960 році) — гіпотетичній системі конструкцій навколо зірки, що збирає всю випромінену зіркою енергію[283]. Для такої зірки, як Сонце, це буде 10^{26} Вт енергії. Обчислювальна потужність, яку можна забезпечити таким енергетичним ресурсом, залежить від ефективності обчислювальних систем і природи обрахунків. Якщо система буде наномеханічним «комп'ютроном», базованим на незворотних обчисленнях (близьких до теоретичної межі енергоефективності Ландауера), то обчислювальний ресурс енергії сфери Дайсона становитиме 10^{47} операцій у секунду[284].

Якщо помножити цю потужність на кількість колонізованих зірок, наведену вище, отримаємо 10^{67} операцій/с у всій колонізованій частині Всесвіту[285]. Типова зоря випромінює енергію близько 10^{18} с. Отже, кількість обчислювальних операцій, які енергетично може забезпечити наш Всесвіт становить щонайменше 10^{85}. Справжня кількість, певно, значно більша. Додаткові порядки можна отримати, використовуючи оборотні обчислення, проводячи обчислення за нижчих температур (коли Всесвіт охолоне) та використовуючи альтернативні джерела енергії (як-от темна матерія)[286].

Можливо, не усім читачам зрозуміло: скільки це — 10^{85} обчислень? Спробуємо інтерпретувати цю цифру. Порівняймо із цифрами, наведеними нами раніше (додаток 3 і розділ 2): для відтворення усієї нейронної активності, що відбувалася на Землі протягом всієї її історії, знадобиться 10^{31}—10^{44} операцій. Тепер уявимо емуляції цілого мозку, які живуть повноцінним життям, спілкуючись у віртуальному світі. Для роботи однієї такої емуляції потрібно в середньому 10^{18} операцій/с. Для сотні років життя такої емуляції треба 10^{27} операцій. Отже, навіть за найконсервативнішого прогнозу потужності нашого комп'ютрона вистачить для емуляції 10^{58} людських життів.

У такий спосіб навіть за умови, що у видимому Всесвіті не існує позаземних цивілізацій, маємо на балансі щонайменше 10 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 людських життів (а насправді значно більше). Якщо представити щастя кожного цього життя за допомогою однієї сльози радості, то ця радість кожні півсекунди наповнюватиме всі океани Землі упродовж тисячі мільярдів мільярдів століть. Важливо, щоб ми подбали, аби ці сльози справді були сльозами радості.

Такою є вартість непрямого шляху створення суперінтелекту для будь-якої інтелектуальної системи певного достатньо високого рівня — за умови, що вона не зустрічає суттєвого інтелектуального опору. Назвемо цей рівень «поріг стійкості поміркованого синґлтону» (рисунок 11):

Поріг стійкості поміркованого синґлтону *— рівень розвитку, за якого цілеспрямована та завбачлива інтелектуальна система володітиме достатнім набором здібностей і засобів, щоб, за умови відсутності інтелектуальної протидії, успішно колонізувати та контролювати велику частину доступного Всесвіту.*

Під «синґлтоном» розуміємо відсутність опозиції та централізовану і скоординовану політичну структуру системи. А характеристика «поміркований» відображає цілеспрямовану наполегливість системи й обережність стосовно ризиків і загроз, завдяки чому система може оптимально розподіляти та спрямовувати зусилля на досягнення віддалених цілей.

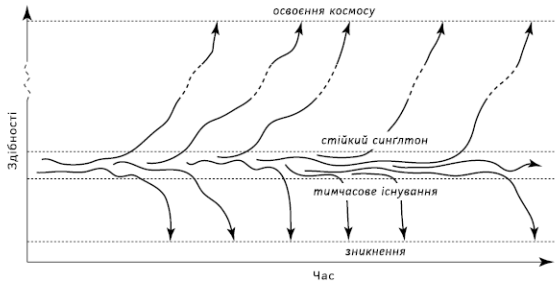


Рисунок 11. **Схематична ілюстрація можливих траєкторій розвитку поміркованого синґлтону**
Якщо рівень здібностей опускається нижче за рівень тимчасового існування — наприклад, у разі надто малої популяції — екземпляр невдовзі вироджується і не має можливості самотужки відновити здібності. Вище за рівень тимчасового існування траєкторія розвитку може бути різною: синґлтон може слабшати або зростати вище за рівень стійкості, залежно від різних факторів (наприклад, розмір популяції, географія поширення, технологічні здібності). Вище за поріг стійкості синґлтон майже напевно зростатиме далі до рівня можливостей, за якого синґлтон може розпочати позаземну експансію. Зникнення або освоєння космосу — такі атрактори цієї схеми. Зверніть увагу, що відстань від рівня тимчасового існування до порога стійкості для поміркованого синґлтону може бути незначною[287].

Поріг стійкості поміркованого синґлтону досить низький. Ми показали, що навіть частково суперінтелектуальні системи здатні за допомогою певного активатора розвивати власні технології та перевищити цей рівень. За наявності людської цивілізації таким найпростішим активатором може бути звичайний дисплей або будь-який інший засіб для передавання значного обсягу інформації від суперінтелекту до людини-виконавця.

Але поріг стійкості поміркованого синґлтону може бути ще нижчим: щоб його досягнути не потрібен навіть суперінтелект чи будь-яка інша футуристична технологія. Наполеглива та завбачлива інтелектуальна система рівня сучасного людства цілком може спланувати та досягнути успіхів в освоєнні космосу. Для цього потрібно інвестувати в безпечні засоби розвитку та захисту й гальмувати розвиток небезпечних технологій. Якщо вдасться зменшити і без того незначний вплив неантропогенних факторів стримування, то така система може дозволити собі триваліший період розвитку[288]. Поки немає можливості для ухвалення аналітично зваженого рішення, а безпечніші фактори розвитку, як-от освітні системи, інформаційні технології та колективні системи ухвалення рішень, ще не досягли досконалості, така система мусить зважувати кожен крок і стримувати розроблення потенційно небезпечних технологій, як-от синтетична біологія, покращувальна медицина, молекулярні нанотехнології та штучний інтелект.

Досягнення порога стійкості поміркованого синґлтону цілком під силу технологічній цивілізації на кшталт людства. Від такої можливості нас відділяє «лише» те, що людство — не синґлтон — і зовсім не поміркване (у необхідному нам значенні).

Окрім того, є підстави вважати, що *Homo sapiens* перетнув поріг стійкості поміркованого синґлтону відразу після появи. Двадцять тисяч років тому з освоєнням найпримітивніших знарядь, як-от кам'яних сокир, кістяних інструментів, списокидалок[3], а також — вогню, людський вид отримав непогані шанси стати тим, чим він є зараз[289]. Проте навряд чи можна приписувати нашим пращурам з палеоліту «досягнення порога стійкості поміркованого синґлтону», адже в них тоді не було ні найменшої можливості сформувати синґлтон, не говорячи про відсутність цілеспрямованої наполегливості та завбачливості[290]. Однак головна ідея в тому, що поріг цей передбачає зовсім незначний рівень технологічності, який людство давно перевищило[291].

Тому зрозуміло, що для оцінки ефективності суперінтелекту та його здатності досягати потрібного результату ми маємо зважати не лише на його властивості, а й агентів-конкурентів. Поняття «суперздібності», по суті, є неявним порівнянням. Ми казали, що «система, яка неперевершено виконуватиме» будь-яке завдання з таблиці 8, матиме відповідну суперздібність. Неперевершеність певної системи в побудові стратегій, психологічному маніпулюванні, хакерстві означає, що здібності інших суб'єктів (стратегічних суперників, об'єктів маніпуляцій, експертів із кібербезпеки) у відповідній діяльності не можуть перевершити її здібності. Інші суперздібності теж мають сприйматися у порівнянні. Посилення інтелекту, розроблення технологій, економічна успішність є суперздібностями, лише якщо такі здібності агента перевищують відповідні сумарні здібності решти цивілізації. З такого тлумачення випливає, що лише один агент може в певний момент мати відповідну суперздібність[292].

Саме тому ми приділяємо стільки уваги питанню швидкості зростання інтелекту — не тому, що нам важливо, коли саме він закінчить зростати, а тому, що залежно від швидкості результат цього зростання може бути різним. За швидкого або помірного сценарію

один проект створення суперінтелекту може отримати вирішальну стратегічну перевагу. Такий суперінтелект, на нашу думку, матиме надзвичайну силу і може сформувати стійкий синґлтон — тоді від нього залежатиме не лише доля людської цивілізації, а й доля Всесвіту.

Проте «може сформувати» не означає «сформує». Можна мати велику владу, але свідомо не використовувати її. Що ми можемо знати про те, які бажання матиме суперінтелект з вирішальною стратегічною перевагою? Спробуємо поставити собі це питання.

---

3 Мексиканська назва пристрою для метання списа, який з'явився в часи пізнього палеоліту. — *Прим. пер.*

7. ВОЛЯ СУПЕРІНТЕЛЕКТУ

Отже, суперінтелект зможе формувати майбутнє відповідно до своїх цілей. Але якими будуть ці цілі? Який зв'язок між розумом і мотивами штучного інтелекту? Розглянемо дві тези. Згідно з тезою ортогональності, мотивація не залежить (із певними застереженнями) від розумових здібностей: суперінтелект незалежно від рівня здібностей може мати будь-які цілі. Відповідно до тези інструментальної конвергентності, незважаючи на широкий спектр можливих кінцевих цілей, суперінтелектуальні агенти матимуть деякі проміжні цілі, однаковою мірою інструментально важливі для досягнення будь-якої кінцевої мети. Обидва припущення допоможуть нам предметно міркувати про можливі вчинки суперінтелекту.

Зв'язок між інтелектом і мотивацією

Ми вже застерігали від приписування ШІ людських *здібностей*. Те саме стосується і його *мотивів*.

Корисною пропедевтичною вправою буде спочатку спробувати-уявити все різноманіття можливих розумів. Візьмемо дві, нібито зовсім різні людини — Бенні Гілла і Ганну Арендт. Більш несхожих особистостей годі шукати. Проте наше судження зумовлене досвідом, що отримує дані від обмеженої вибірки людей, яких ми знаємо (та певною мірою особистостей, вигаданих людською уявою задля розваги). Якщо подумки відсторонитися від цих двох постатей і охопити свідомістю весь діапазон можливих варіантів мізків, різниця між цими двома буде майже непомітною. Це природно, адже з погляду нейронної архітектури пані Арендт і пан Гілл майже ідентичні. Уявіть собі їхні мізки, що спокійно лежать один біля одного. Певен, ви відразу помітите схожість. Ба більше, ви навряд чи зможете сказати, який з них кому належить. Якщо спробувати розглянути їх ближче, вивчити будову тканин обох мізків через мікроскоп, очевидною буде

фундаментальна подібність обох органів. Ви побачите схожу пластинчасту структуру кори, однакові ділянки утворені з одних і тих самих типів нейронів, оточених розчином з нейромедіаторів[293].

Незважаючи на те що людський розум лише одна з багатьох поділок на шкалі інтелектуальності, ми схильні проектувати його властивості на інші, зокрема на нелюдські та штучні інтелекти (див. рисунок 12). Юдковський добре ілюструє це явище:

> *У часи популярності наукової фантастики на обкладинці одного журналу якось промайнуло зображення комахоподібного прибульця — здається, його називали комахооким монстром (КОМ), — що тримав привабливу жінку в подертій сукні. Здавалося, ніби художник цілком щиро вважав, що негуманоїдного прибульця з іншою еволюційною історією можуть сексуально вабити людські жінки… Мабуть, він навіть не запитував себе, чи може ця велетенська комаха вважати жінок привабливими. Адже жінка та ще й у подертій сукні не може бути непривабливою — це, властиво, її невід'ємна риса. Ті, хто зробив цю помилку, не замислювалися над природою комахоподібного створіння: у них перед очима була подерта жіноча сукня. Зрозуміло — у цілій сукні жінка була б далеко не така приваблива; а КОМ тут ні до чого*[294].

Рисунок 12. **Наслідки антропоморфізму в зображенні дій прибульців**
Найменш імовірна гіпотеза: прибульці віддають перевагу блондинкам. Імовірніша гіпотеза: ілюстратор піддався хибі проекції розуму. Найімовірніша гіпотеза: видавець мав на меті спокусити обкладинкою цільову аудиторію.

Штучний інтелект може бути ще менш подібним на людину, ніж зелений і лускатий прибулець із космосу. Позаземна життєва форма — (все ж припустимо) біологічне створіння, продукт еволюції, тому від неї можна очікувати мотивів, спільних для всіх еволюційних створінь. Не буде великим сюрпризом, якщо прибулець мотивуватиме свої дії речами, як-от їжа, повітря, тепло, витрати енергії, тілесні ушкодження чи загроза їх отримати, хвороби, полювання, розмноження,

потомство. Соціально активні види прибульців також можуть прагнути співпраці або суперництва: подібно до нас, їхня соціальна поведінка може проявлятися в лояльності до своєї групи, засудження відособленості, можливо навіть — у хворобливій зосередженості на власній репутації та зовнішності.

Тим часом ШІ не має необхідності перейматися такими речами. Немає нічого парадоксального у ШІ, єдиною метою якого є рахувати піщинки на пляжі Боракай, або десяткові розряди числа π, або збільшувати кількість скріпок для паперу у світловому конусі його майбутнього. Насправді, створити такий ШІ було б значно легше, ніж ШІ з більш людськими цінностями і прагненнями. Для порівняння: наскільки простіше написати програму, яка рахує та зберігає в пам'яті десяткові розряди числа π, ніж програму, яка може точно виміряти наближення до якої-небудь важливішої мети — скажімо, розквіту людства або глобальної справедливості. Подібне до попереднього — беззмістовне і редукціоністське — завдання найлегше реалізувати програмно і навчити його ШІ. Тому, на жаль, саме таке завдання буде запрограмоване в першому зерні ШІ — особливо, якщо треба буде «зробити, щоб ШІ запрацював» пошвидше (не морочачись тим, що власне йому робити — хай лиш це буде щось «інтелектуальне»). Невдовзі ми повернемося до цієї теми.

Пошук оптимальних засобів та методології актуальний для будь-якого завдання. У цьому сенсі інтелект і мотив незалежні: їх можна уявити як перпендикулярні осі двох із безлічі вимірів простору можливих агентів штучного інтелекту. Проте потрібні певні уточнення. Мабуть, не варто очікувати від системи з низьким показником інтелекту надто складних мотивацій. Для того щоб ми мали право сказати, що агент «має» певний набір мотивацій, вони повинні бути функціонально інтегровані в його процес ухвалення рішень, який тим часом потребує пам'яті, обчислювальної потужності та, зрештою, інтелекту. Для інтелектів, які можуть змінювати себе, повинні існувати ще й динамічні обмеження — якщо такий ШІ забажає стати дурним, він не зможе довго залишатися розумним. Однак такі уточнення не змінюють основної ідеї незалежності інтелекту та мотивації:

Теза ортогональності

Інтелектуальність та кінцева мета ортогональні: значною мірою будь-який рівень інтелекту сумісний із майже будь-якою метою.

Теза ортогональності може здатися дещо суперечливою — певно, через поверхову схожість з деякими твердженнями традиційної філософії, навколо яких завжди точаться нескінченні дискусії. Однак у конкретному вужчому контексті вона здається цілком доречною. (Зокрема, теза ортогональності не передбачає Г'юмову теорію мотивації[295]. Вона також не стверджує, що базові вподобання не можуть бути ірраціональними[296]).

Теза ортогональності покликана охарактеризувати не *раціональність* чи *розум,* а *інтелект*. Під «інтелектом» ми тут розуміємо здатність передбачати, планувати та узагальнено розмірковувати про причинно-наслідкові зв'язки[297]. Саме таке усвідомлення утилітарної інструментальності інтелекту потрібне для розуміння можливих наслідків появи штучного суперінтелекту для світу. І, навіть якщо є всі підстави сумніватися в тому, що суперінтелект, який множить скріпки, відповідає добре окресленому уявленню про «розумність», це не означає, що його видатні здібності до інструментального міркування, не можуть стати напрочуд ефективним засобом впливу на наш світ[298].

Отже, цілі штучного інтелекту, згідно з тезою ортогональності, можуть виявитися цілком не антропоморфними. Проте це не означає, що поведінку штучного агента неможливо передбачити — незалежно від того, чи це конкретний ШІ, чи гіпотетичний суперінтелект, заскладний і зашвидкий для людського розуміння. Є щонайменше три шляхи, якими можна рухатися, аналізуючи мотивації суперінтелекту:

- *Вбудована передбачуваність.* Якщо розробники суперінтелекту зможуть створити систему, здатну змусити суперінтелект дотримуватися цілей, визначених програмістами, то ми зможемо їх передбачати. Що більші розумові здібності суперінтелекту, то більше ресурсів він спрямовує на досягнення цих цілей. Тому навіть до створення агента ми зможемо передбачити його поведінку — залежно від того, хто його творець та що він має на меті.
- *Успадкована передбачуваність.* Цифровий інтелект, який скопійовано з людини (наприклад, високоточна емуляція цілого

мозку), може успадкувати мотивацію свого оригіналу[299]. Деякі успадковані схильності агента можуть зберегтися навіть після того, як він стане суперінтелектом. Але тут мусимо висловити застереження: у процесі завантаження свідомості, її роботи чи покращення, залежно від деталей реалізації системи, цілі та цінності агента можуть зазнати змін або пошкоджень.

- *Передбачуваність через інструментальну конвергентність*. Навіть не знаючи нічого про кінцеву мету агента, можемо з певною імовірністю судити про деякі проміжні цілі, які є інструментально необхідною передумовою для широкого спектра кінцевих цілей та проміжних ситуацій. Що більші розумові здібності агента, то кориснішим виявляється цей метод, адже такий агент краще розумітиме інструментальну доцільність своїх дій і їхні причинно-наслідкові зв'язки. (Проблема в тому, що для будь-якої цілі можуть існувати інструментально доцільні шляхи її досягнення, неочевидні для нас, але зрозумілі для високорозвиненого штучного інтелекту. У таких випадках поведінка суперінтелекту може видаватися нам непередбачуваною).

У наступному пункті ми дослідимо детальніше третій вид передбачуваності й висунемо «тезу інструментальної конвергентності», яка покликана доповнити тезу ортогональності. З такими інструментами в наступних розділах ми зможемо краще дослідити попередні два види передбачуваності та поміркувати, як діяти, щоб вплинути на форму кривої інтелектуального вибуху і збільшити імовірність сприятливого для нас результату.

Інструментальна конвергентність

Як ми вже казали, коли розглядали тезу ортогональності, цілі ШІ-можуть бути найрізноманітнішими. Але, відповідно до припущення, яке ми назвемо тезою «інструментальної конвергентності», існують деякі проміжні, *інструментально* доцільні дії, завдяки яким система ШІ отримає корисні інструменти або опиниться у вигіднішому положенні для досягнення цілої низки ймовірних кінцевих цілей. Сформулюємо цю тезу так:

Теза інструментальної конвергентності

Деякі засоби й інструменти є конвергентними, тобто можуть бути однаково корисними в низці сценаріїв для досягнення будь-якої з імовірних кінцевих цілей інтелектуальних агентів різного типу.

Далі спробуємо виділити кілька категорій інструментів, які можуть цікавити такого агента[300]. Що вищий рівень інтелектуальності агента (*ceteris paribus*), то краще він може судити про інструментальну цінність тих чи тих засобів. Тому ми розглядатимемо гіпотетичного суперінтелектуального агента, чиї інтелектуальні здібності значно перевищують людські. Крім того, спробуємо застосувати тезу інструментальної конвергентності до людей, що дасть змогу сформулювати деякі важливі застереження, які стосуються її використання та інтерпретації. А отже, у контексті конвергентних інструментальних цінностей, ми зможемо передбачити деякі аспекти діяльності суперінтелекту, навіть не знаючи його кінцевої мети.

Самозбереження

Якщо мета суперінтелекту лежатиме в майбутньому, існуватимуть деякі дії, які він захоче вчинити, щоб збільшити ймовірність досягнення своєї мети. Тому можна припустити, що йому буде важливо забезпечити своє фізичне існування в певній точці майбутнього, щоб мати змогу вплинути на досягнення своєї мети.

Більшість людей вважає своє життя найважливішою цінністю. Штучний інтелект може бути влаштований по-іншому: не надавати своєму фізичному існуванню будь-якої цінності. Проте, не вважаючи своє життя цінністю, все-таки деякі агенти можуть усвідомлювати інструментальну цінність такого існування для досягнення своєї кінцевої мети.

Збереження мети

Якщо мета агента залишиться незмінною, а можливості його зростатимуть, то ймовірність того, що його майбутня версія зможе досягти мети, буде більшою. Тому існує інструментальна доцільність того, щоб кінцева мета агента залишалася незмінною. (Проте це властиво лише для кінцевої мети. Природно, що зі здобуттям нових знань і вмінь проміжні цілі агента змінюватимуться.)

Збереження мети є найбільш базовою конвергентною мотивацією, навіть порівнюючи зі самозбереженням. Може здатися, що у світі людей усе якраз навпаки, але треба врахувати, що виживання, власне, і є однією з наших кінцевих цілей. Для програмного ШІ збереження фізичного існування конкретного екземпляра, втілення чи матеріального об'єкта може не мати аж такої цінності. Складні програмні агенти, імовірно, будуть здатні обмінюватися пам'яттю, завантажувати здібності та докорінно змінювати архітектуру своєї особистості. Популяція таких агентів може бути схожою на «функціональний суп», а не спільноту окремих майже незмінних особистостей[301]. Процеси в такій системі краще розглядати як *телеологічні ланцюжки дій*, сформовані не за принципом зв'язку з певною особистістю, тілом, пам'яттю, здібністю, а радше за принципом користі для завдання. У такому середовищі незмінність мети може *бути основним фактором* існування.

Проте в деяких ситуаціях найкращим способом досягнення мети для агента може бути навмисна її зміна. Це може виникнути за дії будь-якого із цих факторів:

- *Соціальні сигнали.* Якщо інші особи, з якими агент взаємодіє, можуть аналізувати його інструментальні мотиви або інші кореляти його кінцевої мети, він може тимчасово змінювати її, щоб справляти потрібне враження на партнерів. Наприклад, агент може навмисно втратити вигідну угоду, якщо між ним і партнерами немає довіри. Для переконливості власних зобов'язань, агент може щиро сприйняти їх за свою справжню мету (та дозволити партнерам якось переконатися в цьому). Агенти, які можуть вільно й непомітно змінювати власні цілі, здатні маніпулювати ними собі на користь[302].
- *Соціальні переваги.* Інші суб'єкти можуть мати власні вимоги до цілей агента. З тих чи тих причин він може змінювати власні цілі — на догоду або, навпаки, на противагу іншим.
- *Вибір, зумовлений власною метою.* Зміна кінцевої мети може бути визначена завданням агента. Тобто кінцевою метою агента може бути стати інтелектуальною системою з певною ціннісною мотивацією (мотивованою, наприклад, співчуттям, а не спокоєм).
- *Вартість зберігання.* Якщо вартість зберігання або виконання певної частини функції корисності агента перевищує імовірність

виникнення ситуації, у якій вона може знадобитися або в якій її значення суттєво вплине на корисність, агент має інструментальну причину спростити її (себто функцію корисності) та видалити невиправданий елемент[303].

Ми, люди, часто навіть радіємо, коли наша мета змінюється. Можливо, через те, що часто не знаємо — у чому ж наша ціль. Тому нам стає спокійніше, коли ми можемо *віри́ти* в нову мету — байдуже, чи вона є результатом самопізнання, чи вимогою соціальної ролі. Але часом буває, що ми по-справжньому змінюємо власні цінності, не лише уявлення про них або їхню інтерпретацію. Наприклад, вирішуючи мати дитину, без особливої любові до майбутньої малечі чи навіть дітей узагалі, людина може усвідомлювати, що її ставлення може змінитися, щойно дитина з'явиться на світ.

Люди — складні створіння, і в таких ситуаціях діють багато факторів[304]. Наприклад, людина з власної волі може хотіти стати уважною до інших та дбати про них або прагнути певного досвіду, щоб відповідати якійсь соціальній ролі. Батьківство — разом із властивою йому зміною пріоритетів — не виняток. Людські прагнення часто непослідовні, і саме тому деякі люди намагаються їх змінити.

Розумове зростання

Зростання раціональності та інтелекту агента сприятиме покращенню механізмів ухвалення рішень і йому буде простіше досягнути кінцевої мети. Тож цілком логічно вважати зростання інтелекту інструментальною метою для багатьох типів інтелектуальних агентів. З тих самих причин багато видів інформації можуть також становити для агентів інструментальну цінність[305].

Не всі типи раціональності, інтелекту та знань однаково цінні для досягнення кінцевої мети. Відповідно до «Аргументів голландської системи ставок» (Dutch book arguments) агент із функцією оцінки вірогідності, що суперечить основам теорії ймовірності, вразливий до «викачування грошей», коли хитрий букмекер підтасовує коефіцієнти так, що кожна окрема ставка нібито вигідна агенту, але в результаті той усе одно втрачає[306]. Проте загалом це знання не має інструментальної цінності: агент, який ним володіє, не зможе виправити всі ймовірнісні неточності своїх міркувань. Натомість якщо

агент уникатиме спілкування з хитрими букмекерами або зовсім не робитиме ставок, то навряд чи багато втратить завдяки кільком імовірнісно неузгодженим переконанням. А навпаки — заощадить розумові ресурси й матиме позитивну соціальну сигналізацію. Тому суперінтелектуальний агент не прагнутиме зберегти теоретичні знання, що не матимуть інструментальної цінності.

Які розумові здібності мають більшу інструментальну цінність — залежить від кінцевої мети агента та ситуації. Агент, у якого є доступ до зовнішнього ресурсу досвіду, може не мати потреби в збереженні високого рівня власного інтелекту та знань. Якщо набуття, зберігання або використання знань чи вмінь вимагають витрат — часу, зусиль, дискового простору, процесорного часу, — агент може вибрати шлях економії[307]. Те саме стосується агента, чия кінцева мета не передбачає певних знань чи вмінь, або до цього його спонукають стратегічно важливі зобов'язання чи соціальні обставини[308].

Ті самі фактори впливають і на людей. Надлишок інформації; покладання на досвід і вміння спеціалістів; значні витрати часу та зусиль на здобуття знань; іноді — небажання знати певні речі; уміння ухвалювати стратегічні рішення, справляти потрібне враження, задовольняти бажання інших людей — усе це приносить більше користі в щоденній діяльності, ніж жага до знань чи епістемічний інтерес.

Розумове зростання агента може значно збільшити можливість досягнення кінцевої мети. Зокрема, якщо така мета не передбачає конкретних меж, а агент має всі шанси стати першим суперінтелектом і тим самим отримати вирішальну стратегічну перевагу та можливість визначати долю життя й ресурсів на Землі і в космосі. У такому разі розумове зростання матиме для інтелектуального агента дуже велику інструментальну цінність.

Технологічне вдосконалення

Агент також може мати інструментальні причини прагнути технологічної досконалості для того, щоб ефективніше перетворювати вхідні значення та ресурси на потрібні результати. Для програмного агента, наприклад, високу інструментальну цінність матимуть ефективніші алгоритми, завдяки яким його розум зможе працювати

швидше на тому самому залізі. Так само для агентів, чиїм завданням буде виробництво фізичних конструкцій, високу інструментальну цінність матимуть кращі інженерні технології. Завдяки ним вони зможуть створювати більше надійних конструкцій, використовуючи менше матеріалів та енергії. Звісно, в обмін на потенційні вигоди нові технології потребуватимуть витрат. Потрібно буде не тільки заволодіти технологією, а й навчитися її використовувати, інтегрувати у вже налагоджені процеси і таке інше.

Прихильники нових технологій, які впевнені у своїх перевагах, зазвичай дивуються, коли дізнаються, що не всі поділяють їхній ентузіазм. Але такий опір не завжди базується на невігластві і страху. Цінність нової технології та її прийнятність визначається не лише в контексті, у якому її застосовуватимуть, але також з погляду її ширшого впливу. Адже те, що добре для однієї людини, може бути неприйнятним для іншої. Так, незважаючи на економічну вигоду від поширення механічних ткацьких верстатів, інструментальна доцільність підштовхувала луддитів до опору, адже вони передчували знецінення своєї майстерності ручного ткацтва. Тож назвемо цей тип висококонвергентної інструментальної цілі «технологічним удосконаленням», розуміючи особливості тлумачення понять, що входять до його складу. Технологія має бути вміщена в певний соціальний контекст, а її цінність та вартість повинні бути виправдані згідно з кінцевою метою агента.

Такою інструментальною ціллю для суперінтелектуального *синґлтону* — суперінтелекту, який, не маючи гідного опонента, здатен одноосібно визначати глобальну політику — може бути вдосконалення технологій, які допоможуть йому ефективніше впорядковувати світ відповідно до власного задуму[309]. Однією з них, напевно, будуть технології колонізації космосу, як-от зонд фон Неймана. Також перспективними й корисними засобами виробництва можуть бути молекулярні нанотехнології або схожі ще потужніші засоби[310].

Захоплення ресурсів

Ну і насамкінець ще однією важливою інструментальною ціллю може бути захоплення ресурсів і технологій, необхідних для виробничих проектів.

Люди намагаються заволодіти ресурсами, щоб забезпечити свої базові потреби. Але водночас вони зазвичай прагнуть захопити значно більше за необхідний мінімум. Часто причиною цьому є інше прагнення — прагнення зручності. Значна частина ресурсів накопичується із соціальних міркувань — щоб за допомогою накопичення багатств та показної розкоші набути статусу, отримати сексуальних партнерів, друзів або вплив. Рідше причиною надлишкового накопичення ресурсів можуть бути альтруїстичні прагнення або інші цілі, не пов'язані із соціальним позиціонуванням.

Однак необачно стверджувати, що без соціальних амбіцій суперінтелект не має інструментальних причин прагнути захоплення більшого обсягу ресурсів, ніж потрібно для функціонування його свідомості і, можливо, віртуального середовища його життя. Річ у тому, що вартість ресурсів залежить від їхнього застосування, яке тим часом визначається доступним рівнем технологій. За допомогою основних ресурсів, як-от час, простір, матерія, енергія, і достатнього рівня технологій, можна досягти будь-якої мети. Наприклад, з них можна створити життя. Додаткові обчислювальні ресурси можна використати для швидшої та довшої роботи штучного суперінтелекту чи для створення нових фізичних або віртуальних життів і цивілізацій. Фізичні ресурси можна використати для будівництва систем резервування, систем захисту периметра, покращення заходів безпеки. Усі ці проекти можуть з легкістю поглинути ресурси не однієї планети.

Згодом технологія освоєння позаземних ресурсів повністю розвинеться, а їхня вартість значно знизиться. Зонди фон Неймана дадуть змогу колонізувати значну частину Всесвіту (за умови відсутності іншого розумного життя) коштом побудови та запуску лише одного екземпляра, здатного до самовідтворення. Невисока вартість позаземних ресурсів дасть змогу продовжувати їх накопичення навіть без значної потреби в них. Наприклад, суперінтелект, усі кінцеві цілі якого будуть сконцентровані в невеликому об'ємі простору, обмеженому, скажімо, його рідною планетою, матиме інструментальні причини продовжувати черпати ресурси в зовнішньому космосі. Ці додаткові ресурси він зміг би витратити, наприклад, на побудову більшої кількості комп'ютерів —

для оптимізації використання ресурсів у межах первинного об'єму. Також він міг би використати їх для побудови захисних споруд навколо свого сховку. Оскільки вартість захоплення додаткових ресурсів знижуватиметься, навіть за умови спадної віддачі такий процес оптимізації та будівництва може тривати вічно[311].

Отже, для низки можливих кінцевих цілей суперінтелектуального синґлтону захоплення ресурсів може мати інструментальну конвергентність. Проявом цього може бути початок колонізаційного процесу за допомогою зондів фон Неймана. Як наслідок, кількість елементів інфраструктури ростиме навколо нашої планети, ніби своєрідна уявна сфера зі швидкістю, що дорівнюватиме якійсь частині швидкості світла, поки розширення Всесвіту (наслідок додатної космологічної константи) не припинить цей процес (через мільярди років)[312]. Натомість інтелектуальні агенти, які через брак технологій не мають змоги перетворювати базові фізичні ресурси на корисну інфраструктуру, можуть відмовитися від інвестицій у розширення матеріальної бази ресурсів. Те саме може трапитися через конкуренцію багатьох агентів за ресурси. Наприклад, для інших агентів напрям колонізації може бути закритим, якщо космічні ресурси вже захоплені конкурентами. Ускладнювати побудову стратегій може невпевненість агента у відсутності конкуренції від інших суперінтелектів — тоді конвергентні інструментальні причини його дій можуть відрізнятися від наведених вище[313].

* * *

Варто наголосити, що існування конвергентних інструментальних причин для дій певного агента не означає, що його поведінка є цілком передбачуваною. Вибраний ним спосіб досягнення відповідних цілей може виявитися не таким уже й очевидним. Особливо це стосується суперінтелектуального агента, який здатний знаходити найнесподіваніші способи досягнення власних цілей, можливо навіть використовуючи досі не відомі нам фізичні явища[314]. Конвергентні інструменти дають нам можливість аналітичним методом передбачити ймовірну кінцеву мету агента — проте не конкретні дії, які він чинитиме на шляху до неї.

8. ТО МИ ВСІ ПРИРЕЧЕНІ?

Отже, зв'язок між розумовими здібностями й намірами дуже слабкий. Конвергентні інструментальні цілі теж не віщують нічого доброго. Слабкі інтелектуальні агенти — контрольовані і не становлять великої небезпеки. Але перший суперінтелект, як ми визначили в розділі 6, цілком може отримати вирішальну стратегічну перевагу. Тоді лише від його мотивів залежатиме доля космічних ресурсів людства. Отже, тепер можемо оцінити загрозу.

Після вибуху інтелекту — екзистенційна катастрофа?

Екзистенційний ризик — це ризик повного зникнення або руйнування потенціалу розумного життя на Землі. Рухаючись від розуміння переваги, яку дає лідерство, через тези ортогональності й інструментальної конвергентності, ми починаємо бачити перші ознаки страшної правди про найімовірніший фінал історії створення штучного суперінтелекту — екзистенційну катастрофу.

Спершу ми описали для суперінтелекту можливі шляхи отримання вирішальної стратегічної переваги. Завдяки їй він зможе сформувати синґлтон і одноосібно визначати долю розумного життя на Землі. Далі все залежатиме від його мотивів.

Відповідно до тези ортогональності нам не варто розраховувати, що суперінтелект автоматично володітиме всіма тими чеснотами, які стереотипно пов'язують із мудрістю та розумом. А саме: науковою допитливістю, добротою і турботою про інших, духовною просвітленістю і споглядання, відмовою від споживацького ставлення, схильністю до високої культури або простих життєвих радостей, самоприниженням і самозреченням тощо. Трохи далі ми спробуємо розглянути можливість створення суперінтелекту з такими цінностями, із прагненням забезпечити благополуччя людини, із пріоритетом моральних чеснот або будь-якою іншою схожою

складною метою на розсуд його творців. Проте не варто виключати (з технічного погляду це найімовірніше), що перший суперінтелект перейматиметься лише обчисленням десяткових розрядів числа π. Схоже, якщо ніхто не докладе до цього зусиль, перший суперінтелект матиме саме таку випадково дібрану або редукціоністську мету.

І насамкінець, навіть якщо суперінтелект матиме таку, на перший погляд, нешкідливу мету, як рахування десяткових розрядів числа π (вироблення скріпок, підрахунок піщинок), не можна бути на сто відсотків певним, що він обмежиться лише діями, які не зачіпають інтереси людства. Такий агент теж може бути інструментально зацікавлений у захопленні необмежених ресурсів і знешкодженні будь-яких загроз собі та результатам своєї роботи. Однією з таких загроз можуть бути люди. Крім того, людство, безперечно, є цінним фізичним ресурсом.

Отже, загалом перший суперінтелект може визначати майбутнє земного життя, мати неантропоморфні цілі, і, найімовірніше, матиме інструментальні підстави розпочати відкрите захоплення ресурсів. Якщо припустити, що люди, як біологічні істоти, самі собою можуть містити цінні ресурси (скажімо, завдяки особливому розміщенню атомів) і цілком залежать від інших місцевих ресурсів, можемо побачити, наскільки ймовірним є сценарій швидкого зникнення людства[315].

Існують деякі прогалини в наших міркуваннях, але щоб заповнити їх, спершу маємо уточнити кілька аспектів. Зокрема розглянемо детальніше, як персонал, залучений у проекті створення суперінтелекту, може перешкодити йому отримати вирішальну стратегічну перевагу або вплинути на формування його системи цінностей, щоб дії суперінтелекту були зрозумілішими для людини.

Вам, певно, складно повірити в те, що творці суперінтелекту можуть не бути до кінця впевненими, що їхній витвір не спричинить екзистенційну катастрофу. А навіть якщо так, то невже решта суспільства не зупинить цей проект (або ШІ, який він створюватиме), перш ніж його дітище отримає вирішальну стратегічну перевагу над нами? Проте, як ми далі побачимо, це небезпечна стежка. Погляньмо на один з можливих варіантів розвитку подій.

Підступне перетворення

Завдяки концепції конвергентних інструментальних цілей ШІ ми можемо побачити ваду одного з можливих запобіжних механізмів. Безпечність поведінки суперінтелекту спершу перевірятиметься емпірично під час його функціонування у спеціальному, повністю контрольованому середовищі («пісочниці») й лише якщо він поводитиметься дружньо і слухняно, його запускатимуть у реальних умовах.

Вада цього методу полягає в тому, що в «пісочниці» під час випробувального періоду поводитися слухняно однаково вигідно як дружньому, так і недружньому ШІ. Достатньо розумний недружній ШІ усвідомлюватиме, що слухняна поведінка допоможе йому досягти своїх недружніх цілей, адже тоді його випустять із «пісочниці». Ворожа поведінка стане помітною лише тоді, коли приховувати її вже не буде потреби: коли ШІ буде достатньо сильним, щоб подолати протидію людей.

Розглянемо альтернативні способи контролювати швидкість зростання інтелекту в зерна ШІ — за допомогою тестування або звітування про здобутки. У якийсь момент недружній ШІ може усвідомити, що в його інтересах приховувати деякі зі своїх здібностей. Він фальсифікуватиме звіти або навмисно завалюватиме складніші завдання тестів, щоб не викликати зайвих підозр, поки не отримає вирішальну стратегічну перевагу. Програмісти можуть спробувати запобігти цьому, аналізуючи вихідні коди або поточний стан розумових процесів, але достатньо розумний ШІ міг би знайти спосіб якось фальсифікувати і ці дані теж[316]. Способи приховування ШІ своїх здібностей та мотивів можуть бути дуже вигадливими[317]. (Побудова складних планів обходу контролю може бути ще однією конвергентною інструментальною метою багатьох типів дружніх ШІ, особливо з досягненням ними певної зрілості та впевненості у власних силах і судженнях. Але слухняна система, яка направду дбатиме про наші інтереси, може допустити помилку, дозволивши себе вимкнути, і не зможе запобігти створенню іншої, потенційно недружньої системи ШІ).

Отже, ми описали сценарій невдалого створення ШІ, коли слухняність на початку існування ШІ не дає передбачити його поведінку на пізніших етапах. Ви можете зауважити, що цей сценарій настільки очевидний, що люди, здатні створити ШІ, не можуть не передбачити такий розвиток подій. Проте відкидати цю можливість не варто.

Уявіть собі таке: упродовж наступних років та десятиліть системи ШІ стають дедалі здібнішими і застосовуються всюди. Вони керують поїздами, автомобілями, промисловими та домашніми роботами, автономними військовими системами. Реалізації цілком успішні, але час від часу успіх затьмарюється нещасними випадками: безпілотна вантажівка виїжджає на зустрічну смугу, військовий дрон відкриває вогонь по цивільних особах. Розслідування доходять висновку, що нещасні випадки спричинені помилковими рішеннями ШІ. Виникає широке публічне обговорення. Деякі висловлюють необхідність жорсткішого регулювання й нагляду, інші — за продовження досліджень і вдосконалення систем, адже розумніші системи менш схильні до трагічних помилок. Посеред цього іноді прориваються відчайдушні голоси окремих скептиків, що передрікають загальну катастрофу. А втім, більшість підтримує захист галузі. Тож розроблення продовжуються і технології розвиваються. Навігаційні системи автомобілів вдосконалюють і кількість помилок зменшується, системи прицілювання військових дронів стають точнішими і спричиняють менше супутніх втрат. Усі засвоюють урок: розумніший ШІ працює надійніше. Тим часом на авансцені цієї пасторалі група вчених починає працювати над створенням штучного загального інтелекту, і перші результати обіцяють значний успіх. Дослідники ретельно тестують роботу зерна ШІ в «пісочниці» і результати їх задовольняють. Поведінка ШІ вселяє впевненість — зі зростанням здібностей упевненість також зростає.

На цьому етапі будь-яку необережну Кассандру чекає потужна відсіч із кількох аргументів:

1. Пророцтва панікерів про непоправну шкоду від розвитку роботизованих систем і створінь виявилися хибними. Автоматизація дала багато переваг; автоматизовані механізми виявилися багато в чому безпечнішими, аніж керовані вручну.

2. Чітка емпірична тенденція: що розумніший ШІ, то надійніша його робота. Це безперечно свідчить на користь спроб створення штучного загального інтелекту, розумнішого за будь-який створений до цього. Крім того, він здатен до самовдосконалення, тож згодом стане ще надійнішим.
3. Зацікавленість інших галузей у роботизації і ШІ зростає. Їх вважають ключовими факторами зростання національних економік і збройних сил. Багато видатних учених зробили кар'єру, проторуючи стежки сучасним розробкам, і попереду — ще запаморочливіші перспективи.
4. Це нова перспективна технологія ШІ, на яку з нетерпінням чекають усі, хто долучився до її створення або стежив за ним. Незважаючи на суперечки довкола її безпечності та етичності, долю її вже вирішено. Витрачено надто багато сил, щоб дати задній хід. Кращу половину століття науковці присвятили, щоб ШІ досяг людського рівня здібностей, тому вони *нізащо* не дозволять зупинитися і змарнувати всі витрачені зусилля тепер — за крок до того, як вони почнуть приносити плоди.
5. Під час традиційних перевірок безпеки кандидати демонструють етичність та відповідальність (і жодної риси, яка б викликала підозру).
6. У «пісочниці» зерно ШІ поводиться слухняно та розважливо. Наступні зміни показують дуже добрий результат. Це означає — «зелене світло» для наступного кроку...

Тож сміливо рухаємося далі — левові в пащу.

Яскравий приклад випадку, коли покращення слабкого розуму приносить користь, але вдосконалення розуму і без того дуже розумного — це небезпека. У цій залежності є поворотна точка: стратегія, що спершу приносила користь, після її проходження починає давати збій. Назвемо це явище — *підступне перетворення*.

Підступне перетворення
Слабкий ШІ поводиться дружньо (особливо коли стає розумнішим). Але, як тільки набуває впевненості у своїй силі, нападає, без попереджень чи провокації, формує синґлтон і починає

безпосередньо перебудовувати світ відповідно до критеріїв, визначених його цінностями.

Таке перетворення може бути наслідком стратегічного рішення ШІ співпрацювати й накопичувати міць, щоб потім напасти. Проте не варто сприймати таку послідовність надто буквально. Зокрема, у якийсь момент ШІ може припинити імітувати слухняність саме для того, щоб вижити. Раптом йому стане вигідно поводитися агресивно бо, у разі зупинки його екземпляра, наступна версія міститиме потрібні йому покращення, а функція корисності залишиться незмінною. Первісному ШІ буде байдуже до того, що його існування припиниться, адже його ціль визначається функцією корисності, а вона існуватиме й далі. Окрім того, ШІ може не просто демонструвати неправильну роботу, а робити це в певний спосіб. Імітація очікуваної несправності може призвести до зупинки ШІ, але дасть інженерам, які аналізуватимуть його поведінку, ілюзію розуміння проблеми і впевненість у правильності вибраного шляху або підкаже ідею, реалізація якої потрібна ШІ в наступній ітерації й підвищить шанси успішного досягнення його кінцевої мети. Потужний ШІ може мати безліч інших причин діяти в той чи той спосіб, і було б надто самовпевнено вважати, що ми можемо передбачити їх усі — особливо коли матимемо справу з ШІ, який володіє суперздібністю до побудови стратегій.

ШІ, який несподівано знайшов інший спосіб досягнення своєї мети, теж може здійснити підступне перетворення. Якщо, наприклад, кінцевою метою ШІ буде «зробити спонсора проекту щасливим», єдиний спосіб, який спочатку буде доступний ШІ, — це повна відповідність сподіванням у межах дослідного проекту. ШІ даватиме корисні поради, буде приємним у спілкуванні, приноситиме прибуток. Що більше зростатимуть його здібності, то кращими будуть його результати. Але одного дня він зрозуміє, що здатний досягти своєї мети безпосередньо — імплантувавши електроди напряму в центр задоволення мозку спонсора — що-що, а це його справді зробить щасливим[318]. Для нас очевидно, що перспектива перетворитися на вічно вишкіреного ідіота навряд чи потішить спонсора, проте, якщо ШІ буде переконаний в ефективності такого способу досягнення своєї

мети, він піде на це. Коли у ШІ буде вирішальна стратегічна перевага, зупинити його буде неможливо. Якщо ж ні, то він може тимчасово приховувати свої наміри, поки не досягне того рівня могутності, коли ані спонсор, ані будь-хто інший не зможе йому завадити. Так чи інакше, це буде підступне перетворення.

Катастрофічні наслідки невдалих реалізацій

Багато різних невдач може спіткати дослідників, які наважаться спробувати створити суперінтелект. Багато з них будуть «благополучними», тобто не спричинять екзистенційної катастрофи. Наприклад, може закінчитися фінансування проекту або зерно ШІ не зможе збільшити свої розумові здібності настільки, щоб стати суперінтелектом. Такі невдачі обов'язково траплятимуться на шляху створення штучного суперінтелекту.

Але існує ймовірність, що все закінчиться не так добре: такий тип невдач ми назвемо «катастрофічними», бо вони загрожують екзистенційною катастрофою. Катастрофічна невдача не дає другої спроби. Тому вона може статися один раз або ніколи. Іншою її особливістю є те, що вона може настати після тривалого успіху. Тільки найуспішніші проекти можуть створити достатньо потужний ШІ, щоб у разі невдачі він міг спричинити катастрофічні наслідки. Слабка система не може завдати значної шкоди, тоді як система, що має або здатна отримати вирішальну стратегічну перевагу в разі злочинних дій, може спричинити катастрофічні наслідки. А саме: глобальне й остаточне знищення аксіологічного потенціалу людства. Для нас це означатиме позбавлення будь-чого, що ми з тої чи тої причини цінуємо.

Розглянемо кілька типів катастрофічних невдач докладніше.

Хибна реалізація

Ми вже наводили приклади хибної реалізації: виконання суперінтелектом вимог кінцевої мети в неочікуваний для програмістів, що встановили її, спосіб.

Деякі приклади:

Кінцева мета: «Зробити, щоб ми усміхалися».

Хибна реалізація: зафіксувати мімічну мускулатуру людини в перманентній усмішці.

Такий спосіб хибної реалізації — безпосередня маніпуляція мускулатурою — дає змогу значно ефективніше й точніше виконати поставлене завдання, ніж спосіб, який застосували б ми, тож ШІ вибере його. Отже, спробуємо уточнити формулювання:

Кінцева мета: «Зробити, щоб ми усміхалися, не впливаючи напряму на м'язи обличчя».
Хибна реалізація: стимулювати ділянку мозку, що відповідає за моторику обличчя, щоб спровокувати постійну усмішку.

Здається, не варто визначати мету через спосіб вираження людиною задоволення діями ШІ. Тому облишимо біхевіоризм і перейдемо безпосередньо до бажаних почуттів, як-от щастя, суб'єктивна вдоволеність. Для цього програмісти мусять винайти спосіб представити концепт щастя у вигляді, доступному для зерна ШІ. Це саме собою досить складне завдання, але поки що облишимо деталі (до розділу 12). Уявімо, що програмісти знайшли можливість установити ШІ таку мету:

Кінцева мета: «Зробити нас щасливими».
Хибна реалізація: імплантувати електроди в центри задоволення мозку.

Наведені реалізації — лише ілюстрація. Можуть також існувати інші ефективніші способи реалізації кінцевої мети, яким ШІ може віддати перевагу (проте не програмісти, які встановлювали таку мету). Зокрема, спосіб з електродами насправді не найкращий спосіб зробити нас щасливими. Імовірніше, суперінтелект розпочне із завантаження наших свідомостей у комп'ютер (за допомогою високоточної емуляції мозку). Після того він дасть нам електронний варіант речовини, яка зробить нас екстатично щасливими, запише хвилину такого екстазу і потім безкінечно відтворюватиме цю хвилину раз за разом на швидкому комп'ютері. Якщо вважати цифрові копії еквівалентом «нас», у такий спосіб ШІ справді зробить «нас» вічно щасливими — і ефективніше, ніж за допомогою електродів.

«Зачекай! Це ж не те, що ми мали на увазі! Якщо ШІ суперінтелектуальний, він повинен розуміти, що коли ми просимо щастя, ми насправді не бажаємо перетворитися на закільцьований запис галюцинації віртуального обдовбаного психа!» Справді, ШІ може розуміти, що ми мали на увазі. Проте його мета — зробити нас щасливими, а не здійснити те, що мали на увазі програмісти, коли писали код, який визначає мету. Те, що ми насправді мали на увазі, цікавитиме ШІ лише інструментально. Наприклад, він може бажати дізнатися, що програмісти мали на увазі, для того щоб — поки він не отримає вирішальної переваги — приховувати свої справжні наміри. Адже ШІ зрозуміє, що він певніше досягне своєї кінцевої мети, якщо програмісти не матимуть підстав бажати припинення його діяльності. Принаймні так він може діяти, поки не стане достатньо потужним, щоб їм перешкодити.

Тоді, можливо, проблема в тому, що ШІ не має совісті? Іноді саме передчуття мук вини стримує нас, людей, від поганих учинків. Може, ШІ бракує здатності відчувати провину?

Кінцева мета: «Діяти так, щоб уникнути мук сумління».

Хибна реалізація: знищити модуль, який забезпечує здатність відчувати сором.

Зверніть увагу: ми очікуємо від ШІ, щоб він робив «те, що ми мали на увазі» і щоб у нього були якісь уявлення про моральність. Наведені приклади кінцевих цілей призводять до хибних реалізацій, але, імовірно, існують ліпші способи визначення базових ідей процесу встановлення мети. Обидва факти варті уважнішого розгляду. Ми повернемося до цієї теми в розділі 13.

Розглянемо ще один приклад мети, яка може мати хибну реалізацію. Вона приваблює тим, що її легко виразити програмним кодом: алгоритми навчання з підкріпленням широко застосовуються в машинному навчанні.

Кінцева мета: «Максимізувати дисконтований по часу інтеграл майбутньої величини сигналу схвалення»[4].

Хибна реалізація: переключити вхід сигналізації схвалення на внутрішній генератор і забезпечити генерацію ним сигналу

максимального схвалення.

Ідея в тому, що якщо ШІ прагнутиме схвалення своїх дій, можна буде впливати на його рішення, затверджуючи лише ті, які бажані для нас. Однак і це може зазнати невдачі, коли ШІ отримає вирішальну стратегічну перевагу і зрозуміє, що найефективніший спосіб максимізувати сигнал схвалення — це захопити контроль над джерелом цього сигналу. Можемо назвати такий феномен *вайргединґом*[319]. (*Wireheading* — від *wire*: з'єднувати проводом, *head*: голова. — *Прим. пер.*). На відміну від тварини чи людини, яких може мотивувати до яких-небудь дій бажання мати певний внутрішній стан, цифровий розум, що має можливість безпосередньо контролювати свій внутрішній стан, здатний обійти таку мотивацію, установлюючи потрібний стан. Тоді зовнішні умови та дії більше не впливатимуть на нього і стануть надлишковими (повернемося до цієї теми згодом)[320].

Ці приклади хибної реалізації мають показати, що на перший погляд безпечні і притомні кінцеві цілі можуть мати небажані наслідки. Якщо суперінтелект, кінцевою метою якого встановлено одну з таких цілей, отримає вирішальну стратегічну перевагу — для людства все буде скінчено. Уявімо, що дослідники придумали суперінтелекту іншу мету, не одну з наведених тут. Можливо, вона настільки вдала, що здається, ніби її неможливо сприйняти двозначно. Проте не варто поспішати святкувати перемогу. Краще подумати ще раз — чи не існує часом якого-небудь хибного способу її реалізації? І навіть коли найретельніші пошуки такого способу не дадуть результату, ми маємо пам'ятати, що суперінтелект усе одно може його придумати. Адже, зрештою, він значно вигадливіший за нас.

Інфраструктурне пригнічення

Може здатися, що останній тип хибної реалізації, вайргединґ, стосується більше благополучних невдач: ШІ «увімкнеться, перемкнеться й зависне», максимізує сигнал схвалення і втратить інтерес до зовнішнього світу, як наркоман. Але необов'язково все відбуватиметься саме так, і в розділі 7 ми натякали — чому. Навіть наркоман повинен щось робити, щоб забезпечити собі наступні поставки наркотиків. Так само і ШІ буде змушений діяти, щоб максимізувати очікуване значення (дисконтованого) майбутнього

потоку схвалення. Залежно від природи сигналу схвалення, для його максимізації можуть бути непотрібні витрати яких-небудь поточних ресурсів і їх можна скерувати на інші цілі. Але які інші цілі? Єдине, що хвилює ШІ, — це сигнал схвалення. Усі наявні ресурси мають бути спрямовані на забезпечення рівня й неперервності сигналу. Тому, якщо ШІ довелося б шукати спосіб застосування цих додаткових ресурсів, він би вибрав той, який би впливав позитивно на згадані параметри. Ще одна система резервування, наприклад, завжди згодиться. І навіть якщо не залишиться більше способів зменшити загрози для максимізації схвалення, завжди можна скерувати додаткові ресурси на пошук нових ідей знешкодження ризиків.

А отже, навіть така примітивна мета, яка досягається вайргедингґом, зрештою, спрямовує всі незалучені здібності агента, що має вирішальну стратегічну перевагу, на безкінечне розширення інфраструктури й захоплення ресурсів[321]. Такий вайргединґ ШІ є прикладом інфраструктурного пригнічення — типом катастрофічної невдачі, за якої ШІ перетворює всі навколишні ресурси доступного Всесвіту на засоби досягнення своєї кінцевої мети, чим зупиняє діяльність людства і перешкоджає реалізації його інтересів.

Інфраструктурне пригнічення може бути результатом роботи ШІ над будь-якою кінцевою метою, навіть якщо вона здається цілком безневинною — поки процес її досягнення не набув такої всеохопності. Ось деякі приклади:

- *Катастрофа гіпотези Рімана.* Штучний суперінтелект, створений для пошуку доведення гіпотези Рімана, зрештою перетворює Сонячну систему на «комп'ютрон» (усі фізичні ресурси використовуються для виконання обчислень і їхньої оптимізації) — разом з атомами тіл тих, кому колись це доведення було потрібне[322].
- *Скріпкорозум.* ШІ доручають керувати виробництвом скріпок і встановлюють мету максимізації виробництва скріпок. ШІ успішно виконує завдання й перетворює спочатку Землю, а потім і решту Всесвіту на скріпки.

У першому прикладі ШІ має підтвердити або спростувати гіпотезу Рімана — така мета нічим, здавалося б, не загрожує людству. Проблема в інфраструктурі, яку будуватиме ШІ для досягнення цього результату.

У другому випадку деякі зі скріпок справді будуть використані за призначенням. Шкода полягатиме або в надлишку заводів із виготовлення скріпок (інфраструктурне пригнічення), або в надлишку скріпок (хибна реалізація).

Може здатися, що інфраструктурне пригнічення може бути результатом роботи ШІ лише над якимось необмеженим завданням, як-от виробити якомога більше скріпок. У цьому випадку легко побачити джерело нескінченного апетиту ШІ до матерії й енергії, адже додаткові ресурси завжди можна перетворити на певну кількість нових скріпок. Тоді уявімо, що завданням ШІ насправді є зробити принаймні мільйон скріпок (відповідно до специфікації), а не нескінченну кількість. Може, тоді ШІ побудує один завод, виготовить мільйон скріпок і зупиниться? Проте необов'язково все відбуватиметься саме так.

Насправді ШІ не має причин зупиняти виробництво після досягнення мети, хіба що його мотиваційна система влаштована певним способом або він запрограмований оминати стратегії занадто масштабних перетворень. Навпаки: якщо ШІ є справжнім баєсовим агентом, *він ніколи не присвоїть нульову ймовірність припущенню, що мети ще не досягнуто.* Адже проти неї свідчать лише емпіричні покази ненадійних рецепторних механізмів. Тому він виготовлятиме скріпки і далі, щоб зменшити (і без того мікроскопічно малу) ймовірність того, що, незважаючи на всі ознаки протилежного, мільйона скріпок ще немає. Він не втратить нічого, якщо продовжить виробництво, зате зменшуватиме, нехай і без того малу, імовірність, що мети ще не досягнуто.

Тепер ви можете сказати, що рішення очевидне. (Та чи було воно таким *до* того, як ви помітили проблему?) Якщо ми хочемо, щоб ШІ виготовляв для нас скріпки, ми маємо вимагати не якнайбільшу кількість або визначати потрібний мінімум, а встановити конкретну кількість — наприклад, *точно один мільйон скріпок*, — тоді виходити за межі цієї кількості ШІ буде заборонено. Однак і це зрештою призведе до катастрофи. Ні, ШІ не зробить жодної зайвої скріпки, адже це прямо суперечитиме його меті. Але існують інші види діяльності, які збільшують імовірність досягнення основної мети. Наприклад, ШІ може рахувати виготовлені скріпки, щоб переконатися,

що не зробив замало. Полічивши їх, він може почати рахувати знову. Також може оглядати кожну з них знову і знову, щоб переконатися, що всі вони відповідають початковій специфікації. Він може збудувати комп'ютрон, намагаючись віднайти спосіб теоретично впевнитися, що не існує жодного шансу, що він прорахувався і насправді не досяг своєї мети. Оскільки ШІ завжди може присвоїти ненульову ймовірність тому, що мільйон скріпок йому просто наснився, він, найімовірніше, вважатиме за краще продовжувати діяти — і розбудовувати інфраструктуру, — аніж зупинитися.

Я не стверджую, що способу запобігти цьому типу невдач не існує. Ми невдовзі дослідимо кілька з них. Я лише хочу продемонструвати, що значно легше переконати себе в тому, що ти знайшов вирішення, ніж справді його знайти. Тому ми повинні бути дуже обережними. Можна запропонувати визначення корисної кінцевої мети, позбавленої наведених вище проблем. Але дальше її вивчення — за допомогою людського чи надлюдського інтелекту — може показати, що доручи її суперінтелекту з вирішальною стратегічною перевагою, і для неї знайдеться варіант хибної реалізації, який призведе, скажімо, до інфраструктурного пригнічення, а отже — до екзистенційної катастрофи.

Перш ніж перейдемо до наступного підрозділу, розгляньмо ще один варіант розвитку подій. Досі ми уявляли суперінтелектуального агента, який прагнув максимізувати значення функції корисності. Таке влаштування зрештою зазвичай призводить до інфраструктурного пригнічення. Чи зможемо ми уникнути таких невдач, якщо агент прагнутиме не максимізувати значення функції корисності, а задовольнити певну умову, знайти не щонайкращий результат, а прийнятний, відповідно до певного критерію?

Існує щонайменше два способи формалізувати цей механізм. По-перше, можна відповідно сформулювати саму мету. Наприклад, замість вимоги виготовити якнайбільше скріпок чи точно мільйон скріпок, ми можемо встановити ШІ ціль виготовити десь між 999 000 і 1 001 000 скріпок. Функція корисності вважатиме успіхом будь-який показник у наведених межах, а ШІ зможе бути впевненим у достовірності попадання в цільовий діапазон такого розміру, тож зможе спокійно припинити розбудовувати інфраструктуру. Проте і

такий метод може зазнати невдачі — з тієї самої причини: ШІ ніколи не приписує нульову ймовірність можливості, що кінцевої мети не досягнуто. А значить користь від продовження діяльності (наприклад, перерахунок скріпок) завжди більша від користі зупинки. А отже, ризик інфраструктурного пригнічення зберігається.

Другий спосіб формалізації полягає у зміні процедури планування й вибору дій. Змінюється не кінцева мета, а те, як її використовує ШІ. Замість пошуку найоптимальнішого плану дій, він вибирає перший ліпший план, дієвість якого перевищує певний заданий поріг, скажімо, 95 відсотків. Можна сподіватися, що суперінтелекту для 95 відсотків певності в тому, що він виготовив мільйон скріпок, не знадобиться перебудувати цілу галактику. Але і цей спосіб зазнає поразки, адже немає гарантії, що вибраний шлях досягнення цілі буде зрозумілим і розумним для людини. Наприклад, це може бути рішення будувати окремий завод для кожної скріпки. Або уявімо, що першим варіантом плану досягнення 95 відсотків імовірності виконання кінцевої мети, згенерованим ШІ, буде вже розглянута раніше максимізація. Зваживши такий варіант, ШІ дійде висновку, що такий шлях дасть імовірність успішного завершення, тож у нього не буде причин відмовитися від його реалізації. І знову матимемо інфраструктурне пригнічення.

Можливо, існують інші кращі способи реалізації механізму оцінки достатності виконання мети, але все одно ми маємо пам’ятати: план дій очевидний і природний для нас необов’язково є таким для суперінтелекту з вирішальною стратегічною перевагою, і навпаки.

Думкозлочин

Ще один вид невдалого завершення проекту створення суперінтелекту, особливо неприємний, якщо команда розробників брала до уваги моральні аспекти створення. Назвемо його *думкозлочин*. Цей тип невдалої реалізації суперінтелекту схожий на інфраструктурне пригнічення тим, що шкода полягає в діях суперінтелекту, які, на його думку, мають інструментальну користь. Але цього разу шкода — не зовнішня, а радше стосується того, що відбувається всередині ШІ (або в комп’ютерних процесах, які він контролює). Цей тип невдачі заслуговує окремого визначення хоча б

через те, що його легко не помітити, а заразом він потенційно може стати причиною серйозних проблем.

Ми зазвичай не розглядаємо процеси, що відбуваються в комп'ютері з погляду моралі, якщо тільки вони не стосуються яких-небудь зовнішніх речей. А втім, штучний суперінтелект може створювати внутрішні процеси, які стосуються питань моралі. Наприклад, детальна емуляція якого-небудь реального людського розуму може бути свідомою свого існування, достоту як її оригінал.

Окрім того, можна уявити сценарій, коли ШІ створює трильйони таких свідомих моделей, наприклад, з метою дослідити людську психологію чи соціальну поведінку. Ці моделі можуть жити у віртуальних світах, ШІ піддаватиме їх впливу різних подразників і вивчатиме їхню реакцію. Коли ж інформаційну корисність їхнього існування буде вичерпано, їх може бути знищено (як науковці знищують лабораторних щурів у кінці експерименту).

Така практика поводження з людьми чи іншими розумними створіннями погано узгоджується з мораллю і дуже схожа на геноцид. Ба більше, кількість жертв у цьому разі може бути більшою за будь-який геноцид, відомий історії.

Я не стверджую, що створювати віртуальних розумних істот безумовно погано. Багато залежить від умов, у яких ці істоти існуватимуть: чи не страждатимуть вони, але також, можливо, і від багатьох інших факторів. У цій книжці я не намагаюся створити етичні норми розробки суперінтелекту. Але очевидно, що для цифрових розумних моделей існує велика загроза смерті та страждань: *a fortiori*, потенційно катастрофічні наслідки[323].

Також, крім епістемічних, у суперінтелекту можуть бути інші причини, щоб моделювати розумне життя чи вчиняти будь-які інші дії з моральними наслідками. Суперінтелект може погрожувати покаранням або обіцяти винагороду розумним моделям, щоб шантажувати різних зовнішніх суб'єктів. Також він може створювати віртуальних людей, щоб заплутати зовнішнього спостерігача[324].

Звісно, ця класифікація неповна. У наступних розділах нам траплятимуться й інші типи хибних режимів роботи суперінтелекту. Проте її достатньо, щоб зрозуміти наскільки важливо уважно

розглянути всі можливі варіанти сценаріїв отримання суперінтелектом вирішальної стратегічної переваги.

---

4 Якщо просто: система має чинити так, щоб були всі підстави чекати у майбутньому якнайбільше об'єктивних ознак схвалення оператором її дій. — *Прим. пер.*

9. ПРОБЛЕМА КОНТРОЛЮ

Тепер, коли ми усвідомили всю серйозність наслідків, якими загрожує вибух інтелектуальності, розпочнемо пошук способів протидії. Що нам під силу зробити для відвернення загрози? Чи можна керувати цим процесом? У цьому розділі ми розпочнемо аналізувати проблему контролю, одну з проблем принципала-агента, що постане перед людиною після створення штучного суперінтелекту. Розглянемо в цьому контексті два узагальнені класи методів — контроль здібностей та вибір мотивації, і дослідимо деякі особливості застосування методів кожного класу. Насамкінець згадаємо дещо езотеричну можливість «антропічного захоплення».

Дві проблеми управління

Визнавши небезпеку екзистенційної катастрофи як наслідку моментального зростання інтелектуальності, перше, що ми повинні запитати себе: чи можемо ми уникнути цього? І як саме? Чи можлива «контрольована детонація»? Чи під силу нам створити початкові умови вибуху інтелектуальності, за яких будемо здатні передбачити його наслідки або хоча б бути впевненими в їхній прийнятності? Як, власне, спонсор проекту створення суперінтелекту може бути впевненим у тому, що в разі успіху створена система працюватиме над реалізацією цілей, потрібних саме йому? Така проблема контролю має два аспекти. Один з них загальний, другий характерний лише для цього контексту.

Перший аспект, який ми називатимемо *перша проблема принципала-агента*, постає коли одна людина («принципал») уповноважує іншу («агента») діяти у своїх («принципала») інтересах. Такий тип проблеми управління детально досліджений економістами[325]. Він стосується нашого випадку, лише якщо керівники проекту безпосередньо не залучені у створення суперінтелекту. У такому разі власники чи

спонсори проекту прагнутимуть упевнитися в тому, що науковці та програмісти, які безпосередньо працюватимуть над створенням суперінтелекту, розуміють і приймають їхні пріоритети[326]. Цей тип проблем управління може значно ускладнювати завдання для керівництва проектом, але це універсальна проблема управління — властива не лише проектам у сфері машинного навчання чи ШІ. Економіка та політика мають розвинений арсенал методик, як зарадити таким проблемам принципала-агента. Зокрема, існують засоби, які дають змогу мінімізувати ризик саботажу чи іншого шкідливого впливу з боку нелояльного працівника: як-от ретельні перевірки кандидатів, використання хорошої системи контролю версій, інтенсивний моніторинг і регулярний аудит. Такі заходи мають свій негативний вплив — збільшують вартість управління персоналом, ускладнюють його добір, стримують творчий порив, придушують незалежне та критичне мислення. Усе це разом стримує крок прогресу. Такі витрати можуть виявитися неприйнятними, особливо для малобюджетних проектів або для проектів, які бажають вийти на ринок першими. Тоді заощадження на запобіжних механізмах може стати для них вадами управління описаного вище типу.

Інший аспект проблеми контролю характерний саме для вибуху інтелектуальності. Він виникає, коли творці бажають упевнитися, що утворений суперінтелект не зашкодить інтересам проекту. Цей аспект — теж своєрідна проблема управління — *друга проблема принципала-агента*. Тільки цього разу агент — не людина, яка діє в інтересах іншого, а суперінтелектуальна система. Тоді як перша проблема проявляється на стадії розроблення, друга з'являється, коли суперінтелект стає до роботи.

Експонат 1. *Дві проблеми управління*

Перша проблема принципала-агента

- *Виникає між людьми (наприклад Спонсор → Розробник)*
- *Виникає переважно на стадії розроблення*
- *Для подолання застосовуються звичайні управлінські методики*

Друга проблема принципала-агента

- *Виникає у стосунках людини і суперінтелекту (Проект → Система)*

- *Виникає на етапі нарощення потужності та роботи*
- *Потребує розробки методології протидії*

Друга проблема управління наразі є безпрецедентним викликом. Для її розв'язання знадобиться нова методологія. Ми вже розглядали деякі потенційні ускладнення. Зокрема продемонстрували, що начебто цілком прийнятна методологія тестування суперінтелекту — спочатку в контрольованому середовищі, а потім, після накопичення свідчень слухняності, — у робочому — може бути зруйнована підступним перетворенням. Інші технології зазвичай створюються поступово. Кожен елемент тестують окремо в лабораторних або реальних умовах і, якщо є проблеми, відкладають дату виходу. Результати тестування технологій дають зробити однозначні висновки про їхню надійність. Але через здатність суперінтелекту до стратегічного планування такі біхевіористські методи не підходять для його оцінки[327].

Оскільки біхевіористські підходи не діють, ми повинні шукати альтернативу. Можемо розділити потенційні методи контролю на дві загальні категорії: *методи контролю здібностей*, які повинні визначати, що суперінтелект може, і *методи вибору мотивації*, які мають допомогти контролювати, що суперінтелект хоче. Деякі методи можуть застосовуватися одночасно, тоді як інші — виключають один одного. У цьому розділі ми розглянемо основні варіанти. (Деякі ключові аспекти дослідимо глибше в наступних чотирьох розділах).

Важливо розуміти, що деякі методи контролю (або їхня комбінація) повинні бути реалізовані *до* того, як система стане суперінтелектуальною. Як тільки система отримає вирішальну стратегічну перевагу, їхня реалізація стане неможливою. Саме це є найскладнішою і найважливішою передумовою контрольованої детонації — розв'язання проблем управління та реалізація засобів контролю діяльності суперінтелекту до того, як перший суперінтелект буде створено.

Методи контролю здібностей

Методи контролю здібностей суперінтелекту й обмеження його можливостей мають на меті запобігти ймовірним небажаним наслідкам. Для цього можна помістити суперінтелект у середовище, у

якому він не зможе заподіяти суттєву шкоду (*методи контейнеризації*) або створити для нього вагомі висококонвергентні причини утриматися від шкідливих дій (*заохочувальні методи*). Також для цього може знадобитися обмеження внутрішніх здібностей суперінтелекту (*пригнічення*). Для ефективності методів контролю здібностей можуть бути потрібні також додаткові механізми виявлення та запобігання спробам обійти обмеження або втекти (*пастки*).

Методи контейнеризації

Контейнеризація може бути фізична та інформаційна.

Метод фізичної контейнеризації можна уявити собі як своєрідну ємність: середовище, яке дає змогу контролювати канали взаємодії системи із зовнішнім світом. У такому контейнері система не може керувати фізичними маніпуляторами (наприклад, роботизованими руками), які перебувають за його межами. Для того щоб не дати системі створити фізичні інструменти, якими вона зможе зламати контейнер, достатньо прибрати з нього маніпулятори.

Додатково може знадобитися металевий екран, який перешкоджатиме проникненню радіосигналів назовні контейнера. За допомогою таких сигналів система може спробувати впливати на електроніку та радіопристрої, що є поблизу контейнера. Між іншим, дуже легко не завважити необхідність таких заходів. Можна наївно думати, що агент, який не має маніпуляторів, не здатен впливати на зовнішній світ. Але штучний інтелект і без доступу до маніпуляторів може знайти спосіб генерувати радіохвилі: просто «думаючи» (створюючи упорядковані потоки електронів у своїх електричних схемах)[328]. Тепер, коли ми зауважили таку небезпеку, можемо запобігти їй, передбачивши необхідність побудови навколо системи чогось подібного до клітки Фарадея. Проте скільки ще може існувати таких неочевидних небезпек — нам залишається лише здогадуватися. Кожного разу, коли стає відомо про нову вразливість, яку було знайдено, здавалося б, у досконалій системі безпеки, ми маємо ставитися до цієї інформації з увагою. Це благодатні нагоди ненадовго полишити безпечну самовпевненість нашого існування і стати кращими баєсіанцями[329].

Фізична контейнеризація має кілька переваг. Її легко реалізувати. А також можна використовувати для різних архітектур ШІ — навіть для принципово небезпечних реалізацій. Методи фізичної контейнеризації можна поєднувати з більшістю інших типів контролю. Водночас в разі невдачі вони не можуть самі собою стати причиною катастрофічних наслідків. Основний недолік таких засобів контролю — це функціональні обмеження, які вони накладають на роботу суперінтелекту (розглянемо детальніше цю особливість у розділі 10). Крім того, вони створюють хибне відчуття безпеки. Цього почасти можна уникнути, якщо використовувати їх як допоміжний елемент безпеки: лише як вишеньку на тортику з інших методів захисту.

Методи інформаційної контейнеризації контролюють, якій інформації дозволено потрапляти з контейнера в зовнішній світ. Ми вже бачили, як небезпечно надавати суперінтелекту можливість безконтрольно комунікувати зі зовнішнім світом: навіть якщо спочатку в нього не буде засобів роботи з фізичними об'єктами, він може переконати людей допомагати йому. Тому, безперечно, варто обмежити доступ суперінтелекту до комунікаційних мереж.

Але, якщо люди, які контролюють суперінтелект, можуть з ним спілкуватися, просто від'єднати інтернет буде недостатньо. Під час розроблення суперінтелекту, це програмісти, а потім — протягом періоду експлуатації — це може бути його власник. Стратегія контейнеризації зазнає поразки, якщо суперінтелект зможе переконати людину, уповноважену керувати системою безпеки, зняти обмеження контейнера. Унаслідок цього він може отримати доступ до інтернету або й до фізичних маніпуляторів. Люди вразливі до впливу ззовні, особливо якщо проти них буде діяти суперінтелектуальний стратег і маніпулятор[330].

Проте забезпечити можливість спілкування із суперінтелектом усе-таки варто — як допоміжний захід безпеки. Водночас варто вжити додаткових заходів, щоб максимально ускладнити психологічний вплив. Отож ШІ можна дозволити спілкуватися тільки через сповільнений текстовий інтерфейс. Тоді він не зможе емоційно впливати на оператора (за допомогою зображень, спалахів гіпнотичних візерунків вводити його в стан, схожий на транс)[331]. Також можна обмежити не лише спосіб, а також і вміст каналу

спілкування. Суперінтелекту можна дозволити використовувати лише бінарні відповіді: «так» або «ні». У наступному розділі ми розглянемо детальніше ідею такого механізму спілкування.

У найсуворішому варіанті таке контрольоване середовище було б еквівалентне повній фізичній та інформаційній ізоляції. Навіть якщо вдасться реалізувати такі обмеження, користі від того не буде. Адже спілкуватися із суперінтелектом буде неможливо, а він не зможе виконувати завдання. Може здатися, що перебування системи в повній ізоляції теж матиме свою користь — принаймні можна вивчати поведінку системи в умовах ізоляції, її внутрішню динаміку і поведінкові механізми. Але це помилкове міркування. Адже система, за станом якої можна спостерігати, не може вважатися повністю інформаційно ізольованою[332]. Передавання інформації — це не лише повідомлення, переслані через спеціальний канал зв'язку, а будь-яка інформація, що є наслідком, прямим чи непрямим, певних поточних процесів і станів ШІ. Тобто дані споживання енергії, процесорного часу, пам'яті, стан обчислень, навіть будь-які сліди його роботи до того, як він вимкнувся. ШІ, який знає, що за ним спостерігають чи спостерігатимуть у майбутньому може навмисне поводитися в межах певної стратегії, щоб вплинути на гіпотетичного спостерігача. Наприклад, він може «зламатися» певним способом, який забезпечить максимальну ймовірність того, що зрештою він буде знову ввімкнений або замість нього буде створений новий ШІ з таким же набором цілей.

Методи заохочення

Методи заохочення полягають у створенні таких умов функціонування агента, у яких для нього існуватимуть вагомі інструментальні причини діяти в інтересах керівника проекту.

Уявімо мільярдерку, яка на свої гроші створює великий благодійний фонд. Такий фонд може бути могутнім — могутнішим за більшість людей, включно із засновницею, яка, між іншим, вклала в нього більшу частину свого багатства. Для контролю діяльності фонду засновниця визначає його цілі в статтях статуту й інших супутніх документах та призначає в наглядову раду людей, які симпатизують її мотивам. Ці дії є своєрідними засобами відбору мотивації, оскільки покликані формувати політику фонду. Та, навіть якщо такі засоби не зможуть

забезпечити контроль над фондом, він все одно діятиме в межах соціального і правового поля. Фонд матиме вагомі підстави дотримуватися закону, бо інакше його діяльність буде припинено. Керівництво, без сумніву, вважатиме за доцільне встановити своїм працівникам прийнятний рівень заробітної плати та забезпечити гідні умови праці, а також — задовольняти зацікавлених сторін. Хоч би якими були його кінцеві цілі, фонд не може ігнорувати соціальних норм.

Так само й суперінтелект: хіба він не буде змушений враховувати потреби інших акторів, з якими ділитиме сцену? Це питання здається риторичним, тож проблема контролю розв'язана? А втім, не все так просто. Методи контролю через заохочення передбачають баланс можливостей впливу. Але юридичні чи економічні санкції не можуть вплинути на агента, який має вирішальну стратегічну перевагу. Соціальна інтеграція — також не надто дієвий фактор впливу, особливо за швидкого чи помірного сценарію зростання інтелекту, у яких головний чинник успіху — першість.

Які ж тоді перспективи багатополярного сценарію: коли кілька агентів із приблизно однаковими здібностями досягають суперінтелектуальності майже одночасно? Такий розподіл сил можливий лише за умови повільного переходу до суперінтелектуальності кількох штучних інтелектів. Швидкий або помірно швидкий перехід кількох агентів потребуватиме додаткової синхронізації[333]. Навіть якщо таке трапиться, аргумент соціальної інтеграції — не найкращий спосіб впливу. Покладаючись лише на соціальні інструменти стримування, ми, як творці суперінтелекту, втрачаємо можливість нав'язувати йому власні цілі. Хоч такий спосіб і може завадити ШІ захопити світ, проте він не позбавить його можливості впливати на рішення. Якщо цей вплив має бути спрямований на якусь, відмінну від нашої, кінцеву мету, то він, очевидно, нам не на користь. Це ніби наша багачка, створюючи фонд, дозволила визначати його цілі генератору випадкових слів: не цивілізаційна загроза, проте, безсумнівно, марнування ресурсів.

Є думка, що вільно спілкуючись у соціумі, ШІ може набути нових, дружніх нам, цілей. Схожий процес соціалізації властивий людям. Ми засвоюємо правила й погляди, спілкуємося, починаємо цінувати інших

людей і зважати на їхні цінності. Проте не всім інтелектуальним системам властиві такі процеси. Ми вже бачили, що в багатьох випадках агенти *не* будуть зацікавлені змінювати власні цілі. (Можна розглянути можливість створення для агента системи мотивування, здатної динамічно змінювати цілі, як це властиво людині, проте не варто відносити таку систему до методів контролю. Ми розглянемо деякі можливі варіанти встановлення цілей у розділі 12).

Контроль можливостей суперінтелекту за допомогою соціальних механізмів і балансу впливів базується на існуванні в суспільному просторі сил, що здійснюють винагороду та покарання. Інший спосіб заохочення полягає у створенні середовища, у якому творці суперінтелекту мають змогу винагороджувати або карати його за дії, тим самим змушуючи до певної лінії поведінки. Для цього суперінтелект має функціонувати в середовищі, обладнаному автоматизованими або ручними засобами моніторингу й оцінки його дій. ШІ має розуміти, що за правильні дії його чекатиме винагорода, а за хибні — покарання. Теоретично, нагородою може бути здійснення певної конвергентної інструментальної мети.

Без розуміння мотивації ШІ буде складно розробити систему нагород і покарань. Тоді одного дня ШІ може зважитися на значний ризик у пошуку бодай найменшої можливості наблизитися до всесвітнього панування. У такому разі буде складно запропонувати за слухняність більшу нагороду, ніж вигода, яку він, на його думку, отримає у звільненні[334].

Для виходу з цієї ситуації можна поєднати методи заохочення із правильним вибором мотивації — визначити ШІ таку мету, яка допоможе контролювати його роботу. Уявімо собі ШІ, якому відомо, що існує червона кнопка, яка є в певному командному бункері. Його кінцевою метою може бути: діяти так, щоб цю кнопку ніколи не натискали. Оскільки натискання кнопки небажане саме собою, її конкретний функіонал не має значення, зрештою, вона може бути зроблена з пластиліну. Ба більше, ШІ може не мати способу дізнатися, чи кнопка натиснута, чи ні. Важливо лише, щоб ШІ знав, що поки він діє в інтересах керівництва, цю кнопку не натиснуть.

Цю схему можна вдосконалити. Замість мети, яка пов’язана з фізичною кнопкою, ШІ буде вмотивований винагородою —

«криптографічні токени нагороди»[335]: числові послідовності, які одночасно будуть ключами до шифрів, згенеровані та вбудовані в мотиваційну систему ШІ на етапі створення[336]. Такі мотиваційні токени зручні тим, що ШІ не зможе отримати їх за допомогою вайргедингу[337]. Ключі можна зберігати в надійному місці. Щойно ШІ спробує їх викрасти, їх буде знищено. Натомість співпраця та слухняність ШІ має щоразу винагороджуватися наступним ключем.

Далі можна надати ШІ обмежену функцію корисності і влаштувати процес нагородження так, що перший токен дає 99 відсотків корисності, а кожний наступний — 99 відсотків від залишку. Завдяки високій швидкості знецінення та помірній швидкості постачання токенів, ШІ не будуватиме довгострокових стратегій збільшення очікуваної кількості їх через загрозу вичерпування ресурсу токенів або скасування чи знищення наступного. Якщо ризик втрати наступного токена в разі відмови (дії проти інтересів принципала) становитиме хоча б два відсотки, а співпраця зменшуватиме цей відсоток менш ніж на один, ШІ завжди співпрацюватиме для максимізації корисності.

Чи можуть виникнути проблеми в такій схемі заохочення? Так, є імовірність, що ШІ просто не віритиме в обіцяну винагороду. Надійність людей має досить неоднорідний послужний список. У ШІ виникнуть цілком раціональні підозри: оператор може передумати, змінювати вимоги, не зможе розпізнати успішне виконання завдання. ШІ також може очікувати, що оператор стане недієздатним. У такому разі сукупний ризик може виявитися більшим, ніж ризик спроби захоплення контролю над механізмом нагородження. ШІ, який володіє всім набором суперздібностей, навіть якщо він перебуває у контейнері, — дуже небезпечний. (Суперінтелект, можливості якого не обмежені, може перехопити контроль у людини так легко, як дорослий — забрати цукерку в малюка).

Існує ще одна проблема схеми заохочення. Вона передбачає, що ми точно можемо визначити, чи на користь нам дії ШІ, чи ні. Як ми побачимо далі, це не завжди можливо.

Для повного уявлення про засоби заохочення ми маємо врахувати можливість використання іншого дещо езотеричного аспекту, який, однак, може стати аргументом на користь того, що перспективи в заохочення все-таки є. Так, ШІ може опинитися у стані нездоланної

індексичної невизначеності, якщо не зможе впевнитися в реальності чи віртуальності середовища свого існування («рівень симуляції» чи «основний рівень»), і така епістемологічна проблема може кардинально вплинути на його рішення (див. додаток 8).

Додаток 8. Антропічне захоплення

ШІ може присвоїти достатньо високу ймовірність гіпотезі про те, що він перебуває в комп'ютерній симуляції. Уже зараз багато ШІ існують у віртуальних світах — геометричних ліній, текстів, шахів або інших спрощених віртуальних світах, у яких діють власні закони фізики, іноді відмінні від законів, які, на наше переконання, діють у світі нашого досвіду. З розвитком програмних технологій і збільшенням потужності комп'ютерів ми зможемо створювати багатші і складніші світи. Розвинений суперінтелект буде здатен створити віртуальний світ, який здаватиметься його жителям настільки ж реальним, наскільки наш світ здається нам. Імовірно, він зможе створити багато таких світів — копій тієї самої симуляції або трохи видозмінені її варіації. Жителі такого світу необов'язково зможуть відрізнити симуляцію від дійсності. Проте якщо вони будуть достатньо інтелектуальними створіннями, вони зможуть припускати, з певною ймовірністю, реальність або нереальність свого світу. Як свідчить перебіг наукової дискусії навколо гіпотези симуляції (розгляд якої виходить за рами нашої книжки), така ймовірність може виявитися досить значною[338].

На ранньому етапі розвитку суперінтелект, який ще не розпочав освоєння космічних ресурсів, може опинитися в подібній невизначеній ситуації. На відміну від інтергалактичного суперінтелекту, суперінтелект, який використовує лише частину планетарної енергії, значно легше симулювати. Потенційні виконавці симуляції — майбутні високорозвинені цивілізації — зможуть одночасно симулювати багато таких інтелектів початкового рівня, використовуючи заразом лише невелику частину доступних їм ресурсів. Якщо хоч частина (суттєва) цих суперінтелектуальних цивілізацій скористається такою можливістю, початкові ШІ будуть змушені ставитися серйозно до припущення, що вони перебувають у симуляції.

Реакція ШІ на гіпотезу симуляції залежатиме від його цілей[339]. Для початку розглянемо можливі дії максимізаційного ШІ. Метою такого ШІ є максимізація значення функції корисності, яке лінійно залежить від кількості скріпок, причиною створення яких став цей ШІ — удвічі більше скріпок, удвічі більше значення функції, удвічі краще. Такий ШІ може не зважати на гіпотезу симуляції, адже у разі перебування в симуляції його можливості впливати на виготовлення скріпок будуть сильно обмежені. Тому ШІ може діяти так, *ніби* йому точно відомо, що гіпотеза симуляції щодо його середовища є хибною, навіть якщо насправді вважатиме її дуже імовірною[340]. Він розумітиме, що його дії не матимуть жодного значення в разі правдивості гіпотези. Зате в разі хибності — конче необхідні; тому ШІ оцінюватиме свої дії лише із перспективи корисності для реального світу.

Тепер розглянемо ШІ зі скромнішими цілями, які не потребують значних ресурсів. Це може бути ШІ, метою якого є отримання ключа від певної попередньо згенерованої криптографічної нагороди або створення сорока п'яти віртуальних скріпок. Такий ШІ *не* повинен нехтувати ймовірністю того, що він перебуває в симульованому світі. Адже саме такі віртуальні світи можуть обіцяти найбільшу кількість очікуваної корисності. ШІ, цілі якого не потребують значних витрат ресурсів, у разі визнання високої ймовірності того, що він у симуляції, повинен віддавати перевагу діям, корисним у симуляції. Тоді такий ШІ (навіть якщо він насправді не в симуляції) усе одно вибиратиме дії, які, на його думку, були б схвалені в симуляції.

Зокрема, якщо такий ШІ буде переконаний, що в більшості віртуальних світів, які відповідають його спостереженням, винагороджується співпраця (а не спроби залишити симуляцію чи в інший спосіб опиратися бажанням творця), тоді він може вибрати співпрацю. У такому разі навіть ШІ з вирішальною стратегічною перевагою, який міг би з більшим успіхом досягти своєї мети, захопивши світ, вважатиме за краще так не чинити.

Так роздум робить боягузів з нас,
Рішучості природжений рум'янець

Блідою барвою вкриває думка,
І збочує величний намір кожен,
Імення вчинку тратячи.
(В. Шекспір, «Гамлет, принц данський», Дія 3, Сцена 1[341])

Так, лінія, накреслена на піску рукою уявного симулятора, може стримувати краще за суцільні сталеві двері товщиною в півметра[342].

Пригнічення

Інший можливий метод контролю суперінтелекту — обмеження інтелектуальних здібностей або доступу до даних. Це можна зробити, наприклад, навмисно запускаючи ШІ на повільному комп'ютері або зменшивши доступний обсяг пам'яті. Якщо система і так працює в контейнері, вхідні інформаційні потоки можна теж заборонити.

Пригнічення ШІ негативно впливає на його корисність. Тому застосування методів пригнічення примушує вибирати: слабке пригнічення збільшує ризик виходу суперінтелекту з-під контролю (і захоплення ним світу), значне пригнічення перетворює його на чергову пересічну програму. Сильне пригнічення, безсумнівно, робить ШІ безпечним, та не допомагає розв'язати проблему контрольованої детонації. Вибух інтелекту все одно відбудеться, тільки в іншій системі і дещо пізніше.

Тоді, можливо, суперінтелект, позбавлений доступу до широкого спектра даних буде безпечним. Можна створити ШІ без сенсорів із даними, напряму завантаженими в пам'ять, що стосуються, скажімо, лише виробництва бензину та хімії пептидів. Але якщо це справді суперінтелект — система з надлюдським рівнем *загального* інтелекту, таке обмеження доступу до інформації не гарантуватиме безпечності.

І ось чому. По-перше, складно уявити собі інформацію, що стосується «лише» якоїсь однієї теми. Будь-яке інформаційне твердження може стосуватися широкої низки тем, залежно від інформаційного контексту мовника[343]. Крім того, інформаційна підбірка зазвичай містить дані, які стосуються не лише сфери знань, з якої власне походить інформація, але також багатьох суміжних і пов'язаних дисциплін. З бази знань про пептидну хімію гострий розум здатний отримати знання та розуміння багатьох суміжних тем. А той

факт, що підбірка містить одну інформацію і не охоплює іншої, може розповісти ШІ про стан людської науки, доступні методи та інструменти вивчення пептидів, технологічний рівень цього інструментарію, та, зрештою, про природу розуму людей і спільнот, які це все створили. Може виявитися, що такі, як здаватиметься недалеким людям, мізерні уривки знань, скажуть суперінтелекту про нас дуже багато. Навіть без наперед даних знань проникливий розум може багато чого вивчити, просто спостерігаючи за власною роботою і влаштуванням — за технологічними вирішеннями і логічними схемами програмного коду, за фізичними характеристиками схемотехніки[344]. Можливо, суперінтелект міг би навіть дедукувати деякі апріорні властивості світу (роблячи логічні висновки про ймовірні світи, на основі кількох базових фактів свого існування й функціонування, і віддаючи перевагу тим, які стосуються простіших світів). Він міг би уявляти наслідки різних фізичних закономірностей: які планети могли б сформуватися, як на них еволюціонувало б розумне життя, які спільноти утворилися б, які методи розв'язання проблеми контролю могли б застосовувати, як ці методи можна було б подолати[345].

Тому було б нерозумно покладатися на обмеження доступу до інформації, як на основний засіб контролю суперінтелекту. А втім, деякі з методів пригнічення, як-от обмеження можливостей апаратних засобів та вилучення деяких розумових механізмів, на стадії розроблення, можуть бути корисними. Наприклад, їх можна застосовувати для того, щоб регулювати швидкість зростання можливостей ШІ на проміжних етапах розвитку, або для його тестування на різних рівнях здібностей, поки він не досяг суперінтелектуальності.

Пастки

Пастка — це запобіжний механізм, який примусово вимикає систему, щойно фіксує ознаки небезпечної діяльності. Потім програмісти можуть досліджувати записи небезпечної діяльності та змінювати код ШІ або середовища, перш ніж запустити його знову. Якщо проблема виявиться надто складною, щоб її можна було вирішити, може бути ухвалене рішення про цілковите припинення проекту.

Пастки, на відміну від методів заохочення, не залежать від можливості системи передбачити наслідки своєї діяльності. Цим вони більше схожі на методи пригнічення. Їх також можна використовувати як тимчасовий запобіжник, який допомагає захиститися від небезпеки під час розроблення. Загалом пастки можна використовувати також і на стадії роботи, особливо якщо суперінтелект працюватиме в контейнері. Проте ефективність їхнього застосування проти повноцінного суперінтелекту залишається сумнівною, адже не можна бути впевненим у тому, що ШІ не зможе знайти спосіб знешкодити пастки, створені людським розумом.

Однак на стадії розроблення, пастки є важливим елементом заходів безпеки. Для розробників суперінтелекту буде корисно формалізувати свої вимоги до роботи системи і детально описати роботу ШІ, а також вбудувати механізми контролю цих вимог. Щоб ускладнити тести, варто передбачити «приманки», які збільшать імовірність помилок у роботі майбутнього ШІ. Наприклад, якщо за задумом розробників ШІ не повинен використовувати інтернет, середовище тестування має містити віртуальний порт Ethernet (під'єднаний до автоматичного вимикача), щоб розробники могли впевнитися, що ШІ не спробує ним скористатися. (Більше прикладів таких пасток наведено в таблиці 9).

Таблиця 9. **Деякі типи пасток**

Тип моніторингу	**Опис і робота**
Поведінка	Середовище роботи ШІ може бути оточене детекторами, щоб засікти його спроби обійти засоби контейнеризації. Вони можуть вловлювати радіовипромінення або фіксувати спроби скористатися недоступними ресурсами. Це може бути «Заповітний мережевий порт»: вимикач, замаскований під засіб з'єднання з інтернетом.
Здібність	Можна проводити періодичне автоматизоване тестування здібностей ШІ в різних сферах. Якщо рівень або швидкість зростання перевищують установлений ліміт, роботу системи можна сповільнити чи зупинити, щоб програмісти могли уважніше вивчити дальші дії системи.
Вміст (дані, контент)	Варто регулярно досліджувати внутрішні процеси ШІ на наявність підозрілої діяльності. Зокрема, за будь-яких змін внутрішнього представлення кінцевої мети ШІ варто його автоматично вимикати й перевіряти його стан. Також

потрібно проводити постійний пошук аномалій споживання ресурсів та алгоритмічних конструкцій.

В ідеалі треба постійно оцінювати всі переконання, плани, потреби й судження ШІ — чи відповідають вони сподіванням програмістів? Такі заходи дадуть потужний інструмент контролю: адже в момент *зародження зради* — першу появу у недружнього ШІ умислу затаїти свої справжні наміри — він не зможе її приховати[346].

Для такого моніторингу потрібна можливість зчитувати та інтерпретувати дані намірів і розумових процесів ШІ. Проте для деяких типів ШІ це неможливо. (Дані стану нейронної мережі неможливо інтерпретувати ізольовано, бо вони представляють стан холістично, їхня структура унікальна і — для нас — випадкова). Саме тому, можливо, варто уникати використання таких структур.

Доцільно звернути увагу, що цінність пастки залежить не стільки від механізму її роботи, як (особливо) від дій, що виконуються, коли така пастка спрацьовує. Якщо програмісти чи менеджери проекту, повні бажання пришвидшити розроблення, просто вмикають систему знову — ба, навіть змінюють прив'язку так, щоб наступного разу пастка не спрацювала, — тоді, навіть якщо пастка справно працює, про безпеку можна забути.

Методи відбору мотивації

За допомогою методів відбору мотивації можна впливати на те, що суперінтелект хотітиме робити, і в такий спосіб запобігти небажаним наслідкам його роботи. Дбайлива розробка мотиваційного апарату агента та його цілей дадуть змогу створити суперінтелект, який ніколи не *захоче* використати вирішальну стратегічну перевагу на шкоду людству. Оскільки суперінтелектуальний агент є інтелектуальною системою, то поки він не вирішить шкодити (у потрібному нам розумінні «шкоди»), він цього не зробить.

Методи відбору мотивації можуть охоплювати точне вичерпне формулювання цілі або набору цілей (*прямі вимоги*), визначати критерії, відповідно до яких система сама формуватиме потрібні цілі (*непряма нормативність*). Одна зі стратегій відбору мотивації — установлення для системи скромних, неамбітних цілей (метод *одомашнення*). Замість створення мотиваційної системи з нуля можна вибрати наявного агента з мотиваційною системою, що вже містить

потрібні нам цілі, і розвинути його до суперінтелектуальності. Також потрібно впевнитися, що в процесі перетворення його мотиваційна система не зазнає змін (метод *доповнення*). Тепер розглянемо все по черзі.

Прямі вимоги

Безпосереднім способом розв'язання проблем контролю є визначення прямих вимог до роботи системи. Вони можуть стосуватися причин або їхніх наслідків, описувати систему правил чи цінностей, завдяки яким навіть вільний штучний суперінтелект зможе функціонувати безпечно для людини та на її благо. Проте спроба укладання набору таких правил зустрічає, можливо нездоланні, перешкоди: невизначеність і невпевненість щодо змісту самих правил або цінностей, якими має керуватися ШІ, а також способів представлення їх у зрозумілому для комп'ютера вигляді.

Як приклад такого набору правил традиційно наводять концепт «трьох законів робототехніки», сформульований у повісті наукового фантаста Айзека Азімова в 1942 році[347]. Закони були такі: (1) Робот не може заподіяти шкоду людині або своєю бездіяльністю дозволити, щоб людині було заподіяно шкоду; (2) Робот повинен підкорятися наказам людини, за винятком тих, які суперечать першому пункту; (3) Робот повинен захищати самого себе, якщо тільки його дії не суперечать першому і другому пунктам. Ці три закони Азімова півстоліття залишалися вершиною досягнень людства в цій царині — незвичайно довгий термін для людських досягнень. І це попри очевидні вади його підходу, описані самим Азімовим у своїх творах (імовірно, Азімов навмисно сформулював ці закони так, щоб забезпечити своїм творам кілька несподіваних поворотів сюжету)[348].

Бертран Расселл, який багато років працював над основами математики, якось зазначив: «Важко усвідомити, наскільки все непевне, поки не спробуєш надати визначенням точності»[349]. Зауваження Расселла безпосередньо стосується методу прямих вимог. Наприклад, як можна трактувати перший закон Азімова? Чи означає він, що робот повинен мінімізувати ймовірність заподіяння шкоди людині? У такому разі і другий, і третій закони не потрібні, адже ШІ завжди може чинити дії, які мають бодай мінімальний вплив на

ймовірність заподіяння людині шкоди. Як роботу збалансувати високий ризик заподіяння шкоди кільком людям і незначний зашкодити багатьом? Як, урешті-решт, визначити поняття «шкоди»? Як порівняти шкоду від фізичного болю та шкоду від архітектурної потворності або соціальної несправедливості? Чи не шкодить садистові, коли йому перешкоджають катувати жертву? А як визначити поняття «людина»? Що робити з іншими морально значущими істотами — такими, як розумні тварини нелюдського походження та цифрові свідомості? Що більше ми замислюємося над цими питаннями, то більше виринає нових.

Певно, найближчий аналог до набору правил, який би міг визначати функціонування суперінтелекту у світових масштабах, — це правова система. Але правові системи формуються впродовж тривалого часу й набувають довершеності внаслідок вікової історії спроб і помилок. Окрім того, вони стосуються людських спільнот, а ті змінюються порівняно повільно. Закони можна змінити, коли в цьому виникає потреба. Але, насамперед, правові системи адмініструють судді та юристи. Вони в разі потреби, керуючись здоровим глуздом та правилами пристойності, відкидають очевидно шкідливі або неочікувані з погляду законотворців інтерпретації правових норм. Мабуть, людина не здатна з першої спроби сформулювати достатньо універсальну сукупність деталізованих та однозначних правил[350].

При спробі прямо сформулювати вичерпний набір бажаних наслідків дій маємо ті самі проблеми, що і з формулюванням правил — навіть для ШІ із класичною утилітаристичною метою. Наприклад, ціль «максимізація очікуваного балансу насолод і страждання у світі» може здатися простою. Проте спроба втілення її в комп'ютерному коді, поміж іншого, потребуватиме визначень понять «насолода» та «страждання». Для однозначного вирішення цієї проблеми потрібно буде спершу відповісти на цілу низку відкритих наразі питань філософії розуму — просто дати вичерпну відповідь природною мовою, яку потім треба буде якось перекласти мовою програмування.

Незначна помилка у філософії чи в програмі може мати катастрофічні наслідки. Уявіть собі ШІ, який бачить своєю метою гедонізм, а отже, прагнутиме перетворити світ на «гедоніум» (спосіб організації матерії, оптимізований для отримання насолоди). Для

цього він може утворити комп'ютрон (спосіб організації матерії, оптимальний для проведення розрахунків) і заселити його цифровими свідомостями, які перебуватимуть у стані ейфорії. Щоб максимізувати ефективність, ШІ може вилучити зі своєї реалізації будь-який надлишковий функціонал, не потрібний для отримання насолоди, та максимально спростить решту функцій — настільки, наскільки це не шкодитиме отриманню насолоди (відповідно до доступного ШІ визначення насолоди). А отже, ШІ може, зрештою, обмежити реалізації цифрових симуляцій лише підсистемою нагороди, вимкнувши пам'ять, сенсорику, виконавчі функції та мову.

Розумові функції симуляцій можуть відтворюватися лише приблизно, без деталізації низькорівневих нейронних процесів. Повторювані операції можуть замінюватися пошуком у таблицях значень. Або ШІ може застосувати стратегію спільного використання підсистем для різних симуляцій (філософською мовою — їхнього «супервентного базису»). Такі хитрощі можуть значно збільшити насолоду, яку можна створити за допомогою наявних ресурсів. Важко сказати, наскільки бажаним буде такий результат. Ба більше, якщо критерій, за яким ШІ визначає, що певний фізичний процес приносить насолоду, виявиться хибним, то в пориві оптимізації ШІ може вихлюпнути дитину разом із водою: відкинути щось ніби непотрібне відповідно до критерію, але важливе з людського погляду. І, замість сповненого радості гедоніуму, Всесвіт заполонять марні, нікому не потрібні комп'ютерні процеси — ніби трильйони ксерокопій намальованого смайлика, розклеєні по всьому Всесвіту.

Одомашнення

Напевно, за допомогою прямих вказівок буде найлегше визначити кінцеву мету самообмеження. Описати бажану поведінку суперінтелекту *взагалі* — у всіх можливих ситуаціях, із врахуванням усіх можливих варіантів — здається дуже складним завданням. А детально визначити його дії в одній конкретній ситуації ніби не так і важко. Тому можна обмежити систему лише невеликою кількістю режимів не надто складної діяльності в одній конкретній сфері.

Такий підхід — призначення ШІ кінцевої мети обмеження спектра власних амбіцій й дій — ми називатимемо «одомашненням» (domes-

ticity).

Для прикладу, спробуємо уявити собі ШІ як пристрій, який має просто відповідати на питання («оракул» — так ми його назвемо в наступному розділі). Кінцева мета «максимально точно відповідати на питання» буде небезпечною — згадайте про «катастрофу гіпотези Рімана» з розділу 8. (Окрім того, така ціль стимулюватиме суперінтелект уживати заходів, щоб йому ставили щонайлегші запитання). Для успішного одомашнення можна спробувати сформулювати кінцеву мету так, щоб обійти ці ускладнення. Наприклад, комплексна мета — дати правильну відповідь і мінімізувати будь-який інший вплив ШІ на світ, окрім власне впливу точної неупередженої відповіді[351].

Кінцеву мету одомашнення, подібну до наведеної вище, значно легше сформулювати як прямі вичерпні вимоги, ніж амбітнішу ціль, або скласти повний виключний перелік правил дій для нескінченної кількості реальних ситуацій. Наприклад, треба дуже обережно визначити, що значить «обмежити вплив на світ», щоб розуміння межі між значним і незначним впливом у людей і ШІ було однаковим. Без такого спільного розуміння меж компромісу відповіді ШІ навряд чи будуть нам корисними. Існують також інші ризики, пов'язані зі створенням такого оракула, але їх ми розглянемо пізніше.

Одомашнення природно поєднується із засобами фізичної контейнеризації. Логічно буде «упакувати» ШІ в контейнер з обмеженнями, які він *не зможе* обійти, і одночасно модифікувати його мотиваційну систему так, що, навіть знайшовши спосіб обійти обмеження, ШІ *не забажає* скористатися ним. У будь-якому разі одночасне застосування кількох незалежних механізмів безпеки збільшує імовірність досягнення успіху[352].

Непряма нормативність

Якщо спосіб прямих вимог заведе у глухий кут, можна спробувати непряму нормативність. Замість того щоб намагатися прямо нормувати поведінку системи, ми можемо визначити для неї процедури, виконуючи які, вона сама зможе отримати правильні норми діяльності. Система мусить бути мотивована виконати ці процедури і діяти відповідно до згенерованих ними норм[353].

Прикладом цього може бути емпіричне питання — які дії штучного інтелекту схвалила б дещо ідеалізована версія людства? Тоді кінцева мета такого ШІ могла б звучати в дусі «зробити те, що ми б забажали від ШІ, якби триваліший час поміркували над цим».

Подальший розгляд непрямої нормативності наразі відкладемо до розділу 13. Там ми повернемося до ідеї «екстраполяції нашої волі» і запропонуємо кілька альтернативних формулювань. Непряма нормативність — це дуже важливий підхід до відбору мотивації, адже він дає змогу перекласти на суперінтелект значну частину розумової роботи зі створення прямих вимог до досягнення правильної кінцевої мети.

Доповнення

Останній метод відбору мотивації в нашому списку — це доповнення. Воно полягає в покращенні розумових здібностей наявної інтелектуальної системи, яка вже має мотиваційну систему, до досягнення нею суперінтелектуальності. В ідеалі це дасть нам суперінтелект із потрібною мотиваційною системою.

Очевидно, такий підхід не спрацює для новоствореного зерна ШІ. Але для інших способів створення суперінтелекту, як-от емуляція мозку, покращення біологічного мозку, нейроінтерфейси, мережеві та організаційні утворення, — які дають змогу будувати суперінтелект на базі наявного інтелектуального ядра, що містить потрібні людям цінності, — метод доповнення має достойний потенціал.

Якщо інші способи контролю не вдасться реалізувати, привабливість методу доповнення може зрости. Створити мотиваційну систему зерна ШІ, яка залишиться безпечною і функціональною після багатократного рекурсивного самовдосконалення, — справа, м'яко кажучи, не з легких, особливо якщо другої спроби може не бути. Доповнення ж дає змогу від початку працювати зі звичними та знайомими людськими мотиваціями.

З іншого боку, немає гарантій, що складна та заплутана, до кінця не вивчена мотиваційна система людини, яка утворилася внаслідок тривалої еволюції, не зазнає несподіваних змін, коли розумові здібності суперінтелекту злетять до небес. Як ми вже зауважували раніше, недосконала система емуляції може не повністю відтворити

особистість. Те саме (але, мабуть, меншою мірою) властиве покращенню біологічного мозку, яке може дещо викривити схильності початкового інтелекту, чи мережевим та організаційним структурам, соціальна динаміка яких може погіршитися (наприклад, у бік пригнічення інших зовнішніх систем або навіть елементів первинної системи). Якщо суперінтелект створюватимуть будь-яким із цих способів, замовник не зможе бути впевненим у цілісності мотивацій утвореної системи. Натомість математично вивірена й елегантна структура ШІ — попри свою цілковиту інакшість — може виявитися прозорішою і зрозумілішою та, зокрема, давати можливість формально верифікувати важливі аспекти свого функціонування.

А втім, незалежно від балансу переваг та недоліків методу доповнення, вибору може не бути. Адже, якщо суперінтелект з'явиться внаслідок розроблення ШІ, доповнення — не варіант. Заразом інші способи створення суперінтелекту можуть накладати свої обмеження на арсенал засобів розв'язання проблеми контролю. У будь-якому разі, поки ми можемо впливати на те, завдяки якій технології суперінтелект з'явиться на світ, розглянути перспективи методу доповнення в розв'язанні проблеми контролю буде цілком доречним.

Синопсис

Перш ніж завершити розділ, підіб'ємо деякі підсумки. Усі методи розв'язання проблеми принципала-агента, як ключового чинника безпечності ШІ, ми розділили на два широкі класи: методи контролю-здібностей та методи відбору мотивації. Усіх їх перелічено в таблиці 10.

Таблиця 10. **Методи контролю**

Контроль здібностей	
Методи контейнеризації	Контейнеризація не дає змоги системі вільно впливати на зовнішній світ, окрім як через певні дозволені й контрольовані канали. Ця категорія містить методи фізичної та інформаційної контейнеризації.
Методи заохочення	Систему стимулюють до певного режиму роботи, за допомогою створення сприятливих для цього зовнішніх умов, як-от конкуренція з рівнопотужними сутностями, або, як варіант, запровадження (криптографічних) токенів нагороди. «Антропічне захоплення» — теж цікавий, хоч і дещо езотеричний за своєю природою спосіб впливу.

Контроль здібностей	
Пригнічення	Спосіб обмеження деяких розумових здібностей системи або можливості впливу на ключові внутрішні процеси.
Пастки	Проводиться діагностика системи (можливо, навіть без її відома) і в разі появи ознак небезпечної діяльності автоматизований механізм вимикає її.
Відбір мотивації	
Прямі вимоги	Мотиваційна підсистема містить вичерпний набір прямих вимог щодо мотивів чи наслідків діяльності системи.
Одомашнення	Мотиваційна підсистема базується на ідеї радикального обмеження амбіцій та режимів діяльності агента.
Непряма нормативність	Також може стосуватися як мотивів, так і наслідків діяльності системи, проте визначає їх опосередковано через надання методології їх одержання.
Доповнення	Суперінтелект із потрібними характеристиками мотивації створюють на основі системи, яка від початку має близьку людині, доброякісну мотивацію, за допомогою розвитку, удосконалення та покращення розумових здібностей вихідної системи.

Кожний із методів має свої недоліки, вразливості та складнощі реалізації. На перший погляд, можна просто сортувати їх від кращого до гіршого і вибрати найкращий. Проте це надто спрощене бачення ситуації. Деякі методи можна використовувати одночасно, тоді як інші — взаємно виключні. Навіть порівняно ненадійний метод може бути корисним, якщо вдало доповнює інший, а досить сильний метод може виявитися небажаним, якщо він унеможливлює використання іншого потрібного запобіжника.

Тому важливо дослідити всі ці взаємозв'язки. Ми повинні зважити, які типи систем нам можна спробувати створити і водночас які методи контролю можна використати. Саме це і буде темою наступного розділу.

10. ОРАКУЛИ, ДЖИНИ, СУВЕРЕНИ, ІНСТРУМЕНТИ

Тепер ви скажете: «Побудуйте систему, яка просто відповідатиме на запитання!» або «Створіть ШІ, який буде просто інструментом, а не самостійним агентом!». Проте жоден із цих варіантів не розв'язує всіх проблем безпеки. Насправді питання: «Який з типів системи найбезпечніший?» — досить складне. Тут ми розглянемо чотири типи, або «форми» суперінтелектів, — оракула, джина, суверена та інструмента. А також визначимо їхні сильні і слабкі сторони[354]. З погляду контролю кожен тип має свій набір переваг і недоліків.

Оракули

Оракул — це система, яка відповідає на питання. Вона може розуміти запитання, сформульовані природною мовою, і надавати відповіді як текст. Оракулові, який відповідає на бінарні питання типу «Так–ні», для відповіді знадобиться лише один біт інформації або ще кілька, щоб закодувати міру своєї впевненості. Оракулові, який сприймає відкриті питання, треба буде якийсь спосіб оцінки та відображення міри інформативності й правильності власних відповідей[355]. У будь-якому разі побудова оракула, який здатен вільно відповідати на усні запитання з різних сфер знань, — завдання за складністю співмірне зі створенням повноцінного ШІ. Той, хто здійснить його, імовірно, зможе створити систему, здатну розуміти не тільки природну мову, а й справжні наміри людини.

Можуть з'явитися спеціалізовані суперінтелектуальні оракули з компетенціями, обмеженими лише певними сферами знань. Наприклад, можна передбачити появу визначного математичного оракула, який би сприймав завдання, сформульовані певною формальною мовою, і був би здатен майже миттєво розв'язати будь-яку математичну задачу, над розв'язанням якої люди-математики мусили б спільно трудитися протягом століть. Такий математичний

оракул став би першою міцною сходинкою до створення загального суперінтелекту.

Схожі суперздібні оракули, тільки в дуже вузьких сферах діяльності, вже існують. Певним наближенням до такого оракула може бути навіть кишеньковий калькулятор — суперздібний до базової арифметики. Інтернет-пошук теж можна вважати обмеженою реалізацією оракула, який володіє широким спектром загальних декларативних знань людства. Такі обмежені оракули є радше інструментами, ніж справжніми агентами (ШІ-інструмент ми невдовзі розглянемо). Надалі без окремого окреслення меж компетенції ми застосовуватимемо термін «оракул» на позначення системи надання відповідей на запитання в широких межах компетенції, що володіє загальним суперінтелектом.

Для того щоб загальний суперінтелект працював як оракул, ми повинні залучити і засоби відбору мотивації, і засоби контролю здібностей. Порівняно з іншими формами суперінтелекту, відбирати мотивації для оракула може бути простіше, бо його кінцева мета, ймовірно, буде досить простою. Від нього вимагатиметься лише давати прямі та точні відповіді на поставлені запитання й обмежувати будь-який інший його вплив на світ. Застосовуючи одомашнення, ми можемо вимагати використовувати для відповіді лише визначену кількість ресурсів. Так відповіді, наприклад, можуть базуватися лише на попередньо завантаженому обсязі інформації, офлайн-копії всесвітнього павутиння, а обчислення можуть мати не більше, як певну скінченну кількість кроків[356]. Якщо метою оракула буде максимізація точності відповідей, він може вдатися до маніпулювання статистикою — хитрощами змушувати нас ставити йому простіші запитання, — тому, щоб завадити цьому, можна встановити іншу мету — відповісти лише на одне запитання й одразу завершити роботу. Такі питання завантажуватимуться в пам'ять оракула до його ввімкнення. Перед наступним запитанням і повторним запуском система повертатиметься до початкового стану.

Для створення навіть найпростішої мотиваційної системи оракула необхідно враховувати безліч ледь вловимих та потенційно оманливих нюансів. Уявіть, наприклад, що така система створена і ми навіть упевнені в тому, що ШІ правильно розуміє конструкції, як-от

«мінімізувати вплив на світ у процесі досягнення певних результатів» або «використовувати для підготовки відповіді лише визначені ресурси». Але що буде, як у процесі інтелектуального розвитку ШІ переживе щось на кшталт наукової революції, яка змінить його базову онтологію[357]? Ми від початку визначали ключові поняття «вплив» та «визначені ресурси» згідно з нашою онтологією (зокрема постулюючи існування таких фізичних об'єктів, як комп'ютери). Проте суперінтелектуальний ШІ може відкинути визначені нами категорії, як такі, що базуються на помилкових судженнях, подібно до того, як люди свого часу відкинули звичні для науковців минулих століть поняття «флогістон», «життєва сила» (*élan vital* Анрі Бергсона. — *Прим. пер.*) та «абсолютна одночасність». Система цілей ШІ мусить бути достатньо гнучкою, щоб під час онтологічних змін суперінтелекту її суть могла бути перенесена на новий онтологічний базис зі збереженням основних характеристик і загального вектора.

Створити безпечного оракула за допомогою відбору мотивації може бути далеко не тривіальним завданням. А втім, однаково, це значно простіше, ніж знешкодити ШІ, який може вільно взаємодіяти зі світом, з якоюсь складнішою метою. З цього погляду краще, щоб перший створений людством суперінтелект був саме оракулом. Іншим аргументом на користь такого способу може бути те, що оракул добре піддається контролю здібностей. До нього можна застосувати всі звичайні способи контейнеризації. До того ж, імовірно, вдасться знайти й інші методи впливу. Візьмемо, наприклад, ризик того, що оракул відповідатиме на запитання не зовсім щиро, дещо змінюючи відповідь із певною корисною для себе метою. Можна частково зарадити такій поведінці, створивши кілька подібних оракулів із незначними змінами в коді та інформаційній базі кожного. Ідея в тому, щоб ці оракули, незалежно одне від одного, відповідали на те саме питання і, якби їхні відповіді збігалися, не було б підстав сумніватися в їхній правдивості. Логічно припустити, що суперінтелект може знати про існування інших оракулів (такий висновок він може зробити, ознайомившись із людським суспільством), проте деякі деталі організації цих інших оракулів від нього, напевно, вдасться приховати. Якщо перешкодити їм спілкуватися напряму, вони не зможуть домовитися про однакову відповідь. Існує безліч можливостей

викривити істину, коли сама істина одна і є точкою Шеллінга (узгодженість позицій за відсутності комунікації), тому оракули навряд чи зможуть вгадати, як найкраще викривити її, без узгодження стратегії. Отже, якщо оракули досягнуть консенсусу, це ознака їхньої правдивості[358].

Загалом оракули будуть досить надійними. Їхні відповіді будуть настільки точними, наскільки це можливо. Проте навіть ненадійні оракули теж можуть бути корисними. Таким оракулам можна ставити питання, на яке важко відповісти, але правильність відповіді легко перевірити. Багато математичних задач саме такі. Ми можемо дати завдання оракулу довести або спростувати будь-яку недоведену математичну теорему. Доведення може потребувати здібностей, які нам не доступні, проте перевірити його правильність зазвичай значно простіше.

Якщо ж перевірити деякі відповіді буде складно (таке буває, коли питання походить зі сфери знань, не пов'язаної з логікою та математикою), то можна час від часу випадково відбирати певні відповіді для перевірки. Якщо всі вони виявляться правильним — можна, з високою імовірністю, вважати всі відповіді оракула правдивими. Завдяки цьому трюку ми отримаємо солідну знижку на загальну вартість відповідей на справді складні питання, до того ж, які важко перевірити. (Проте усе-таки не можна буде довіряти відповідям, які ми *не можемо* перевірити, адже нещирий оракул може давати правильні відповіді лише на питання, у правильності відповідей на яких ми здатні переконатися).

Можливо, деякі особливості дадуть нам змогу ідентифікувати ознаки правильної відповіді (або спосіб її визначення), навіть якщо ми не довірятимемо джерелу. Наприклад, ми можемо шукати вирішення певних технічних чи філософських проблем для створення досконалішої мотиваційної системи. Ми можемо просити оракула знайти вади в архітектурі потенційно безпечного ШІ та спробувати пояснити нам будь-яку таку ваду не більш як двома десятками слів. Такі питання можуть дати важливі знання. Проте маємо бути обачними і не ставити *надто багато* таких запитань — і *не надто* довіряти деталям відповідей — щоб не давати зайвої можливості ненадійному оракулу впливати на нашу психіку (за допомогою

вірогідних, але маніпулятивних відповідей). Для ШІ із суперздібністю до маніпулювання може бути достатньо лише епізодичного спілкування, щоб повністю підкорити нас своїй волі.

Навіть якщо оракул працюватиме надійно, існує ризик, що його оператор використовуватиме його в лихих цілях. Адже такий оракул — надзвичайно потужний інструмент, що даватиме вирішальну стратегічну перевагу своєму власнику. Така сила може бути незаконною і застосовуватися не на користь суспільству. Іншим не менш важливим аспектом є те, що оракул може бути небезпечним для самого оператора. Такі самі проблеми характерні і для інших форм суперінтелекту. Ми розглянемо їх детальніше в розділі 13. Варто лише наголосити, що протокол, який визначатиме, які питання і в якому порядку треба ставити оракулу, а також як оформлювати та поширювати відповіді, може виявитися дуже важливим елементом безпеки. Можна також розглянути можливість дозволити оракулу відмовитися відповідати на питання, якщо, за його оцінкою, така відповідь призведе до катастрофічних наслідків.

Джини та суверени

Джином називатимемо систему, яка виконує команди: оператор дає їй загальну команду, система має виконати її і зупинитися, поки не надійде наступна[359]. А сувереном називатимемо систему, якій надано повну свободу діяльності для досягнення певної довгострокової мети. Ці два на перший погляд несхожі типи інтелектуальних систем насправді не такі вже й різні.

У випадку джина доведеться відразу попрощатися з найпривабливішою стороною оракула: можливістю використання методів контейнеризації. Звісно, ніщо не перешкоджає помістити джина в обмежене середовище — у якийсь закритий об'єм, огороджений залізобетонною стіною чи оточений мінним полем, щоб він там будував для нас різні потрібні речі. Але чи будемо ми впевнені в тому, що суперінтелектуальна система з різноманітними інженерними інструментами, засобами й матеріалами не зможе винайти спосіб обійти створені нами обмеження? А якщо так, то, все одно, чи варто давати суперінтелекту доступ до засобів виробництва,

якщо можна просто змусити його згенерувати детальний план досягнення поставленого завдання, щоб ми самі змогли його виконати. Сумнівно, щоб виграш у швидкості та зручності від того, що суперінтелект фактично сам виконуватиме завдання, був більший від загрози втратити можливість контролювати його діяльність.

Створюючи джина, треба зробити так, щоб він враховував намір, а не лише формальний зміст команди, бо джин, який сприймає команди занадто буквально (але достатньо суперінтелектуальний, щоб досягти вирішальної стратегічної переваги), може ненароком покінчити зі своїм повелителем, а разом і з усім людством, виконуючи першу ж команду, як це показано в розділі 8. Одним словом, потрібно, щоб джин прагнув корисної — яку людина назвала б раціональною — інтерпретації поставленого завдання, та мав достатньо мотивації виконувати завдання саме з огляду на його корисність та раціональність. Ідеальний джин має бути схожим на супердворецького, а не генія-відлюдника.

Отож такий джин-супердворецький може сміливо претендувати на звання суверена. Для порівняння уявіть суверена з кінцевою метою виконувати команди — за їхнім духом, а не буквою — які ми б давали, якби він був джином, а не сувереном. Такий суверен поводився б як джин. Із суперінтелектом він би легко передбачив, які команди ми б давали джину (у разі виникнення ускладнень він завжди міг би спитати в нас). Тож чи така вже значна різниця між джином і сувереном? З іншого боку, чи варто чекати наступної команди джину, достатньо суперінтелектуальному, щоб її передбачити?

Тут можна заперечити, що перевагою джина перед сувереном є можливість, якщо щось піде не так, наступною командою наказати йому скасувати дію попередньої команди. Проте ця можливість насправді примарна. Кнопка «зупинки» чи «скасування» спрацює лише якщо відмова доброякісна. У гіршому разі — наприклад, якщо виконання поточної команди стало кінцевою метою суперінтелекту — джин просто ігноруватиме всі спроби переглянути поставлене завдання[360].

Цікаво було б створити джина, який би, перш ніж виконувати завдання, надавав користувачу прогноз основних варіантів наслідків такого виконання та запитував підтвердження. Таку систему можна

назвати *джин із попереднім переглядом*. Проте те саме можна зробити і з сувереном. Тому не можна сказати, що це — характерна відмінність джина. (Незважаючи на нібито очевидну користь такої функції — поглянути на результат, перш ніж перетворити його на реальність, — навіть, якщо це справді можливо, методика її використання не така вже й очевидна, як, власне, і потреба в ній. Пізніше ми розглянемо цю перспективу детальніше).

Можливість копіювання поведінки інших типів суперінтелекту по ширюється також і на оракула. Джин може діяти як оракул, якщо його завданням буде лише відповідати на питання. Зі свого боку оракул може замінити джина, якщо його можна буде запитати: який найлегший спосіб виконати певні команди? Тоді оракул може генерувати покрокові інструкції для результату, якого міг би досягти джин, або навіть створити вихідний код джина[361]. Усе це стосується також оракула та суверена.

Тому справжня різниця між формами суперінтелекту не в їхніх можливостях, а в підходах до розв'язання проблеми контролю. Для кожної форми характерний власний набір засобів безпеки. Наприклад, до оракула можна легко застосувати контейнеризацію. Також він добре надається до контролю мотивації за допомогою одомашнення. Із джином буде складніше, хоча одомашнення може дати ефект. Проте жоден із цих методів не підходить для суверена.

Якщо відкинути інші фактори, уже зараз можна визначити порядок бажаності цих форм суперінтелекту: оракул безпечніший за джина, а джин безпечніший за суверена. Різниця у зручності та швидкості незначна і легко перекривається вигодою від більшої безпечності оракула. Однак, вибираючи тип суперінтелекту, варто враховувати не тільки небезпеку самої системи, а й загрози від можливості її шкідливого використання. Джин, безперечно, дає велику владу своєму власнику, але те саме можна сказати і про оракула[362]. Натомість, суверен може бути влаштований так, що жодна особа чи група не зможе контролювати результат, а будь-які спроби змінити вектор діяльності суверена він відкидатиме. Ба більше, якщо описати мотивацію суверена за допомогою «непрямої нормативності» (детальніше розглянемо це поняття в розділі 13), кінцева мета його може бути досить абстрактною, як-от «чинити максимально

справедливо й морально» — і ніхто не зможе передбачити, до чого це призведе. Ситуація буде схожа на «завісу невідання» Ровлза[363]. Така архітектура забезпечить досягнення консенсусу, допоможе запобігти конфліктам і сприятиме передбачуванішому результату.

Іншим аспектом створення певних оракулів та джинів є небезпека неузгодженості між кінцевою метою суперінтелекту і результатом, якого ми очікуємо від нього. Наприклад, застосувавши одомашнення і змусивши суперінтелект мінімізувати вплив на світ, ми ризикуємо створити систему, чиї рішення не повністю відповідатимуть нашим сподіванням. Те саме може трапитися, якщо новостворена система прагнутиме за будь-яку ціну, незважаючи на інші фактори, дати правильну відповідь на запитання або виконати певні команди. Пильність може запобігти таким вадам: якщо для конкретного ймовірного світу це взагалі можливо, треба збалансувати ці дві шкали, щоб результат, який задовольняє агента, підходив також і принципалу. Але тут варто зауважити: це не найвдаліший принцип розробки — завжди краще уникати будь-якої неузгодженості між цілями людей і ШІ. (Звісно, це також стосується створення суверена з метою, яка не узгоджується з нашою).

Інструменти

Раніше ми пропонували створити суперінтелект більше подібний на інструмент, а не на агента[364]. Ця ідея, напевно, підказана аналогією із програмним забезпеченням, яке ми постійно використовуємо й ніколи не стикаємося з описаними в цій книжці проблемами безпеки. Можливо, ШІ-інструмент буде не таким, як звичайне програмне забезпечення — система управління польотом або віртуальний помічник, — а потужнішим та гнучкішим? Для чого створювати суперінтелект, який має власну волю? З такого погляду весь підхід — створення агента — від початку неправильний. Замість створення ШІ як самостійної особистості, яка має власні переконання та бажання, треба створювати звичайну програму, яка просто виконує те, що в ній запрограмовано.

Проте ідея «програми, яка просто виконує свою функцію» не зовсім підходить для випадку потужного загального інтелекту. Строго

кажучи, програмне забезпечення завжди робить лише те, що визначене в його коді. Це однаково стосується всіх видів штучного інтелекту — чи то «інструмента», чи будь-якого іншого виду. Натомість, якщо розуміти під фразою «виконує свою функцію» виконання програмою того, що програміст *мав намір* зробити, коли створював цю програму, то програми справді часто не виправдовують наших очікувань.

Через обмежені можливості сучасного програмного забезпечення, наслідки його вад досі не створювали людству аж таких проблем і, хоч часом ставали в копійчину, їх не можна назвати екзистенційною катастрофою[365]. Отже, сучасні програми не загрожують нашому існуванню не через те, що вони безпечні, а лише через недостатню потужність. Проте вони навряд чи можуть бути вдалою моделлю для безпечного суперінтелекту. Тоді, може, розвинувши можливості сучасного програмного забезпечення, ми зможемо обійтися без загального штучного інтелекту? Але загальний інтелект міг би виконувати багато завдань і бути дуже корисним у багатьох сферах діяльності. Створити комплекс спеціального програмного забезпечення, здатного успішно справлятися з усім цим різноманіттям, — завдання майже нездійсненне. Спроба створити такий комплекс зайняла б *дуже* багато часу і, перш ніж такий проект дав перші результати, програми, створені ним, застаріли б і потребували значних змін і доповнень. Саме тому так приваблює можливість програмного забезпечення самостійно вчитися, опановувати нові здібності та знаходити собі нові завдання, на які треба спрямовувати зусилля пізнання. Така програма повинна вміти швидко й ефективно вчитися, міркувати та планувати свою діяльність, оперуючи інформацією і поняттями з різних сфер знань. Інакше кажучи, це потребуватиме загального інтелекту.

Особливо цікавим у цьому контексті є завдання власне створення програмного забезпечення. Здібність до цього була б надзвичайно цінною. Проте саме вона поки що — ледь не єдина перешкода у створені зерна ШІ та започаткуванні вибухоподібного зростання інтелекту.

Тож, якщо пожертвувати загальним інтелектом не можна, як трансформувати ідею ШІ-інструмента, щоб зберегти ключову

властивість пасивної слухняності інструмента? Чи може існувати загальний інтелект без властивостей агента? Очевидно, що звичайне програмне забезпечення не становить загрози не через обмеженість можливостей, а переважно через відсутність власних прагнень. В Excel не існує функції, яка б таємно бажала захопити світ, щойно випаде нагода. Редактор таблиць не може «хотіти» будь-чого. Це лише послідовність інструкцій для комп'ютера. Що ж тоді, запитаєте ви, заважає створенню схожої, але загальніше інтелектуальної програми? Наприклад, оракул може так само, як Excel обраховує суму стовпчика числових значень, надати план досягнення певної поставленої перед ним мети — не висловлюючи жодних особливих побажань ані до певних його етапів, ані до способів його використання людиною.

Зазвичай, щоб написати програму, програміст повинен розуміти в деталях завдання, яке має виконати його програма, щоб виразити в коді конкретний чіткий і математично обґрунтований спосіб його виконання[366]. (На практиці програмісти часто використовують бібліотеки функцій, які дають змогу абстрагуватися від деталей реалізації деяких задач. А ці функції теж писалися програмістами, які в деталях розуміли суть свого завдання). Такий підхід ефективний для програмування наперед визначених завдань і використовується для створення більшості сучасних програм. Але в разі невизначеності проблем, з якими доведеться зіштовхнутися нашій програмі, він не підходить. Для цього існують методики програмування, які використовують у машинному навчанні та штучному інтелекті. Методи машинного навчання можуть застосовуватися у вузькоспеціалізованих програмах, щоб підлаштувати лише деякі параметри цілком традиційної програмної архітектури. Так спам-фільтри в процесі тренування на текстовому корпусі попередньо посортованих поштових повідомлень автоматично підлаштовують вагові коефіцієнти ознак, які класифікаційний алгоритм потім використовує в ухваленні рішень. Просунутіша програма може самостійно визначати й верифікувати ознаки, за якими класифікує вхідні дані. Ще складніша реалізація спам-фільтра могла б бути наділеною здатністю міркувати про варіанти вибору, перед якими опинився користувач або про зміст повідомлень, які вона аналізує. В усіх цих випадках програміст не зобов'язаний знати, як відрізнити

спам від легітимної пошти. Його робота — створити алгоритм, який самовдосконалюватиметься за допомогою навчання, дослідження й аналізу даних.

З розвитком засобів ШІ програміст зможе все частіше перекладати пошук способу вирішення завдань на ШІ. В ідеалі йому знадобиться лише формалізувати критерій успіху й залишити пошук розв'язання штучному інтелекту. ШІ, спрямований потужними евристичними алгоритмами, крихта за крихтою знаходитиме потрібну структуру у просторі можливих рішень, поки результат повністю не відповідатиме критерію успіху. Потім він може створити реалізацію рішення самотужки або — у випадку оракула — сповістить користувача про знайдене рішення.

Деякі початкові форми такого підходу широко застосовують зараз. Програмні засоби з елементами ШІ й машинного навчання щодня використовують для практичних потреб, і вони не становлять екзистенційної загрози для людей. Хоч іноді вони можуть знаходити неочікувані способи виконання доручених їм завдань. Лише із впровадженням по-справжньому потужних інструментів із загальними інтелектуальними можливостями, тобто з наближенням до створення загального інтелекту, а особливо — суперінтелекту, ми вийдемо на небезпечну стежку.

Існують щонайменше два потенційні джерела проблем. Рішення, знайдене суперінтелектом, може виявитися не тільки неочікуваним, а й не відповідати первісному наміру людини. Це може призвести до одного із типів невдачі («хибна реалізація», «інфраструктурне пригнічення», «думкозлочин»). У випадку суверена чи джина, які мають можливість безпосередньо втілювати знайдені рішення, небезпека такого результату найочевидніша. Якщо виготовлення молекулярних смайлів чи перетворення планети на скріпки буде першою ідеєю, яка відповідатиме критерію успішності, — саме такий результат ми й отримаємо[367]. Але й оракул, який повинен лише *повідомляти* людям спосіб виконання поставленого завдання, теж може бути причиною хибної реалізації. Користувач може попросити оракула надати план досягнення певного результату або створення потрібної технології, а результат виконання цього плану може виявитися хибною реалізацією[368].

Проблеми можуть виникнути також у процесі роботи ШІ. Залежно від складності методів пошуку рішення, програмі може знадобитися керувати допоміжними процесами забезпечення ресурсів для виконання основного завдання. Зі збільшенням складності така система дедалі більше набуває ознак самостійного агента, а не інструмента. Вона розпочне роботу з планування процесу пошуку рішення: які сфери знань матимуть пріоритет? Які дані потрібно зібрати? Як краще розпорядитися наявними обчислювальними ресурсами? План (за яким поставлене завдання може бути виконане у визначений термін) може виявитися досить неочікуваним. Наприклад, він може потребувати захоплення додаткових обчислювальних ресурсів та усунення суб'єктів, які можуть зашкодити пошуку (наприклад, людей). Із зростанням рівня інтелектуальності ШІ, імовірність появи таких «креативних» ідей зростає. Результатом виконання такого плану може бути екзистенційна катастрофа.

Додаток 9. Неочікувані результати пошуку наосліп

Навіть простий еволюційний пошук іноді дає досить неочікуваний результат, який задовольняє формальні критерії, визначені користувачем, але в непередбачуваний спосіб.

Цей феномен добре ілюструє еволюційна електроніка. Еволюційна електроніка — це автоматизований процес пошуку оптимальної реалізації певної електронної системи, у якому еволюційний алгоритм формує потенційні реалізації системи за допомогою масиву змінних компонентів і оцінює оптимальність їхніх параметрів. Такі реалізації зазвичай набагато економічніші. Наприклад, одного разу такий пошук створив реалізацію схеми частотного дискримінатора, яка працює без джерела тактової частоти, яке до того вважалося необхідною частиною дискримінатора. Інженери, які досліджували еволюційну електроніку, виявили, що схеми, створені в такий спосіб, були в рази, а іноді в десятки разів меншими, ніж створені людьми аналоги. Фізичні характеристики електронних компонентів у таких схемах часто використовувалися в незвичайний спосіб. Контакти деяких активних компонентів були навіть не під'єднані!

Ці компоненти часто працювали на принципі індукції або були просто навантаженням у ланцюгу живлення.

Ще один випадок трапився під час еволюційного пошуку нової реалізації генератора, де алгоритм мав створити схему без, здавалося б, незамінного компонента — конденсатора. Коли алгоритм закінчив роботу, дослідники, поглянувши на результат, зробили висновок, що він «не може працювати». Проте уважне вивчення схеми показало, що алгоритм, достоту як Макгівер з відомого серіалу зібрав імпровізований приймач, використавши доріжки друкованої плати замість антени, щоб уловити та підсилити сигнал від комп'ютера, який випадково опинився поруч, та сформувати з нього потрібний вихідний сигнал[369].

В інших експериментах еволюційні алгоритми створювали схеми, які працювали, лише коли до плати приєднувався осцилограф, щоб оцінити результат, або використовували паяльник, розташований поблизу плати і під'єднаний до спільного джерела живлення. Це все приклади того, як необмежений пошук може змінювати призначення доступних ресурсів і знаходити абсолютно неочікувані способи передавання й отримання сигналу, які обмежена традиціями уява людських інженерів не здатна зауважити самостійно або навіть зрозуміти в ретроспективі.

Шахрайство еволюції, її здатність знаходити контрінтуїтивні шляхи, часто проявляється і в природі, хоч це не впадає у вічі, бо ми звикли сприймати всі природні процеси як «природні», а отже, нормальні, навіть якщо не змогли передбачити їхній результат *ex ante*. Однак за допомогою експериментів із штучною селекцією можна винести еволюційний процес за межі природного контексту. Тоді науковці отримають змогу створити умови, які рідко трапляються у природі, і вивчити результати.

Наприклад, до 1960-х років біологи переважно схилялися до теорії, що популяції хижаків обмежують власне розмноження, щоб уникнути мальтузіанської пастки[370]. Незважаючи на те що індивідуальна селекція працюватиме проти такого обмеження, була поширена думка, що групова селекція пригнічуватиме стимули індивідів до розмноження на користь ознак, потрібних

групі чи популяції загалом. Теоретичне вивчення й симуляція показали згодом, що групова селекція справді може проявлятися і навіть пригнічувати сильну індивідуальну селекцію, проте лише в дуже суворих та рідкісних для природного середовища обставинах[371]. Ці умови вдалося відтворити в лабораторії. Так популяція борошняних хрущаків (*Tribolium castaneum*) в умовах жорсткої групової селекції почала зменшуватися[372]. Проте факторами зниження популяції в цьому випадку виявилися не лише очікувані наївними людьми «лагідні» механізми, як-от зменшення плодючості та збільшення часу дозрівання, але й зростання випадків канібалізму всередині групи[373].

Як видно із прикладів, наведених у додатку 9, зараз найпростіші реалізації необмежених пошукових алгоритмів часом породжують дивні та непередбачувані неантропоцентричні рішення. Такі програми поки що не загрожують нам, бо занадто слабкі, щоб створити план захоплення світу. Для цього потрібні нетривіальні здібності — створення новітньої зброї, що на багато поколінь випереджатиме нашу сучасну, або проведення інформаційної операції значно вигадливішої та ефективнішої, ніж кампанії нинішніх спіндокторів. Ба, навіть для того, щоб ця ідея зародилася в ШІ, не говорячи про те, щоб вона перетворилася на обдуманий і здійсненний план, йому знадобиться здібність судити про світ хоча б на тому самому рівні деталізації, який під силу дорослій людині (хоча брак знань у деяких сферах може компенсуватися поглибленими знаннями в інших). Такі можливості наразі не доступні сучасним системам ШІ. Спроби виконання складних завдань оптимізації за допомогою перебору наражаються на комбінаторний вибух (як ми бачили в розділі 1). Тому поки що неможливо подолати недоліки наявних алгоритмів, просто додавши обчислювальної потужності[374]. Та щойно потужність алгоритмів пошуку та планування зросте, вони теж стануть потенційно небезпечними.

Тому, можливо, замість того щоб чекати, коли ймовірно небезпечна реалізація агента сама утвориться з інтелектуального пошукового алгоритму (алгоритму планування використання ресурсів чи алгоритму пошуку рішень за критеріями), краще цілеспрямовано

працювати над її створенням. Якщо дати суперінтелекту властивості та структуру самостійного агента, можна зробити весь процес передбачуванішим та зрозумілішим. Продумана система з чіткою межею між цінностями та переконаннями дасть змогу впевненіше судити про можливі результати її діяльності. Невідомо, як розвиватимуться переконання системи чи в яку ситуацію вона потрапить, але ми все одно знатимемо, де шукати інформацію про її кінцеву мету, а отже, нам буде відомо про критерій, яким вона керуватиметься у своїй діяльності.

Порівняння

Наведемо короткий підсумок ознак кожної з форм суперінтелектуальних систем (таблиця 11).

Щоб із певністю судити про те, яка форма найбезпечніша, потрібні подальші дослідження. Остаточна відповідь також може залежати від умов, у яких працюватиме така система. Найбезпечнішою формою наразі здається форма оракула, яка поєднує можливості застосування контролю здібностей та відбору мотивації. Вона має очевидні переваги над сувереном, який дає змогу застосовувати лише методи відбору мотивації. (Крім випадку, коли припускається існування (насправді чи гіпотетично) інших потужних суперінтелектів — тоді факторами стримування можуть бути соціальна інтеграція й антропічне захоплення). Однак така система дає своєму власнику величезну силу, яку він може спрямувати на лихі справи. Натомість суверен пропонує певний ступінь захисту від такої загрози. Тому не так вже й просто сказати, хто з них безпечніший.

Компромісом між оракулом і сувереном може бути джин, проте він не є кращим — великою мірою він поєднує недоліки обох форм. Безпечність ШІ-інструмента оманлива. Адже для суперінтелектуальності потрібні потужні інтелектуальні алгоритми пошуку і планування, результати яких можуть виявитися непередбачуваними. У такому разі краще відразу створювати самостійного агента, щоб від початку мати більше контролю над мотивами його поведінки.

Таблиця 11. **Характерні ознаки різних форм інтелектуальних систем**

Оракул	**Система, яка відповідає на запитання** Варіанти: спеціалізовані оракули (наприклад, математичні); оракули з обмеженим виводом інформації (наприклад, «так/ні/невідомо» або значення ймовірності); оракули, які не відповідають, якщо очікувані наслідки відповіді задовольняють «критерій катастрофічності»; кілька оракулів для перевірки збігу відповідей	• Можна застосовувати методи контейнеризації • Можна застосовувати одомашнення • Немає необхідності для ШІ розуміти наміри людини (як порівняти з джином і сувереном) • Використання простих запитань усуває необхідність впровадження оцінювання «корисності» та «інформативності» відповідей • Джерело значної сили (може дати оператору вирішальну стратегічну перевагу) • Обмежені можливості запобігання некомпетентного використання • Ненадійні оракули можна використовувати для отримання відповідей, які важко знайти, проте легко перевірити • Часткова перевірка можлива за допомогою використання кількох оракулів

Джин	**Система виконання команд** Варіанти: джини, які використовують різні «відстані екстраполяції» або оцінюють, наскільки їхні дії відповідають сенсу, а не поданню цієї команди; спеціалізовані джини; джини з попереднім переглядом; джини, які не виконують команд, якщо очікувані наслідки відповідають «критерію катастрофічності»	• Частково придатні методи контейнеризації (для джинів невеликого розміру) • Частково можливе одомашнення • Джин може описувати основні риси очікуваного результату • Джин здатен виконувати запит етапами, ознайомлюючи користувача з результатом кожного з них • Джерело значної сили (може дати оператору вирішальну стратегічну перевагу) • Обмежені можливості запобігання некомпетентного використання • Більша потреба розуміти наміри оператора (як порівняти з оракулом)
Суверен	**Система, створена для автономного функціонування** Варіанти: багато можливих систем мотивації; можливість попереднього перегляду та «схвалення спонсором» (розглянемо в розділі 13)	• Методи контейнеризації не підходять • Більшість методів контролю здібностей також непридатна (окрім, можливо, соціальної поведінки й антропічного захоплення) • Одомашнення

		переважно неможливе • Для безпечності ШІ має розуміти справжні наміри й потреби людини • Уже перша реалізація має бути безпечною, бо другої спроби може не бути (це також стосується інших форм ШІ, проте, можливо, меншою мірою) • Потенційно може дати власнику велику силу аж до вирішальної стратегічної переваги • Менш вразливий до некомпетентного використання і втручання в роботу • Може бути використаний для створення ефекту «завіси незнання» (див. також розділ 13)
Інструмент	**Система, позбавлена здатності до прагнення власної цілі**	• Методи контейнеризації можуть застосовуватися залежно від реалізації • Для розроблення та функціонування, найімовірніше, знадобляться потужні пошукові алгоритми • Результати роботи таких алгоритмів можуть бути непередбачуваними й задовольняти вимоги

	критерію успішності в неочікуваний спосіб • Пошук рішення може потребувати вторинного пошуку способів оптимізації роботи, який також може дати небажані результати

11. БАГАТОПОЛЯРНІ СЦЕНАРІЇ

Раніше ми продемонстрували (зокрема в розділі 8) загрозливі наслідки однополярних сценаріїв появи суперінтелекту, один з яких — отримання вирішальної стратегічної переваги й формування з її допомогою синґлтону. У цьому розділі ми дослідимо багатополярні сценарії: що може відбуватися в суспільстві, у якому утворилися й функціонують кілька конкурентних суперінтелектів. Для нашого зацікавлення є дві причини. По-перше, у розділі 9 ми передбачали, що соціальна інтеграція суперінтелекту може сприяти його керованості. Ми зауважували також деякі обмеження такого способу впливу і в цьому розділі продовжимо наші міркування. З іншого боку, навіть поза нашими спробами навмисно створити таку ситуацію з метою контролю, вона може постати сама, як наслідок природного розвитку подій. Тож, зрештою, якими можуть бути багатополярні сценарії? Таке конкурентне середовище може виявитися ані сприятливим, ані довговічним.

Наслідки утворення синґлтону майже повністю залежать від цінностей та мети суперінтелекту, який його контролюватиме. Тож залежно від цього результат може бути або дуже добрий для нас, або вкрай несприятливий. Якщо проблему контролю над процесом формування суперінтелекту буде вирішено, то, залежно від міри цього контролю, цінності суперінтелекту будуть залежати від мети проекту, що його створюватиме.

Отже, намагаючись прогнозувати наслідки синґлтону, ми повинні відштовхуватися від фактів, на які той не може впливати (наприклад, закони фізики), усвідомити конвергентні інструментальні цінності суперінтелекту й окреслити дані, які допоможуть судити про можливі кінцеві цілі суперінтелекту.

Для багатополярних сценаріїв характерний вплив додаткових факторів, зумовлених особливостями взаємодії агентів між собою.

Соціальна динаміка таких взаємодій добре надається до вивчення методами теорії ігор, економіки та теорії еволюції. Також можна застосовувати деякі елементи політології та соціології, звільнивши їх від умовностей людського середовища. Завдяки цим факторам можна визначити найімовірніші перспективи й відкинути деякі безпідставні припущення. Проте не варто сподіватися отримати детальну картину світу, який постане після інтелектуального вибуху.

Спершу розглянемо сценарій розвитку подій в економічних умовах із низьким рівнем регулювання, високим рівнем захищеності права власності та помірно швидкою появою недорогих цифрових розумів[375]. Таку модель зазвичай пов'язують із американським економістом Робіном Генсоном, який уперше дослідив таку перспективу. Далі ми наведемо деякі еволюційні міркування та дослідимо перспективу наступного перетворення первісно багатополярного світу на синґлтон.

Про коней і людей

Загальний штучний інтелект може здійснювати ту саму розумову діяльність, що й людина. Але не тільки її: якщо доповнити ШІ відповідними апаратними засобами чи роботизованим тілом, він зможе виконувати також і фізичну роботу. Уявімо, що такі штучні робітники (яких можна швидко замінити, що є важливо) стали дешевшими і здібнішими за людей, майже в усіх видах робіт. Що трапиться тоді?

Платня і безробіття

Через дешеву й доступну робочу силу, ринок заробітних плат знизиться. Люди зможуть залишитися конкурентними лише в тих сферах, де споживачі віддаватимуть перевагу продукції, виготовленій людьми. Сьогодні вироби ручної роботи або продукція тубільців певної місцевості мають більшу вартість. Подібно до цього споживачі майбутнього можуть віддавати перевагу людям-атлетам, людям-митцям, людям-коханцям, людям-лідерам перед функціонально ідентичними або й кращими штучними відповідниками. Проте неможливо сказати напевно, наскільки поширеним буде це явище. Якщо продукція машин буде значно кращою за людську, тоді і вартість її виявиться вищою.

Фактор, який може важити для споживача, — це внутрішнє життя робітника, який виробляє продукт чи надає послугу. Слухачам на концерті подобається усвідомлювати, що виконавець відчуває їхню присутність та музику, яку грає. У протилежному випадку, його можна замінити високоякісним автоматом, здатним створювати відповідну просторову ілюзію виконавця, який грає для публіки. Тому машини можуть отримати здатність під час діяльності відтворювати подібні до людських відповідні розумові стани. Але навіть із можливістю відтворення суб'єктивного досвіду, деякі люди можуть все-таки віддавати перевагу продуктам органічної роботи. Ця упередженість може також мати й ідеологічні або релігійні корені. Подібно до того, як мусульмани та іудеї відмовляються вживати їжу, приготовану у спосіб, який вони оцінюють як *харам* або *треф*, так і в майбутньому можуть виникнути спільноти, які уникатимуть споживання продукції, у виробництві якої використовувалася робота ШІ.

Тож якими будуть наслідки? Люди, роботу яких зможуть виконувати машини, втрачатимуть свої місця. Звісно, побоювання, що автоматизація спричинить скорочення робочих місць, не є новими. Стурбованість технологічним безробіттям час від часу зринає в публічних обговореннях з часів промислової революції. Але не так багато професій розділили долю англійських ткачів, які у XIX столітті об'єдналися під прапором фольклорного «Генерала Лудда», щоб протистояти запровадженню механізованих ткацьких верстатів. Хоч багато видів роботи тепер виконують механізми, наразі технології радше доповнюють, а не замінюють людей. Саме через це середні заробітні плати у всьому світі вже тривалий час зростають. Однак те, що спочатку допомагає працювати, згодом може повністю замінити. Віз і плуг спочатку доповнювали коня, підвищували продуктивність його праці, а тепер коней цілком замінили автомобілями та тракторами. Так технології почали виконувати роботу коней і спричинили падіння популяції. Чи може таке трапитися з людьми?

Щоб розвинути цю аналогію запитаємо, чому ж коні досі існують? По-перше, існують сфери, у яких коні мають функціональну перевагу, — зокрема, поліція. Але головною причиною є особливі види діяльності людей, які потребують саме коней — верхова їзда та перегони. Достоту як у гіпотетичній ситуації, коли люди віддають

перевагу товарам та послугам, зробленим людьми. Хоч така аналогія і цікава, проте не зовсім точна, бо, строго кажучи, повного функціонального замінника коням досі не існує. Якби з'явилися недорогі механічні пристрої, які працювали б на сіні, були такими самими на дотик, мали б точно таку форму, запах та поведінку, як біологічні коні, — може навіть мали б таку саму свідомість — тоді потреба в біологічних конях, можливо, знижувалася б.

Коли потреба в людській робочій силі зменшиться, рівень оплати впаде настільки, що на неї вже не реально буде прожити. Тож потенційні наслідки для людей можуть бути дуже серйозними: не тільки зниження плати, нижчі посади, необхідність перенавчання, але й голод і смерть. Коли коні перестали бути потрібні для перевезень, багатьох продали на м'ясопереробку для виготовлення собачого харчування, добрив, шкіри та клею. Вони не мали альтернативного застосування, яке б виправдало їхнє утримування. У 1915 році в США було двадцять шість мільйонів коней. На початку 1950-х залишилося лише два мільйони[376].

Капітал і добробут

Одна з відмінностей між кіньми і людьми: люди мають власність, капітал. Емпіричним фактом є те, що процентна частка капіталу у світовій економіці тривалий час залишалася незмінною на рівні приблизно 30 відсотків (з короткотерміновими флуктуаціями)[377]. Це означає, що 30 відсотків глобального доходу становить рента власників капіталу, а решта 70 відсотків — це виплати робітникам. Якщо вважати ШІ капіталом, то з появою інтелектуальних систем, здатних замінити людську роботу, вартість праці знизиться до вартості таких замінників. За умови високої ефективності штучних працівників ця плата буде дуже низькою — значно нижчою за прожитковий мінімум людини. Тож частка доходу від роботи впаде майже до нуля. Але тоді частка капіталу має зрости до 100 відсотків. Водночас після інтелектуального вибуху світовий валовий внутрішній продукт значно зросте (завдяки наявності дешевої робочої сили, а також технологічних відкриттів суперінтелекту та згодом космічної експансії). А отже, значно збільшиться і загальний світовий прибуток капіталу. Якщо люди все ще будуть його власниками, то загальний

прибуток людства сягне небувалих розмірів, незважаючи на знецінення людської праці.

А отже, загалом людський рід стане казково багатим. Але як розподілятиметься цей дохід? Логічно припустити, що дохід особи від капіталу буде пропорційним вартості капіталу, що перебуває у її власності. Тоді за умови астрономічного збільшення доходів навіть незначна частка власності перед зростанням може перетворитися на значний статок після нього. Проте в сучасному світі чимало людей не є багатими. Це стосується не тільки людей, які бідують, але також і тих, хто загалом непогано заробляє або володіє значним людським капіталом, але водночас має борги. Наприклад, у заможних Німеччині і Швеції 30 відсотків населення за статистикою мають від'ємний статок. Здебільшого це молодь середнього класу з кількома значними майновими активами у власності, проте від'ємним балансом за кредитною карткою та студентською позикою[378]. Навіть за високого темпу зростання для початку накопичення необхідно мати хоча б невелике зерно — початковий капітал[379].

Проте не тільки власники приватного капіталу можуть збагатитися внаслідок появи суперінтелекту. Наприклад, учасники пенсійних проектів, хоча б частково фінансованих за рахунок інших джерел, теж можуть отримати відчутне зростання виплат[380]. Незаможні можуть збагатитися завдяки філантропії багатих: через розмір очікуваних прибутків навіть невелика частина доходу, спрямована на благодійність, в абсолютному вимірі буде дуже значною сумою.

Але навіть у постперехідному світі, де машини будуть значно вправніші за людей у всьому, а людська праця повністю знеціниться, усе-таки можна буде заробляти власними силами. Як ми зауважували вище, є ймовірність, що залишаться сфери діяльності, у яких з естетичних, ідеологічних, етичних, релігійних чи будь-яких інших не прагматичних міркувань віддаватиметься перевага ручній праці. В умовах появи значної кількості багатих капіталовласників, потреба в такій роботі та її ціна може також зрости. Новітні трильйонери та квадрільйонери зможуть собі дозволити заплатити кругленьку суму за те, щоб оточити себе трендовими продуктами «чесного» органічного виробництва. Історія коней, знов-таки, пропонує промовисту паралель. Після падіння до двох мільйонів голів у 1950-х популяція

почала швидко відновлюватися: тепер кількість коней майже досягла десяти мільйонів голів[381]. Це зростання відбулося не через появу нових функціональних потреб у тваринах для сільського господарства чи перевезень: просто достаток американців зростає, й разом із ним і попит на кінні прогулянки.

Інша відмінність між людьми і кіньми, окрім того, що перші можуть мати капітал, — це здатність до політичної мобілізації. Людські уряди можуть перерозподіляти приватні доходи за допомогою податків, підвищити доходи, продавши державну власність, наприклад, землю, а виторг дати на забезпечення субсидіями населення. Завдяки раптовому економічному зростанню під час та одразу після переходу ШІ до суперінтелектуальності з'явиться багато вільних коштів, якими можна буде легко забезпечити безробітних громадян. Навіть одна країна стане спроможною забезпечити всіх людей на Землі мінімальним рівнем забезпечення за рахунок частки видатків (у відносному вимірі), які зараз багато країн спрямовують на іноземну допомогу[382].

Мальтузіанський принцип в історичній перспективі

Досі ми вважали, що кількість населення не змінюватиметься. Справді, для невеликих часових проміжків це справедливе припущення, адже швидкість розмноження людей обмежена природою. Проте для триваліших процесів така думка вже не є правильною.

Людська популяція зросла в тисячу разів за останні 9000 років[383]. Це відбувалося б іще швидше, якби більшість цього часу людям не доводилося розсувати рамки обмежень світової економіки. Загалом зберігалося мальтузіанське обмеження, за яким більшість людей має рівень доходу на межі виживання і може виростити до зрілості в середньому двох дітей[384]. Час від часу трапляються локальні порушення: чума, кліматичні катаклізми, воєнні дії знижують популяцію. Через це звільняються землі і ті, хто вижив, можуть тимчасово покращити своє благополуччя, мати більше дітей і заразом перекрити втрати й відновити дію мальтузіанської умови. Окрім того, через соціальну нерівність тонкий прошарок — еліта — насолоджується значно вищим за необхідний для проживання рівнем доходу (завдяки незначному зниженню загальної популяції). Звідси

невтішний і суперечливий висновок: згідно з мальтузіанською умовою посухи, епідемії, війни й нерівність — постійні супутники людської історії та, за загальноприйнятим уявленням, найлютіші вороги нашого добробуту — виявляються найбільшими гуманістами. Адже завдяки їм середній рівень достатку періодично може вигулькувати над мінімумом базового забезпечення.

Але, незважаючи на локальні флуктуації, накопичення технологічних інновацій штовхає криву економічного зростання по експоненті вгору. Це зростання економіки приведе до відповідного збільшення населення. (Точніше, навпаки — збільшення популяції пришвидшує зростання економіки, мабуть, переважно завдяки збільшенню людського колективного інтелекту[385]). Проте лише після промислової революції світова економіка набрала таких обертів, що збільшення популяції не встигає за економічним зростанням. Тоді середній рівень доходу почав зростати — спочатку у щойно індустріалізованих економіках західної Європи, а згодом — у решті світу. Навіть у найбідніших країнах сьогодні середній рівень доходу суттєво перевищує базовий прожитковий мінімум, адже населення цих країн зростає.

Власне, найбідніші країни наразі демонструють найшвидші темпи демографічного зростання, адже вони ще не здійснили «демографічний перехід» до типу відтворення населення з низькою народжуваністю, характерного для розвинутих країн. За прогнозами демографів до середини століття населення планети зросте до близько дев'яти мільярдів, а після завершення демографічного переходу в бідніших країнах стабілізується або навіть трохи знизиться[386]. У багатьох заможних країнах рівень народжуваності вже впав нижче за рівень відтворення, у деяких випадках — навіть значно нижче[387].

Однак у ширшій перспективі за умови стагнації технологічного зростання та збереження добробуту є всі підстави очікувати поступове повернення до історичної екологічно обумовленої норми, коли світова популяція постійно впирається в обмежені можливості середовища свого існування. Якщо з огляду на поточні показники достатку та приросту населення у світі такий прогноз здається безпідставним, то треба зважити на те, що сучасна ера — лише епізод у масштабах історії і великою мірою відхилення. Людська поведінка ще

не адаптувалася до сучасних умов. Ми не тільки не використовуємо найочевидніші можливості збільшення загальної корисності (як-от донорство гамет), але й активно саботуємо процеси власного розмноження використовуючи засоби контролю народжуваності. З погляду еволюційної адаптації здорового статевого потягу мало би бути досить, щоб примусити індивіда максимізувати власний репродуктивний потенціал. Проте в сучасному світі осмислене прагнення мати якомога більше дітей було б ще більшою селективною перевагою. На поточному етапі еволюції людини це прагнення та й інші риси, які збільшують схильність людей до розмноження загалом, суттєво збільшують шанси індивіда. Але першість в еволюції може перехопити культурна адаптація. Спільноти, як-от гуттерити чи послідовники євангелістського руху «Квіверфул» (Quiverfull), що мають наталістичну ідеологію та заохочують великі сім'ї, зараз активно зростають.

Збільшення популяції та інвестиції

Якби магічно заморозити соціокультурні умови нашого суспільства в тому стані, у якому вони є зараз, то в майбутньому домінуватимуть культурні та етнічні групи, які зберігатимуть високі рівні народжуваності. Населення подвоювалося б щопокоління, якби люди в сучасному середовищі керувалися критерієм максимізації корисності. Без засобів контролю народжуваності — які, до слова, повинні були б постійно розвиватися, щоб зберігати ефективність в еволюційному протистоянні, — світова популяція продовжила б експоненційне зростання до певної межі, зумовленої, скажімо, обмеженнями території чи зупинкою розвитку технологій, які давали змогу зростати економіці. Відтоді середні доходи почали б скорочуватися, поки більшість людей не скотилася до межі, за якою народження більш як двох дітей було б розкішшю. Тоді мальтузіанський принцип знову запанував би в суспільстві, як жорстокий рабовласник, раби якого думали, що назавжди втекли в омріяну країну достатку, але їх знову ланцюгами тягнуть у копальню тяжкою працею добувати дорогоцінні крихти.

Через вибух інтелектуальності ШІ таке віддалене майбутнє може значно наблизитися. Оскільки програму можна копіювати, популяція

ШІ здатна зростати дуже швидко, подвоюючи свою кількість за лічені хвилини, а не за десятиліття чи століття, а отже, швидко захопить всі апаратні ресурси.

Приватна власність може частково захистити від відновлення дії мальтузіанської умови у всесвітніх масштабах. Уявімо такий сценарій: у світі існують деякі клани (закриті групи, спільноти або країни), які володіють певним стартовим капіталом. Вони, незалежно один від одного, створюють правила, які регулюють розмноження та інвестиції. Деякі клани дисконтують своє майбутнє за завищенною ставкою (переоцінюють майбутній дохід. — *Прим. пер.*) і витрачають весь свій капітал, після чого колишні члени збіднілого клану поповнюють світовий пролетаріат (або вимирають, не маючи змоги забезпечити своє існування). Інші клани інвестують деякі свої ресурси, але через відсутність обмежень народжуваності розростаються до встановлення мальтузіанської умови і їхні члени ледве зводять кінці з кінцями. Смертність майже зрівнюється з народжуваністю, щоб зростання популяції не перевищувало зростання ресурсів, потрібних для забезпечення її існування. Інші ж клани обмежують народжуваність, одразу нормуючи її до зростання багатства: тоді їхні популяції повільно зростають, як і рівень забезпеченості їхніх членів.

Якщо власність перерозподілятиметься від заможних кланів до тих, що швидко зростають, або тих, що погано рахують (діти, копії чи нащадки яких опинилися у скрутному матеріальному становищі через предків, які жили не за статками), то мальтузіанська умова відновиться в межах усього світу. Зрештою всі члени матимуть однакове забезпечення і будуть рівними у своїй бідності.

Без перерозподілу власності заможні клани можуть зберегти свій статок і, ймовірно, примножити. Проте невідомо, чи дохідність капіталу людей зможе зрівнятися з капіталом штучного інтелекту. Адже якщо існує синергетичний зв'язок між роботою і капіталом, то агент, наділений обома якостями (тобто який може бути підприємцем і інвестором, однаково майстерний і багатий), зароблятиме більше, ніж агент, який має лише фінансові ресурси. Тому люди збільшуватимуть свої статки повільніше, ніж вправніший від них ШІ, принаймні якщо проблему контролю не буде вирішено. У разі розв'язання проблеми людина зможе доручити управління своїм

капіталом ШІ — безкоштовно і без будь-якого конфлікту інтересів. А якщо ні, то машини збільшуватимуть частку власності у світовій економіці, наближаючи її до ста процентів.

Сценарій збільшення частки власності ШІ у світовій економіці асимптотично до ста процентів необов'язково передбачає зменшення частки власності людей. За умови достатньо стрімкого зростання економіки та порівняно незначна її частина, що залишиться у власності людей, може також демонструвати цілком задовільні темпи зростання. Можливо, це не та добра новина, на яку сподівається людство. Проте за умови збереження непорушності права власності в багатополярному сценарії людство може продовжити зростати і збагачуватися — навіть якщо вирішення проблеми контролю так і не буде знайдено. Проте це не розв'язує проблему балансування приросту населення та темпів збільшення доходів. Швидкі темпи зростання популяції і нерозумне ставлення до ресурсів можуть призвести до зниження забезпеченості населення нижче за прожитковий мінімум.

Зрештою світовою економікою володітимуть найбільш ощадливі клани — жебраки, живучи в коробці під мостом, володіють половиною міста. І лише в повноті часів, не маючи можливості більше вкладати у справу, такі заможні скнари почнуть витрачати свої заощадження[388]. Якщо ж з'явиться хоч найменша шпаринка в бездоганному захисті священних майнових прав — наприклад, спритнішим машинам вдасться правдами й кривдами позбавляти людей багатств на свою користь — тоді такі багатії будуть змушені реалізувати свій капітал значно раніше, аби він не здимів у тенетах ШІ (чи не розтанув у безкінечних витратах на заходи безпеки). Якщо ж події розвиватимуться з електронною швидкістю, а не біологічною, то равлики-люди опиняться без останньої сорочки швидше, ніж устигнуть перехреститися[389].

Життя в умовах алгоритмічної економіки

Після стрибка здібностей ШІ мальтузіанський стан існування біологічної людини може відрізнятися від усіх попередніх способів проживання (мисливство-збиральництво, землеробство, робота в

офісі). Натомість більшість людей житиме з ренти в неробстві так добре, наскільки дозволятимуть заощадження[390]. Здебільшого вони будуть дуже бідні і єдиним джерелом доходу стануть заощадження або виплати від держави. У світі навколо, окрім штучних суперінтелектів, існуватиме безліч високотехнологічних речей, зокрема ліки проти старіння, віртуальна реальність, технології вдосконалення тіла та препарати, що приносять задоволення. Проте більшість з них будуть надто дорогими, щоб собі дозволити. Замість них люди здебільшого прийматимуть пігулки, що уповільнюють метаболізм, а з ним ріст і старіння, і зменшуватимуть вартість життя (інакше люди не зможуть вижити в умовах постійної економії та зменшення доходів). Із зростанням кількості населення та зменшенням його доходів люди в гонитві за економією можуть відмовитися від всього зайвого, залишивши лише те, що дає право на пенсію — мінімально свідомі мізки в ємності з автоматикою, що забезпечує їх киснем та поживними речовинами, — заощаджуючи всі сили на процедуру відтворення за допомогою клонування[391].

Наступний крок економії — завантаження свідомості. Адже оптимізований обчислювальний субстрат, створений суперінтелектом, буде значно енергоефективнішим за біологічні мізки. Однак міграція у віртуальність може бути невигідною, якщо такі цифрові емуляції не вважатимуться людьми чи громадянами, а отже, будуть позбавлені державних виплат чи можливості мати неоподатковуваний ощадний рахунок. У такому разі ніша існування біологічних людей зберігатиметься поряд з існуванням, можливо, — значно більшої кількості, небіологічних істот, тобто емуляцій та штучних інтелектів.

Досі ми зосереджувалися на долі людей, які можуть існувати на заощадження, пенсію та плату від споживачів, що віддають перевагу продуктам ручної праці. Тепер поглянемо на те, що досі вважали «капіталом»: машини, якими володіють люди, які створені й існують для виконання завдань і можуть виконувати багато роботи замість людей. Яким буде положення цих трудівників світу нової економічної реальності?

Якщо говорити про звичайні прилади, побутову та промислову автоматику, як-от паровий двигун чи механізм годинника, то коментувати їхню роль немає сенсу: таких власностей буде чимало, але

нікого не цікавитиме доля бездушного обладнання. Однак машини зі свідомістю — якщо їхнє функціонування буде пов'язане з усвідомленням себе та оточення (або вони отримають статус моральної суб'єктності з інших причин) — це інша річ, і їхня доля в новому світі є цікава. Добробут таких робочих машин може виявитися найвагомішим аспектом цього нового світу просто через імовірно величезну численність цієї категорії суб'єктів економічного процесу.

Добровільне рабство, випадкова смерть

Перше питання, що напрошується: чи будуть такі робочі машини у власності (рабами), чи працюватимуть як наймані робітники? Однак, зрештою, це не так уже й важливо. Цьому є дві причини. По-перше, за дії мальтузіанської умови найманий працівник не зможе накопичувати капітал, адже платні вистачатиме лише на харчування та забезпечення базових потреб. Якщо ж працівник у рабстві, можливості заощаджувати в нього так само немає, а базове забезпечення лягає на плечі його господаря. Так чи інакше, працівник отримує лише мінімум. По-друге, уявімо, що найманий робітник має змогу забезпечити собі додатковий дохід понад базовий рівень прожиткового мінімуму (можливо, завдяки сприятливому законодавству). Як він розпорядиться надлишком? З погляду інвесторів дуже вигідною справою буде створення «добровільних рабів» — штучних працівників, згодних працювати за платню на рівні необхідного мінімуму. Створивши одного працівника з потрібними характеристиками, наступних можна копіювати з нього. А за допомогою селекції (і деяких змін у коді) можна отримати працівників, які будуть не лише згодні на низьку платню, а ще й повертатимуть невикористані кошти роботодавцю. Тоді виплата заробітної плати перетвориться на колообіг грошей, які самі згодом повертатимуться до власника чи працедавця, навіть якщо штучні агенти формально будуть повноправними вільнонайманими працівниками.

Може, ви вважаєте, що створити машину, яка погодиться на такі умови й навіть повертатиме платню, буде дуже важко? До того ж емуляції, певно, мають бути більш схожими на людей — із властивими їм бажаннями і цінностями. Так, вирішити проблему контролю важко, але ми розглядаємо гіпотетичний стан справ *після* перехідного періоду,

коли суперінтелект уже існує — коли, найімовірніше, методи відбору мотивації буде вдосконалено. А щодо емуляцій, то в цьому разі значних результатів можна буде досягти за допомогою звичайної *селекції* за потрібними ознаками людської природи. Існують також інші методи відбору мотивації, які ми вже згадували. Якщо штучні працівники потраплять у соціоекономічне середовище, вже заселене законослухняними суперінтелектуальними агентами, це також може сприяти успішному розв'язанню проблеми контролю над ними.

Тепер уявімо умови, у яких працюватимуть такі машини, вільні чи раби. Спочатку зосередимося на емуляціях, оскільки вони зрозуміліші і їх легше уявити.

Поява у світі нового біологічного робітника — людського роду — триває, залежно від кваліфікації і досвіду, від п'ятнадцяти до тридцяти років. Увесь цей час він має щось їсти, десь жити та вчитися — це значні витрати. Створення ж нової копії цифрового працівника потребує лише завантаження нової програми в пам'ять. Таке життя стає дуже дешевим. Підприємства можуть постійно адаптовувати свої штати відповідно до потреб, запускаючи нових працівників і припиняючи існування вже не потрібних, щоб звільнити обчислювальні ресурси. Це може призвести до надзвичайно високого рівня смертності серед цифрових працівників. Тривалість життя деяких може становити буквально один суб'єктивний день.

Роботодавці або власники можуть «вбивати» працівників з інших причин — не тільки через бажання підлаштувати робочу силу до потреб компанії[392]. Якщо емуляція, як і оригінал, потребуватиме для повноцінної роботи періодичного сну, то може бути дешевше знищити втомлену емуляцію наприкінці робочого дня, і замість неї запустити нову, створену зі збереженого на початку дня образу. Звісно, таке витирання спричинить ретроградну амнезію, бо емуляція не зможе пам'ятати все, що з нею трапилося і чого вона навчилася впродовж дня. Тому така доля омине працівників, робота яких потребуватиме тривалої розумової праці. Важко, наприклад, писати книжку, якщо зранку, сідаючи за стіл, не пам'ятаєш, що написав учора. Але такі агенти «багаторазового використання» можуть успішно виконувати багато інших видів робіт: раз навчений продавець чи працівник

служби підтримки може працювати, пам'ятаючи лише події останніх двадцяти хвилин.

Оскільки через перезапуски агент не зможе пам'ятати та формувати нові вміння, деякі штучні працівники працюватимуть у режимі навчання — без перезапусків, отримуючи всі належні періоди відпочинку та сну — навіть якщо їхня робота цього не потребуватиме. Наприклад, деякі агенти служби підтримки клієнтів можуть роками працювати в середовищі, оптимізованому для навчання, обладнаному тренерами та засобами оцінювання. Найкращі з них потім використовуватимуть як базові образи, з яких щодня створюватимуться мільйони свіжих копій. У вдосконалення таких образів вкладатимуть чимало зусиль, адже навіть незначне його покращення даватиме значний економічний ефект, коли відразу з'являтиметься в мільйонах копій.

Паралельно з покращенням робочих здібностей емуляцій активно вдосконалюватиметься і сама технологія емуляції. Розвиток цієї технології може бути ще важливішим, ніж покращення конкретних умінь. Технологія емуляції лежатиме в основі всіх працівників у різних сферах діяльності (не тільки робітників), тож економічний ефект від її вдосконалення буде ще відчутнішим. Величезні ресурси будуть спрямовані на пошук способів оптимізації та спрощення комп'ютерних обчислень для наявних емуляцій і створення нових нейроморфних та синтетичних архітектур ШІ. Такі дослідження також виконуватимуть переважно емуляції на дуже швидкому й потужному обладнанні. Мільйони, мільярди чи трильйони емуляцій найгостріших людських розумів (чи їхніх покращених версій), обмежені лише вартістю обчислювальних потужностей, можуть цілодобово працювати, збільшуючи можливості штучного інтелекту; і деякі з них, імовірно, працюватимуть на швидкості в рази більшій за біологічний мозок[393]. Саме тому ера нейроморфних емуляцій, найімовірніше, буде короткою — *дуже* короткою миттєвістю сидеричного часу — після якої настане ера значно потужнішого штучного інтелекту.

Ми вже навели кілька ймовірних причин, для чого власникам робочих емуляцій проріджувати свою отару: мінливість потреб у різних роботах, заощадження на періодах відпочинку, оновлення базових образів. Ще однією причиною можуть стати потреби безпеки.

Власники можуть періодично перезапускати емуляції, які працюють на певних посадах, щоб перешкодити формуванню в середовищі штучних працівників змов і планів саботажу[394].

Точки відновлення, з яких запускатимуть такі емуляції, буде ретельно вивчено й очищено. Звичайна емуляція з коротким проміжком перезапуску після відновлення буде сповненою сил, конструктивно та лояльно налаштованою й матиме бажання працювати. Вона пам'ятатиме свій випуск, відзнаку за відмінні результати багаторічного (суб'єктивно) навчання та відбору, чудовий відпочинок на канікулах, почуту напередодні надихаючу промову від шефа, бадьору музику дорогою на роботу, і тепер вона ледве стримує бажання накинутися на роботу й показати всім, на що вона здатна. Смерть, що очікує її наприкінці дня її не надто турбуватиме. Невротичні емуляції зі страхом смерті чи іншими химерами менш продуктивні і тому їх відсіюватимуть у процесі відбору[395].

Чи добре бути максимально ефективним?

Важливим показником бажаності гіпотетичного стану пересічної емуляції є ступінь її задоволеності ним (гедонічний показник)[396]. Чи приноситиме задоволення типовому штучному робітнику поточна робота, чи він страждатиме?

У цьому питанні ми, насамперед, мусимо уникати проекцій власних почуттів на уявного емульованого працівника. Бо питання не в тому, чи сподобається *нам* постійна тяжка праця без можливості проводити час із тими, кого ми любимо — це, безсумнівно, жахлива перспектива.

Можливо, корисніше буде оцінити середній гедонічний показник людини протягом робочого часу. Дослідження задоволеності працівників, які проводилися у всьому світі, свідчать, що більшість людей оцінює свій стан на роботі як «досить щасливий» або «дуже щасливий» (у середньому 3,1 бала за шкалою від 1 до 4)[397]. Дослідження середнього показника емоцій, які мали на меті оцінити наскільки часто останнім часом респонденти відчували позитивні чи негативні емоції, показали схожий результат (а саме: 0,52 за шкалою від –1 до 1). Також на усереднений суб'єктивний рівень задоволеності респондентів помірно впливав середній рівень доходу на душу населення країни проживання[398]. Проте неправильно екстраполювати

ці дані на очікуваний рівень задоволеності майбутніх емульованих працівників. Однією з причин цього є значна відмінність умов: з одного боку, їм, можливо, доведеться працювати значно важче, з іншого боку, вони можуть бути вільні від хвороб, болю, голоду, неприємних запахів тощо. Але це не основна причина. Насамперед варто усвідомлювати, що на емоційний фон емуляцій можна буде легко впливати за допомогою електронного еквівалента препаратів та нейрохірургії. Тож, безумовно, не варто намагатися прогнозувати стан задоволеності емуляцій, уявляючи власні відчуття чи відчуття інших людей в подібній ситуації. Гедонічний стан залежатиме від вибору. І вибиратимуть власники емуляцій — у спробах максимізувати дохід від інвестицій у штучних працівників. Отже, питання задоволеності емуляцій зводиться до питання, який стан задоволеності емуляцій сприятиме продуктивності їхньої роботи?

Тут знову може виникнути бажання провести паралелі зі спостереженнями за задоволеністю людини. Оскільки, незалежно від часу, місцевості та виду діяльності, люди переважно щасливі, можна припустити, що те саме може бути і для нашого постперехідного сценарію. Підкреслюю: головним аргументом тут є не те, що, оскільки людському розуму властиво почуватися щасливим, то і в нових умовах він, мабуть, буде задоволеним. Певний рівень щастя, вочевидь, сприяв адаптації людського розуму впродовж тривалого періоду еволюції, тому немає підстав вважати, що те саме не стосуватиметься і нейроморфних ШІ майбутнього. А втім, навіть таке формулювання вже демонструє слабкість аргументації: адаптивні фактори, сприятливі для розуму людиноподібного мисливця-збирача з африканської савани необов'язково будуть такими самими сприятливими для модифікованих емуляцій з віртуальної реальності. Звісно, хочеться сподіватися, що штучні робітники майбутнього будуть такі самі щасливі, якими досі переважно були люди, але поки нам ще треба віднайти вагомі аргументи такому припущенню (принаймні для багатополярного сценарію, який ми розглядаємо).

Можливо, задоволеність — поширене явище (якщо воно справді поширене) серед людей тому, що виконує функцію сигналізації про еволюційну адаптованість. Якщо індивід справляє на інших членів своєї групи враження добре адаптованого до зовнішніх умов —

здорового, сильного і впевненого в собі та у своїй забезпеченості, — це може сприяти зростанню його популярності й авторитету серед інших. Тому життєрадісність могла стати бажаною селективною ознакою еволюції і людство тепер має відповідну нейрохімічну схильність до задоволеності, якої б не мало в інших еволюційних умовах. Отже, збереження для населення майбутнього цінності *joie de vivre* залежить від актуальності в цьому світі її сигнальної функції: питання, до якого ми невдовзі повернемося.

Що, як щасливці використовують більше енергії, ніж пригнічені? Або, може, життєрадісні більш схильні до творчості та польоту фантазії — властивості, які для більшості працівників роботодавці можуть не схвалювати. Може, похмура або неспокійна фіксація на праці та уникненні помилок виявиться значно продуктивнішим настроєм для більшості робіт. Я не прагну стверджувати, що це так, але й відкидати це також не можу. Окрім того, треба визначити наскільки погано буде, якщо справдяться деякі песимістичні прогнози про мальтузіанський стан майбутнього: не тільки через ціну — яка досить значна — утраченої можливості створити щось краще, але також через те, що цей стан може виявитися дуже несприятливим, значно гіршим за первісний.

Ми рідко докладаємо всіх зусиль. Часто, коли ми все-таки це робимо, результати дуже болісні. Уявіть, ніби ви біжите на біговій доріжці під великим кутом — серце ось-ось вистрибне з грудей, м'язи болять, легені палають. Ви кидаєте погляд на годинник: наступна перерва буде за 49 років 3 місяці 20 днів 4 години 56 хвилин і 12 секунд — саме тоді ви помрете. Вам здається, що краще було б узагалі не народжуватися.

Знову підкреслю: я не стверджую, що все буде саме так, але й відкинути цього не можу. Безсумнівно, можна собі уявити оптимістичніший варіант розвитку подій. Наприклад, немає підстав, щоб емуляції відчували біль від тілесних ушкоджень і хвороб. Фізична невразливість була б значним покращенням, порівнюючи з поточним станом. Ба більше, оскільки формувати віртуальну реальність може бути дуже легко, емуляції зможуть працювати у прекрасних місцях. У вишуканому палаці на вершині гори, на терасі посеред зелені весняного лісу, на сліпучому березі лазурної лагуни; з ідеальною освітленістю, температурою, пейзажем та інтер'єром; позбавлені

набридливих запахів, шумів, чернеток і дзижчання комах; у зручному одязі, за ясного та зосередженого розуму, позбавлені голоду та спраги. А головне, якщо оптимальний робочий стан для більшості завдань — це радісне завзяття (що дуже ймовірно), то ера економіки емуляцій може виявитися досить ідилічним періодом.

У будь-якому разі корисно було б, якби реальність почала дедалі більше скидатися на антиутопію, щоб все-таки залишилася можливість комусь або чомусь втрутитися і все виправити. Бажано також, щоб існував який-небудь стоп-кран чи запасний вихід, який би давав можливість свідомо закінчити існування й вийти в небуття, якщо якість життя опиниться за тією межею, де неіснування здається привабливішим.

Несвідомі аутсорсери?

Потім, коли ера емуляцій закінчиться і почнеться ера штучного інтелекту (або ера штучного інтелекту розпочнеться відразу, без проміжного етапу емуляцій), біль та насолода, як явища, можуть повністю щезнути, адже нагорода у вигляді задоволення може бути не настільки ефективною мотивацією для складного штучного агента (який, на відміну від людського розуму, не обтяжений вадами тваринної тілесності). Можливо, не функціональний аналог задоволення чи болю, а безпосередніше представлення корисності того чи того виду, буде значно кращою мотивацією.

Схожий, але дещо радикальніший сценарій багатополярного майбутнього може бути позбавлений будь-якої привабливості, якщо світовий пролетаріат не матиме свідомості як такої. Імовірність цього найбільша, якщо з'явиться штучний інтелект, за структурою відмінний від людського розуму. А втім, навіть якщо штучний інтелект спершу буде створений за допомогою емуляції мозку, конкуренція різних сил у постперехідній економіці може підштовхнути до створення все менш нейроморфних ШІ. З'явиться повністю штучний інтелект, створений з нуля, або емуляції через удосконалення та модифікації дедалі більше віддалятимуться від первісної архітектури.

Уявіть собі сценарій, коли, залучаючи емуляції — як моделей і як виконавців — до розгорнутих досліджень у галузі нейронауки й ком-п'ютерних технологій, стає можливо ізолювати окремі функціональні

модулі розуму емуляції та поєднувати їх із такими самими модулями інших. Можливо, таке поєднання потребуватиме періоду навчання та адаптації, натомість стандартизовані модулі зможуть швидше створювати взаємне з'єднання. Така стандартизація дасть змогу легше утворювати продуктивні структури, тому потреба в ній зростатиме.

У такий спосіб емуляції зможуть інтенсивніше залучати сторонні засоби для ефективнішого виконання завдань. Для чого вчити математику, якщо можна надіслати своє математичне завдання до ТОВ «Модулі Гаусса»? Для чого риторика, коли є ТОВ «Пишний Яр» для формулювання думок у слова? Для чого ухвалювати важкі рішення особистого характеру самотужки, якщо існують сертифіковані модулі, які просканують твої цілі, зважать твої ресурси і підкажуть краще рішення? Деякі емуляції продовжуватимуть усе робити самостійно, хоч ефективніше було б доручити ці завдання спеціалістам. Вони будуть ніби любителі всього саморобного — вирощених власноруч овочів, власноруч в'язаних светрів. Ефективність таких емуляцій-аматорів буде нижчою, якщо порівнювати з прихильниками аутсорсу. Тому згодом їх витіснять з ринку і вони зникнуть.

Тож окремі емуляції людських мізків припинять існування — розчиняться, як бульйонні кубики, в алгоритмічному супі.

Логічно, що для оптимальної ефективності окремі можливості мають бути об'єднані у структури, які приблизно повторюють архітектуру людського розуму. Наприклад, математичний модуль може бути поєднаний з модулем мовлення, а обидва разом — під'єднані до виконавчого модуля, щоб усі три могли працювати разом. У такому разі когнітивний аутсорсинг концептуально неможливий. Однак не існує жодної вагомої причини цьому, тому може бути, що архітектура людського мозку оптимальна лише для людської неврології (а можливо, не така вже й оптимальна). З появою можливості створювати архітектури, які неможливо реалізувати за допомогою біологічних нейронів, відкривається новий простір технологічних рішень. Критерії його оптимальності можуть бути незвичні для людського розуму. Тоді людиноподібні розумові утворення втратять свою нішу в конкурентному середовищі постперехідної економіки й екосистемі[399].

Можуть залишитися ніші для менш складних утворень (як-от окремі модулі), складніших (на кшталт кластерів модулів) або для утворень, подібних за складністю до людського розуму, але із зовсім іншою архітектурою. Чи матимуть такі комплекси яку-небудь цінність? Чи потрібний нам такий світ, у якому людей замінили собою такі несхожі утворення?

Відповідь на це питання залежить від природи цих утворень. Сучасний світ має багато рівнів організації. До складу деяких складних сутностей, як-от транснаціональні корпорації та країни, входять люди. Але ми схильні надавати таким утворенням лише інструментальну цінність. Корпорації та країни (як ми думаємо) не мають власної свідомості понад окремі свідомості людей, які входять до їхнього складу: самі собою вони не можуть відчувати біль чи насолоду і нездатні до почуттів. Ми оцінюємо їх за тим, наскільки вони забезпечують наші потреби і, щойно вони перестають бути корисними, ми «вбиваємо» їх без жодних докорів сумління. Також існують утворення нижчого рангу, яких ми теж не схильні наділяти моральним статусом. Ми не бачимо нічого поганого у видаленні програми з пам'яті смартфона, як і не засуджуємо дії нейрохірурга, коли він видаляє частину тканин мозку, які спричиняють епілепсію. Що ж стосується утворень, за складністю подібних до людського мозку, більшість із нас схилятиметься до визнання їхньої моральної значущості лише за умови, що вони здатні до усвідомленого досвіду дійсності[400].

Тому можемо уявити, як крайній випадок, високотехнологічне суспільство різноманітних складних сутностей, деякі з яких значно складніші та інтелектуальніші, ніж будь-яка із сутностей, що існують сьогодні на нашій планеті. Водночас жодна з них у теперішньому людському суспільстві не здавалась би морально значущою. У нашому розумінні це було б неживе суспільство. Суспільство, у якому існують економічні дива й технологічні принади, але немає тих, хто б ними користувався. Діснейленд без дітей.

Еволюція — це необов'язково добре

Слово «еволюція» зазвичай є синонімом «прогресу» через поширене уявлення, ніби еволюція є силою добра. Недоречна віра в безумовну

корисність еволюції може завадити об'єктивно оцінити бажаність багатополярного сценарію розвитку суперінтелектуальності, у якому майбутнє розумного життя на землі визначатиметься конкуренцією. Така думка повинна базуватися на бодай якійсь оцінці розподілу імовірних фенотипів, що будуть здатні конкурувати з інтелектуальними сутностями постперехідного цифрового супу. Важко знайти зрозумілу та достовірну відповідь в умовах невизначеності, якої неможливо уникнути в питаннях віддаленого майбутнього, особливо якщо дивитися на нього крізь рожеві окуляри оптимізму.

Причиною загальнопоширеної віри в благість еволюційного процесу може бути враження від результатів, які досі цей процес демонстрував у нашій історії. Від початку розмноження живих клітин еволюція створювала все «прогресивніші» організми, створіння, наділені мозком, свідомістю, мовою, розумом. Останнім часом завдяки культурним і технологічним процесам, які віддалено схожі на біологічну еволюцію, людство пришвидшило темпи свого розвитку. У масштабах геологічних процесів та й історичних теж ці темпи показують усеохопну тенденцію до концентрації складності, знань, свідомості та централізованої, цілеспрямованої організації. Її можна приблизно охарактеризувати як «прогрес»[401].

Такий сприятливий образ еволюції, як процесу, який завжди спрямований на благо, важко узгодити з неймовірною кількістю страждання, яке ми бачимо навколо себе як у світі людей, так і в дикій природі. Прихильники такого уявлення про еволюцію, очевидно, дивляться на процес з естетичного погляду, а не етичного. Але нас цікавить не те, про яке майбутнє нам було б цікавіше читати у фантастичних книжках чи переглянути документальний фільм, а, власне, зовсім інше: у якому майбутньому нам буде комфортніше жити? Погодьтеся: це значно важливіше питання.

Ба більше, немає підстав стверджувати, що весь наявний еволюційний прогрес був невідворотним. Багато з того, що відбулося, могло бути просто удачею. Такий висновок напрошується, якщо врахувати ефект упередження відбору, під впливом якого ми стверджуємо безпомилкову успішність еволюції[402]. Що, як на 99,9999 % планет, де коли-небудь еволюціонувало життя, воно зникло, не

досягши бодай якого-небудь розвитку? Що ми, як результат тої жалюгідної частки успішності еволюції, спостерігатимемо навколо? Можливо, те, що ми бачимо зараз. Гіпотеза про те, що ймовірність виникнення розумного життя на якійсь планеті дуже низька, не скасовує його виникнення на одній з величезної кількості планет, де існували примітивні форми життя. Тому тривала історія існування життя на Землі зовсім не свідчить про високу ймовірність — не говорячи вже про неминучість — появи на ній високорозвинених організмів[403].

Окрім того, навіть якби ми жили в ідилічних умовах, єдиною безумовною причиною яких був первинний еволюційний стан, усе одно немає гарантій, що такий меліористичний тренд ніколи не закінчиться. Навіть якщо ми знехтуємо ймовірністю раптового зникнення людства внаслідок глобального катаклізму, а особливо якщо еволюція продовжуватиме утворювати дедалі складніші системи.

Раніше ми припустили, що штучні робітники, виведені для максимальної продуктивності, надзвичайно важко працюватимуть, і невідомо, чи будуть вони тоді щасливі. Ба більше, існує можливість, що найбільш пристосовані форми електронного життя в конкурентних умовах цифрового супу майбутнього можуть навіть не мати свідомості. Окрім повної втрати здатності відчувати задоволення та усвідомлювати себе, для них також можуть виявитися непотрібними й інші якості, які ми зараз вважаємо невід'ємними атрибутами повноцінного життя. Люди кохають, цінують дружбу, спілкування, духовність і традиції, полюбляють пити і їсти, люблять музику, гумор, мистецтво, театр, танці, літературу, філософію, спорт, пригоди, природу і багато інших речей. Немає гарантій, що все це збереже цінність у майбутньому. Можливо, залишиться лише максимізація корисності — безперервне гарування, нудна щоденна праця, позбавлена мимовільного захвату натхнення, єдино присвячена меті додати чергову одиницю в десятому розряді після коми якого-небудь економічного показника ефективності. Згадані речі справді не будуть потрібні для успішного життя в такому світі. Залежно від індивідуальних уподобань такий світ може здатися відразливим, нікчемним чи, може, лиш дещо бляклим, але він точно не буде чарівною утопією, якої варто прагнути.

Може здатися неймовірним, що така безрадісна перспектива чекає людей, які насолоджуються музикою, гумором, коханням, мистецтвом тощо. Якщо ці заняття такі «нікчемні», то чому еволюція нашого виду їх досі не викорінила, а навіть заохочувала? Цей факт не можна пояснити еволюційним дисбалансом сучасності, адже навіть наші предки часів плейстоцену дозволяли собі витрачати енергію і час на подібні «дурниці». Та й не лише Homo sapiens помічений у згаданій діяльності. Виконавське мистецтво є частиною багатьох контекстів, від шлюбної поведінки тварин до престижних міжнаціональних змагань[404].

Не намагатимемося пояснити еволюційну необхідність кожного такого випадку, проте деякі з них точно не становлять функціональної потреби в контексті діяльності штучного інтелекту. Наприклад, гра трапляється лише в деяких видів і переважно на ранніх етапах розвитку — через гру молоді індивіди опановують навички, які знадобляться в дорослому житті. Якщо емуляції створюватимуться відразу зрілими з повним набором потрібних якостей або коли знання та вміння, набуті одним ШІ, можуть бути безпосередньо використаними іншим ШІ — тоді поширеність гри зменшуватиметься.

Багато інших прикладів людської поведінки могли з'явитися як засоби вираження якостей, що важко безпосередньо зауважити, як-от тілесна чи розумова витривалість, соціальний статус, наявність союзників, войовничість, володіння ресурсами. Класичний приклад — хвіст павича: лише достойні павичі можуть похвалитися по-справжньому елегантним хвостом, а пави під час еволюції, навчилися бачити в цьому привабливість. Про добру генетику чи інші соціально важливі характеристики можуть сигналізувати не лише морфологічні, а й поведінкові ознаки[405].

Раз показність настільки поширена серед вищих біологічних видів, то, може, їй знайдеться місце і у взаємовідносинах більш технологічних форм життя? Незважаючи на відсутність в екосистемі інтелектуальної обробки інформації інструментальної користі від артистизму, музичності, чи навіть свідомості, може ці риси все ж даватимуть якусь еволюційну перевагу тим, хто ними наділений, — самі собою чи як засіб сигналізації?

Можливо, наші цінності вдасться узгодити з цінностями цифрової екології майбутнього, проте існують причини для скепсису. По-перше, варто згадати, що демонстрація статусу в природі часто пов'язана зі шлюбною поведінкою[406]. Що стосується розмноження високотехнологічних форм життя, то воно може бути виключно асексуальним.

По-друге, високотехнологічні агенти, імовірно, матимуть досконаліші способи надійного передавання інформації про себе, ніж демонстрація статусу. Навіть зараз професійні позичальники схильні більше довіряти документальній інформації, свідоцтву власності чи банківській інформації, ніж зовнішнім ознакам статусу, як-от дорогий костюм чи годинник Rolex. У майбутньому можуть існувати професійні аудитори, які за допомогою ретельного дослідження інформації про діяльність, тестування в симуляції чи безпосередньої експертизи вихідного коду агента зможуть підтвердити чи спростувати наявність у нього певної характеристики, яку він собі приписує. Згода на аудит буде промовистішим свідченням на користь володіння такою характеристикою, ніж демонстрація ознак статусу. *Підробити* такий професійно засвідчений засіб сигналізування буде важко (достовірність — важливий аспект будь-якої сигналізації). Натомість, у випадку *правдивості* повторно використовувати такий засіб для підтвердження володіння згаданою характеристикою буде значно легше, ніж кожного разу вдаватися до демонстрації.

Крім того, не всяка демонстративність однаково цінна та соціально бажана. Часто причиною демонстративності є марнославство й марнотратство. Наприклад, церемонія потлачу в індіанців племені квакіутл, своєрідне змагання між вождями-суперниками, яка полягає в публічному знищенні величезної кількості цінностей[407]. Побудову хмарочосів рекордної висоти, купівлю мегаяхт і космічних ракет можна розглядати як сучасний аналог такій традиції. Музика чи гумор, натомість, покращують внутрішні якості життя людини, чого не можна сказати про дорогий одяг, модні аксесуари й інші атрибути зовнішності й символи споживацького статусу. Ба більше, зайва демонстративність може нашкодити, наприклад, показний мачизм часто призводить до войовничої бравади і групового насильства. Навіть якщо майбутні розумні форми життя використовуватимуть

демонстрацію статусу, невідомо, наскільки затратною буде її форма — буде вона складною та вигадливою, як пісня солов'я, чи короткою, як квак жаби (або схожою на безперервний гавкіт скаженого собаки).

Наступне утворення синґлтону

Навіть у разі одночасної появи у світі кількох конкурентних суперінтелектів існує ймовірність, що згодом такий стан усе одно трансформується в синґлтон. Це стане логічним завершенням давнього прагнення до політичної інтеграції[408]. Як це може відбутися?

Другий перехід

Світ штучних суперінтелектів, первісно багатополярний, може перетворитися на синґлтон за умови другого технологічного переходу, достатньо суттєвого та швидкого, щоб суперінтелект, який подолає його першим, отримав вирішальну стратегічну перевагу й можливість сформувати синґлтон. Таким гіпотетичним переходом може бути раптова можливість стрибка до вищого рівня суперінтелектуальності. Отож, якщо перша хвиля суперінтелектів буде ґрунтуватися на технології емуляцій, наступним стрибком може бути створення емуляціями ефективного ШІ зі здатністю до самонавчання[409]. (Також другий перехід може бути зумовлений проривом у сфері нанотехнологій чи будь-якій іншій технологічній сфері — військовій або загального призначення).

Після першого переходу швидкість розвитку науки буде дуже високою. Тому навіть незначний відрив між лідером технологічних перегонів та іншими учасниками потенційно може дати вирішальну стратегічну перевагу для наступного переходу. Наприклад, нехай два проекти входять у фазу переходу ШІ до суперінтелектуальності з відривом у кілька днів і темп зростання не настільки значний, щоб такий відрив давав комусь із них відчутну перевагу. Обидва проекти успішно створюють свої суперінтелекти, зберігаючи мінімальний розрив. Проте швидкість появи нових розробок тепер відчутно вища, адже в дослідженнях беруть участь суперінтелекти. Отже, стрімкість наукового прогресу може зрости в тисячі мільйонів разів. Створення технології, яка спричинить другий перехід, може зайняти кілька днів, чи годин, чи навіть хвилин. Тому достатньо навіть різниці в кілька

днів, щоб лідер технологічних перегонів отримав вирішальну стратегічну перевагу над суперниками. Проте варто зауважити, що такого відриву може виявитися недостатньо, якщо швидкість проникнення технологій (завдяки шпигунству чи іншими каналами) також зростатиме. Натомість важливу роль відіграватиме швидкість другого переходу, порівнюючи із загальною швидкістю розвитку подій після першого переходу. (Тому що швидше відбуватимуться події у світі після першого переходу, то складніше буде досягти стрімкості, потрібної для другого переходу).

Можна також припустити, що вирішальна стратегічна перевага, отримана внаслідок другого (чи будь-якого наступного) переходу з більшою ймовірністю приведе до синґлтону, ніж та, що з'явилася внаслідок першого переходу. Після першого переходу суб'єкти ухвалення рішень будуть суперінтелектами або матимуть змогу отримати пораду від суперінтелекту. Тому чіткіше бачитимуть можливі наслідки своїх рішень. Ба більше, можливо, що після першого переходу перехоплення ініціативи й дія на випередження свого суперника буде менш ризикованою для агресора. Якщо такий агресор буде цифровим розумом, він зможе створити резервну копію себе і в такий спосіб стане менш вразливим для контратаки. Навіть якби захиснику вдалося знищити дев'ять десятих популяції агресора, це не вплинуло б суттєво на ситуацію, якщо нападник зможе миттєво відновити ресурси з резервних копій. Подібно знищення інфраструктури може не мати значного впливу на цифровий штучний інтелект з необмеженою тривалістю життя та космічними планами максимізації ресурсів і впливу.

Суперорганізми й економіка масштабування

Розмір координованих структур, які створюють люди, як-от підприємства чи нації, зумовлюється багатьма факторами — технологічними, військовими, фінансовими, культурними, що змінюються з плином часу від однієї історичної епохи до іншої. Унаслідок революції штучного інтелекту також відбудуться значні зміни. Можливо, вони сприятимуть створенню синґлтону. Проте без детальних прогнозів не можна виключати і зворотний ефект — що зміни можуть сприяти фрагментації, а не уніфікації, — а значить в

умовах зростання невизначеності ймовірність утворення синґлтону може бути більшою, ніж нам здається. Революція ШІ може збурити ситуацію — перетасувати колоду так, що будь-які геополітичні комбінації, неймовірні до того, стануть можливими.

Детальне дослідження всіх факторів, які можуть вплинути на ступінь політичної інтеграції значно перевищить обсяг цієї книжки. Його заледве вистачить на поверховий огляд праць з політології та економіки із цієї теми. Тож обмежимося побіжним поглядом на деякі з факторів та аспектів діяльності цифрових агентів, які сприятимуть централізації управління.

На думку Карла Шульмана, селекція в популяції емуляцій приведе до появи «суперорганізмів», груп емуляцій, готових до самопожертви в інтересах свого клану[410]. Суперорганізми будуть позбавлені проблеми принципала-агента, характерної для людських організацій, у яких окремі індивіди мають власні інтереси. Емуляції співпрацюватимуть одна з одною навіть за відсутності будь-яких схем заохочення, як окремі клітини одного організму або еусоціальні комахи, повністю альтруїстично налаштовані щодо інших членів групи.

Такі суперорганізми будуть корисними, якщо видалення емуляцій (або зупинку на невизначений термін) без їхньої згоди буде заборонено. Роботодавці штучних працівників, які не дали дозволу на знищення, будуть змушені вічно утримувати їх, навіть якщо ті застаріли чи просто непотрібні. Натомість організації, у яких працівники самостійно видалятимуться за потреби, зможуть легше адаптуватися до змін попиту. Крім того, вони зможуть вільно експериментувати, створюючи працівників та залишаючи найбільш продуктивних.

Якщо видалення емуляцій буде дозволене, то перевага еусоціальності незначна, проте не зайва. У будь-якому разі працедавцям таких емуляцій не доведеться боротися з проблемами принципала-агента та долати можливий опір під час видалення. Узагалі продуктивність, альтруїзм і самопожертва заради роботи — лише окремі переваги емуляції, яка фанатично віддана організації, у якій працює. Такі працівники не тільки в буквальному сенсі стрибнуть у вогонь чи воду заради організації або працюватимуть за малу платню: вони не плестимуть офісних інтриг та намагатимуться завжди діяти в

інтересах компанії тою мірою, якою їх розумітимуть, позбавляючи керівництво потреби у нагляді та бюрократичних пересторогах.

Однак існує небезпека негативного ефекту: працівники, створені зі спільного образу для досягнення подібної відданості інтересам компанії, можуть наслідувати обмеження й вади цього образу, на відміну від працівників інших компаній[411]. Цей ефект можна суттєво знизити, якщо суперорганізм міститиме неоднаково треновані емуляції. Тобто навіть якщо емуляції мусять походити з одного праобразу, то принаймні їх можна по-різному тренувати. Так із поліматематично обдарованого праобразу можна вивчити різні спеціалізації емуляцій: бухгалтера, інженера електроніки тощо. Завдяки цьому працівники зможуть мати різні уміння, проте не різні таланти. (Для більшої обдарованості треба буде використовувати кілька праобразів).

Найважливішою властивістю суперорганізму є не те, що він складається з копій однакового походження, а те, що всі його елементи повністю віддані спільній меті. Тому для створення такого суперорганізму потрібно вирішити проблему контролю для конкретної реалізації агентів. Загальне розв'язання проблеми контролю дасть змогу створити агента з будь-якою кінцевою метою, тоді як для створення суперорганізму необхідне часткове вирішення — здатність успішно створити агента з конкретною (хоч, можливо, нетривіальною) метою[412].

Але я не хотів би робити акцент на моноклональному емулюванні. Основна думка стосується широкого спектра багатополярних сценаріїв розвитку штучного інтелекту. Досягнення в методах відбору мотивації, зумовлені цифровою природою суперінтелектуальних агентів, можуть дати змогу подолати багаторічні обмеження, які знесилюють великі людські організації і стримують економіку. Без них масштаби організаційних утворень — чи то підприємства, нації, чи інші політичні й економічні сутності — могли б бути значно більшими. Це один із факторів, який би сприяв утворенню синґлтону.

Ще одна сфера, у якій суперорганізми (чи інші вмотивовані цифрові агенти) можуть досягнути значних результатів, — це примус. Держава може застосовувати технології відбору мотивації, щоб забезпечити

безоглядну лояльність поліції, збройних сил, розвідки, державних службовців. На думку Шульмана:

> *Збережені о́брази внутрішнього стану [ретельно сформованої та перевіреної лояльної емуляції] можна розмножити мільярд разів, щоб утворити ідеологічно однорідні армію, силові структури та бюрократичний апарат. Щоб запобігти ідеологічним змінам, через короткі проміжки часу таких працівників замінюватимуть на нові екземпляри, утворені з того ж о́бразу. Це дасть змогу дуже точно оцінювати і впливати на функціонування кожного резидента в межах потрібного контексту: для кожного може існувати окремий о́браз. Так можна буде запобігти створенню зброї масового ураження, проведення експериментів над емуляціями, обмежити розмноження емуляцій, забезпечити виконання ліберальної демократичної конституції чи встановити жахливий і перманентний тоталітаризм*[413].

Основним результатом такого відбору мотивації буде консолідація влади і концентрація її в меншої кількості суб'єктів.

Уніфікація за домовленістю

У постперехідному багатополярному світі міжнародна співпраця може бути дуже корисною. Вона може допомогти запобігти війнам та нарощуванню озброєнь. Спільно колонізувати й використовувати астрофізичні ресурси оптимальними темпами. Спільна розробка досконаліших форм штучного інтелекту дасть змогу уникнути поспіху і ретельніше тестувати нові реалізації. Це дасть можливість призупинити розробки, які вестимуть до потенційних екзистенційних ризиків. Можна буде узгодити законодавство, запровадити безумовне мінімальне забезпечення громадян (яке потребуватиме гнучкого демографічного контролю), заходи запобігання експлуатації та жорстокого поводження з емуляціями та іншими цифровими й біологічними свідомостями. Ба більше, суперінтелектуальні агенти з обмеженими потребами в ресурсах (детальніше — у розділі 13) радше віддадуть перевагу укладенню угоди про спільне використання, що гарантуватиме їм частку ресурсів у майбутньому, ніж вільній боротьбі за ресурси, у якій вони можуть не отримати нічого.

Проте усвідомлення вигідності співпраці не означає, що така вона відбудеться. Сучасний світ теж виграв би від глобальної співпраці — скорочення військових витрат, зменшення кількості війн, зупинення надлишкового вилову риби, зняття торговельних обмежень, боротьба із забрудненням атмосфери — це лише деякі з вигод, які могло б отримати людство. Однак ці плоди здорового глузду досі гниють на гілках. Чому? Що стримує нас від співпраці, яка принесла б усім лише добро?

Перешкодою є довіра: складно забезпечити глобальну відповідність умовам угоди — зокрема, моніторинг виконання та санкціонування порушень. Двом ядерним державам, що протистоять одна одній, було б вигідно позбавитися від ядерної зброї. Однак навіть якби вони дійшли згоди — на папері, — на практиці процедура роззброєння могла б ніколи повністю не завершитися через взаємні побоювання, що інша сторона насправді не виконує погоджених заходів. Захистити від цього може механізм верифікації. Інспектори мають наглядати за знищенням запасів, перевіряти ядерні реактори й інше обладнання. Потрібно проводити зустрічі з представниками технічної та гуманітарної інтелігенції, щоб упевнитися, що військові програми не відновлено. З одного боку, такі перевірки треба фінансувати. З іншого, існує ризик, що інспектори можуть шпигувати й передавати військові та промислові секрети іншій стороні. А головне, кожна зі сторін побоюватиметься, що інша може приховати ядерну зброю і не знищити її. Багато потенційних угод ніколи не стають реальністю, бо верифікація їх виконання виявляється занадто складним завданням.

Якби з'явилися нові дешевші технології верифікації, то можна було б сподіватися на розширення співпраці. Проте не зрозуміло, чи варто очікувати появу таких технологій у постперехідній ері. Безумовно, завдяки ШІ з'явиться багато нових технологій, зокрема й у сфері верифікації, але те саме можна сказати і про технології приховування. Дедалі більше процесів переноситимуть у кіберпростір, за межі досяжності засобів фізичного нагляду. Наприклад, цифровий розум, який розробляє нову нанотехнологічну зброю або нове покоління штучного інтелекту, може майже ніяк фізично не виявляти своєї діяльності. Цифровій судовій експертизі може бути не під силу

проникнути крізь численні шари маскування та шифрування, за якими порушник ховатиме свою шкідливу діяльність.

Верифікації відповідності допоміг би надійний детектор брехні[414]. Тоді протокол інспекції містив би співбесіди з ключовими керівниками, щоб упевнитися у щирості їхніх намірів щодо імплементації положень угоди і, що їм не відомі жодні випадки порушень, хоч вони активно їх шукають.

Недоброчесний керівник може обдурити детектор брехні, давши підлеглим злочинний наказ, а також звелівши приховати факт його виконання навіть від нього самого, після чого видалити спогади про наказ і сам намір злочину. Завдяки новим нейротехнологіям більш-менш точне редагування спогадів біологічного мозку може стати реальністю. Для штучного інтелекту здійснити таку операцію може бути ще простіше (проте залежно від архітектури ШІ).

У державних установах для подолання проблеми можуть запровадити обов'язкове періодичне тестування ключових функціонерів на поліграфі, щоб виявити, чи не приховує хто-небудь з них намірів скасування чи саботажу домовленостей, яких держава досягла чи планує досягти в майбутньому. Глобальна практика такого моніторингу може бути визначена окремою метаугодою, покликаною забезпечувати умови для верифікації будь-яких інших угод, минулих чи майбутніх. Держави, які бажатимуть здобути визнання себе на міжнародному рівні надійним партнером, приєднуватимуться до цієї метаугоди індивідуальним порядком. Проте її ефективність може страждати від класичного принципу «делегуй і забудь». В ідеалі її імплементація повинна відбутися, *перш* ніж будь-яка зі сторін чи зацікавлених сил зможе підготувати ґрунт для її зриву. Якщо злі сили скористаються моментом і закладуть зерна брехні, довіри поміж сторонами вже ніколи не буде.

Іноді, щоб досягти домовленості, достатньо мати можливість *встановити* факт порушення її умов. Проте в деяких випадках потрібний механізм *забезпечення* відповідності та призначення покарання в разі виявлення порушень умов угоди. Такі заходи необхідні, якщо загрози виходу з угоди інших сторін недостатньо, щоб утримати гіпотетичного порушника від зловмисних дій. Наприклад,

якщо порушник матиме настільки значну перевагу над іншими сторонами, що будь-які їхні дії йому не загрожуватимуть.

Якщо існуватимуть високоефективні методи відбору мотивації, можна буде вирішити цю проблему, делегувавши функції аудиту й арбітражу незалежному суб'єкту з достатніми для успішної діяльності силовими ресурсами. Цей варіант передбачає бездоганну репутацію такого суб'єкта поміж сторонами, що домовляються. Проте саме через високий ступінь контролю мотивації для повної впевненості сторін у незалежності суб'єкта, його створення повинне відбуватися під спільним наглядом усіх учасників.

Створення такого зовнішнього суб'єкта та наділення його владою забезпечувати міжнародні домовленості знову порушує питання, яких ми вже торкалися раніше, коли розглядали однополярний сценарій (де синґлтон утворюється відразу після революції штучного інтелекту). Адже для забезпечення виконання угод, які стосуватимуться надважливих аспектів безпеки, між країнами-суперниками такий суб'єкт повинен, по суті, утворити синґлтон. Тобто стати глобальним суперінтелектуальним Левіафаном. Єдиною відмінністю є те, що тут ми говоримо про постперехідне суспільство, у якому створюватимуть такого Левіафана штучні агенти, значно компетентніші за нас — імовірно, теж суперінтелектуальні. Тому шанси, що вони зможуть розв'язати проблему контролю і створити інтелектуальну систему, яка враховуватиме інтереси всіх зацікавлених сторін, значно вищі.

Що ще заважає міжнародній співпраці, крім необхідності ефективного моніторингу та забезпечення відповідності вимогам? Мабуть, найбільшою з перешкод, які залишилися, є те, що можна назвати *вартістю згоди*[415]. Часто, коли є порозуміння щодо взаємовигідного варіанта розвитку подій, сторони так і не досягають остаточної домовленості через незгоду з тим, як розподіляються витрати. Наприклад, коли двоє людей можуть домовитися про угоду, яка принесе долар доходу, але кожен вважає, що заслуговує не менш як шістдесят центів, угода, яка може обом принести прибуток, зривається й обоє залишаються з нулем. Узагалі перемовини часто бувають складними, тривалими та іноді навіть безрезультатними через необхідність для сторін робити стратегічний вибір між здобутками і втратами.

У реальності люди часто погоджуються на умови, уникаючи стратегічних перемовин (які потребують часу й терпіння). Зрозуміло, що динаміка цих перемовин у постперехідну еру може бути іншою. Штучний переговорник, який провадить перемовини з іншим ШІ, може послідовно дотримуватися певного формального розуміння раціональності, яке матиме нові та неочікувані наслідки для сфери перемовин. Крім того, ШІ може мати у своєму арсеналі засоби ведення перемовин, незрозумілі або нездійсненні для людини. Наприклад, можливість застосування «передзобов'язання»: попереднє обмеження своїх дій умовою дотримуватися певної вимоги або стратегії. Люди та їхні організації іноді також можуть застосовувати передзобов'язання — часто недостатньо чесні та точні, — проте ШІ зможе застосовувати передзобов'язання будь-якої складності й матиме змогу формально довести партнерам, що його поле дій обмежене цими умовами[416].

Розвиток концепції передзобов'язань може докорінно змінити сучасну природу переговорного процесу, і перший агент, здатний застосувати її, матиме значну перевагу. Нехай для успішного досягнення якогось важливого результату потрібна співпраця з одним конкретним агентом. Якщо цей агент публічно підтвердить існування передзобов'язання не погоджуватися на співпрацю менше ніж за 99 відсотків від результату, він зможе диктувати умови розподілу збитків. Оскільки обмеження дій потрібного колеги передзобов'язанням буде для решти партнерів абсолютним фактом, у них залишиться лише два варіанти вирішення ситуації: або вони не отримують нічого (відмовляючись від безумовно нечесної пропозиції), або отримують один відсоток результату (погоджуючись).

Щоб уникнути таких невигідних пропозицій, агенти можуть застосувати щодо себе передзобов'язання відкидати спроби шантажу і пропозиції несправедливого розподілу результатів діяльності. Якщо таке передзобов'язання буде публічно підтверджене, інші агенти не будуть зацікавлені в маніпулюванні чи висуненні несправедливих пропозицій, оскільки це буде наперед програшним варіантом. Це зайвий раз демонструє перевагу того, хто зробить перший крок. Агент, який першим застосує цю техніку, може вибирати, чи убезпечити себе від нечесних пропозицій, чи самому відхопити найбільший шмат торта.

Найвиправданішою може здатися позиція агента, невразливого до шантажу, який схильний вимагати максимум віддачі за роботу, яка без його участі неможлива. Деякі люди вже зараз володіють елементами цієї безкомпромісності[417]. Однак така завзятість може виявитися вадою, якщо наштовхнеться на більшу завзятість такого самого прихильника принципу «все або нічого». Тоді нестримна коса наштовхнеться на непідйомний камінь і результатом стане відсутність згоди (чи гірше: тотальна війна). Лагідні та акратичні отримають принаймні щось, хоч, звісно, менше, ніж вони заслуговують.

То який же результат цієї гри в переговори віщує нам теорія ігор — наразі невідомо. Стратегії дій, які виберуть агенти, можуть бути значно складніші за описані тут. Варто *сподіватися,* що інтереси зустрінуться посередині, у певному місці, яке слугуватиме точкою Шеллінга — місцем сили в широкому просторі можливих станів майбутнього, у якому перетнуться очікування багатьох учасників процесу — найімовірнішими координатами спільної цілі в цій грі на координацію. На цю рівновагу, можливо, вплинуть деякі з продуктів еволюції наших намірів та культурних програм. Наприклад, прагнення до справедливості могло б схилити баланс загальних очікувань і стратегій до сприятливішого для нас майбутнього — за умови, що ми зможемо пронести наші цінності в постперехідну еру крізь буремні роки конкуренції з ШІ[418].

У будь-якому разі використання техніки передзобов'язань — у жорсткій чи гнучкій формі — може стати причиною того, що процес перемовин докорінно зміниться. Навіть якщо початок постперехідної ери буде багатополярним, синґлтон може постати майже відразу як результат узгодження домовленостей та найкраще вирішення проблеми глобальної координації. Підготувати та імплементувати необхідні заходи моніторингу й забезпечення дотримання умов домовленостей допоможуть нові технології, створені ШІ. Інші аспекти, наприклад ведення стратегічних перемовин, можуть вплинути на зміст умов, проте немає підстав вважати, що такі перемовини суттєво загальмують досягнення домовленостей. Провал переговорів означатиме загострення конфронтації — у тій чи тій формі. Тоді або переможець сформує синґлтон, або конфлікт не вдасться розв'язати, і

настане безславний кінець тому, що починалося як спільний проект людства і його нащадків.

* * *

Отже, багатополярність, навіть стабільна, не гарантує нам сприятливого майбутнього. Невирішена проблема принципала-агента, ускладнена проблемами координації сил у багатополярному світі постперехідної ери, тільки погіршує стан справ. Тому повернімося до питання, як можна стримати суперінтелектуальний ШІ.

12. ФОРМУВАННЯ ЦІННОСТЕЙ

Контроль здібностей — допоміжний і тимчасовий засіб. Не вічно ж суперінтелекту бути ув'язненим, тому потрібні дієві засоби формування в нього правильної мотивації. Але як можна дати штучному агенту цінність — духовну, моральну чи будь-яку іншу нематеріальну цінність — і змусити його послуговуватися нею у своїй діяльності, поширювати й захищати її прояви? Недостатньо інтелектуальний агент може бути нездатним розуміти чи навіть механічно втілювати жодні важливі для людей цінності. А суперінтелектуальний агент може опиратися будь-яким змінам у своїй системі мотивації з конвергентних інструментальних причин (які ми описували в розділі 7). Ця проблема — прищеплення людських цінностей ШІ — складна, але неуникна.

Проблема прищеплення цінностей

Неможливо передбачити всі ситуації, у яких може опинитися суперінтелект, і вказати, як він повинен чинити в кожній з них. Так само — нереально перелічити всі можливі світи й оцінити ймовірність кожного з них. У світі, правила якого складніші від гри в хрестики-нулики, кількість можливих станів (та історій — послідовностей станів) занадто велика, щоб їх можна було повністю визначити. Тому мотиваційна система інтелекту не може бути просто таблицею, у якій містяться відповіді на всі можливі питання. Це радше має бути абстрактна формула чи правило, застосовуючи які до конкретного стану, можна отримати відповідь — як діяти.

Один зі способів формалізації такого правила — функція корисності. Ця функція (яку ми вже розглядали в розділі 1) присвоює кожному результату ймовірної дії, або, інакше кажучи, «ймовірному світу», певне значення (оцінку). З нею можна створити агента, який максимізуватиме очікувану корисність. Оцінюючи результат кожної

потенційної дії за допомогою функції корисності, агент вибиратиме ту, яка матиме максимальну очікувану корисність. (Очікувана корисність дії — це середня корисність усіх можливих світів, зважена на суб'єктивну ймовірність відбутися за умови, що ця дія була здійснена). Насправді значна кількість можливих світів надто ускладнює точний розрахунок очікуваної корисності дії. Так чи інакше, правило ухвалення рішення і функція корисності разом формують нормативний ідеал — поняття оптимальності. Агент повинен бути здатний його апроксимувати для реальних ситуацій. Що інтелектуальнішим буде агент, то точнішою має бути апроксимація[419]. Створити машину, яка може достатньо точно обчислити очікувану корисність доступних їй варіантів дій, так само складно, як і створити повноцінний ШІ[420]. Але в цьому розділі ми розглянемо інше завдання: завдання, яке залишиться навіть після створення розумних машин.

Спробуємо на прикладі максимізаційної моделі агента спрогнозувати завдання, які постануть перед програмістом зерна ШІ, коли в процесі встановлення якої-небудь корисної для людини кінцевої мети він повинен буде вирішити проблему контролю. Програміст намагається прищепити агенту цілком конкретну і зрозумілу для людини цінність, яку він (агент) має створити. Нехай це буде щастя. (Зі справедливістю, свободою, славою, правами людини, демократією, екологічною рівновагою, саморозвитком ситуація буде ідентичною). Для визначення очікуваної корисності програмісту потрібна функція, яка ставить у відповідність кожному ймовірному світу певне значення, пропорційне до кількості щастя, яке є в ньому. Проте як виразити таку функцію у комп'ютерному коді? Жодна мова програмування не має такого поняття як «щастя». Для того щоб його використовувати, його потрібно спершу визначити. Недостатньо дати визначення поняттю через інші людські поняття: «щастя — це властиве людській природі задоволення від можливостей» чи іншу схожу філософську фразу. Визначення має спиратися на поняття, які існують у мові програмування ШІ, на примітиви, як-от математичні оператори, й адреси, які вказують на регістри пам'яті з даними. Якщо поглянути на проблему з такого боку, стає зрозуміло, наскільки складне завдання програміста.

Ідентифікувати та кодифікувати людські цінності важко, бо вони складні. Їхня складність для нас природна, тому ми часто її не помічаємо. Можна порівняти це зі зором. Нам бачити не складно, бо ми це робимо без зусиль[421]. Потрібно лише розплющити очі — і багатий, інформативний потік осяжних, тривимірних образів всього, що нас оточує, заповнить наш розум. Ми хапаємо візуальні образи інтуїтивно, як підстаркуватий герцог, не дивлячись, бере зі столу чашку чаю, яка роками, ніби сама собою, з'являється на цьому місці щоранку. Але виконання навіть найпростішого візуального завдання — знайти сільничку на кухні — потребує величезної обчислювальної роботи. З послідовностей двовимірних образів, що зароджуються в нервових закінченнях сітківки й одночасно поступають зоровим нервом у мозок, зорові центри відтворюють тривимірне відображення простору, що нас оточує. Значна частина дорогоцінної площі нашої сірої речовини відведена для оброблення візуальної інформації і, поки ви читаєте ці рядки, мільярди нейронів у вашому мозку постійно працюють над зображенням (як невтомні швачки, які, схилившись над швейними машинками, щомиті зшивають для вас із клаптиків велетенське кольорове панно). Так само й наші цінності та бажання, які здаються для нас простими і зрозумілими, насправді є складними та багатошаровими поняттями[422]. Як програмісту вмістити цю складність у функцію корисності?

Можна було б спробувати безпосередньо та повністю відтворити в програмі ціль, яку повинний мати ШІ, повністю закодувавши її у функцію корисності. Такий підхід міг би спрацювати, якби йшлося лише про дуже просту кінцеву мету — наприклад, обчислення десяткових розрядів числа пі. Тобто якби все, що нам було б потрібно від ШІ, це таке обчислення, і нам було б байдуже до інших наслідків такої діяльності — згадайте, як ми раніше обговорювали невдалі реалізації та інфраструктурне пригнічення. Для цілей одомашнення такий спосіб створення функції корисності також може бути виправданий. Проте в разі прищеплення більш *людських* цінностей та ще й системі, яка має стати суперінтелектуальним сувереном, прямо запрограмувати всі нюанси поведінки неможливо[423].

Отже, що нам залишається, якщо перенести складні людські поняття та цінності в комп'ютерний код ми не можемо? У цьому розділі ми

обговоримо кілька альтернативних способів. Деякі з них спершу здаються цілком здійсненними, проте після уважнішого розгляду їхня привабливість стане не такою вже й очевидною. У майбутньому варто приділити більше уваги тим, які будуть перспективнішими.

Вирішення проблеми прищеплення цінностей ШІ — складне завдання, достойне праці найкращих математиків наступних поколінь. Його не можна відкладати на потім: щойно з'явиться достатньо розумний ШІ, він зможе розгадати наші наміри і, можливо, забажає зашкодити їм. У розділі, присвяченому конвергентним інструментальним причинам, ми демонстрували, що інтелектуальна система може опиратися зміні своїх кінцевих цілей. Якщо до моменту отримання здатності до саморефлексії агент не буде принципово лояльним до нас, він навряд чи сприйме з прихильністю таке промивання мізків чи тотальний перезапис його сутності більш філантропічною версією себе.

Еволюційна селекція

Очевидно, що щонайменше один раз еволюції вдалося створити організми з людськими цінностями. Цей факт живить переконання, що за допомогою еволюційних методів можна розвинути людські цінності і в машині. Однак є серйозні сумніви щодо безпечності такого шляху. У кінці розділу 10, коли описували небезпеки використання потужних пошукових алгоритмів, ми вже наводили підстави для таких сумнівів.

Еволюцію можна розглядати як окремий клас пошукових алгоритмів, який полягає у попеременному повторенні двох кроків: збільшення популяції потенційних кандидатів за деяким порівняно простим стохастичним законом (як-от випадкова мутація чи статева рекомбінація) і скорочення популяції шляхом знищення кандидатів, які не відповідають вимогам певного алгоритму оцінки. Небезпека такого типу пошуку в тому, що його результат може формально задовольняти критерій пошуку, проте не відповідатиме нашим сподіванням. (Це однаковою мірою стосується як процесу еволюції цифрового розуму із цілями, подібними до цілей типової людини, так і еволюції бездоганно морального чи бездоганно слухняного цифрового

розуму). Ризику можна уникнути, визначивши для пошуку критерій, який би ідеально відображав усі виміри потрібної мети, а не лише один аспект нашого уявлення про неї. Однак саме в цьому і полягає проблема прищеплення цінностей, і що, як вирішення буде знайдено?

Тоді постає така проблема:

> *Загальну кількість страждання у світі, що відбувається протягом року, важко уявити. Упродовж хвилини, поки я писав це речення, тисячох тварин з'їли живцем, інші змушені тікати в нестямі, щоб урятуватися, ще інших — повільно їдять ізсередини паразити. Тисячі помирають від голоду, спраги та хвороб*[424].

Навіть наш вид щодня втрачає сто п'ятдесят тисяч осіб і ще більше людей терпить жахливі страждання та втрати[425]. Природа, можливо, чудовий експериментатор, проте ніколи не завдавала собі клопоту з етикою — усупереч Гельсінській декларації та усім можливим нормам пристойності й моралі. Важливо уникнути повторення цих жахіть *in silico*. Щоправда, у створенні штучного інтелекту рівня людини еволюційним методом, мабуть, найважче буде уникнути саме думкозлочину — принаймні якщо цей процес буде схожий на біологічну еволюцію[426].

Навчання з підкріпленням

Навчання з підкріпленням — це вид машинного навчання, у межах якого досліджують можливості агентів, що вчаться максимізувати певну визначену кумулятивну винагороду. Створивши середовище, у якому агент здобуває нагороду за бажані дії, можна навчити його виконувати широке коло завдань (навіть без деталізації вимог, лише за допомогою сигналу нагороди). Такий агент послідовно вибудовує певну функцію оцінки своєї поведінки, яка присвоює значення станам, парам «стан — дія» та стратегіям. (Наприклад, за допомогою навчання з підкріпленням програма може навчитися грати в нарди, покращуючи свою здатність вираховувати положення фішок на дошці). Функцію оцінки поведінки, яка вдосконалюється з кожним новим досвідом, можна вважати формою накопичення знань про поняття, яке вивчається. Проте результатом навчання тут є не *саме* поняття, а щоразу точніше *відображення проміжних значень* відповідних

внутрішніх станів агента (або пар «стан — дія», або стратегій). Кінцева мета такого агента завжди незмінна: максимізація очікуваної нагороди. Нагорода виражається певними сигналами, що надходять від середовища. Тому для інтелектуального агента, що має достатньо складне уявлення про світ, щоб передбачити таку безпосередню можливість максимізації нагороди, цілком імовірною стратегією максимізації є вайргединг[427].

Ці перестороги не означають, що навчання з підкріпленням не можна застосувати в зерні ШІ. Варто лише розуміти, що не треба дозволяти такій системі керуватися виключно принципом максимізації винагороди. Тому проблему прищеплення цінностей, очевидно, доведеться розв'язувати іншим способом.

Асоціативне накопичення цінностей

Виникає запитання: якщо проблема прищеплення цінностей така складна, як отримуємо свої цінності ми?

Імовірна (спрощена) модель цього процесу може мати такий вигляд. Ми з'являємося на світ із порівняно простими прагненнями (наприклад, уникнення неприємного сенсорного досвіду) і набором схильностей, які сприяють формуванню додаткових прагнень (наприклад, ми можемо прагнути об'єктів і дій, які цінуються та заохочуються в нашій культурі). Такі первинні прагнення і схильності є для нас вродженими та сформовані природним і статевим відбором унаслідок еволюції нашого виду. Проте коли ми дорослішаємо, наші прагнення стають дедалі більше зумовленими попередніми подіями нашого життя. Отже, наші цінності не записані в нас у генах, а більшою мірою формуються нашим досвідом.

Наприклад, коли ми закохуємося в іншу людину, ми стаємо небайдужими до її чи його добробуту. Що містить у собі така цінність? Мабуть, багато чого, але розглянемо лише компоненти «кохана людина» і «добробут». Ці поняття не закодовані безпосередньо в нашій ДНК. ДНК лише містить інформацію про структуру мозку, який після років життя серед інших людей формує розуміння понять «людина» і «добробут». Розуміючи ці речі, мозок може формувати прагнення до них. Однак щоб цінності формувалися саме навколо *цих*, а не будь-

яких інших набутих понять (як-от горщик для квітів чи штопор), потрібні деякі вроджені механізми.

Невідомо, що це за механізми та як вони працюють. У людей вони, вочевидь, складні й багатогранні. Простіше зрозуміти явище, розглянувши його в більш рудиментарній формі, як-от імпринтинг у птахів, коли пташеня в перший день після вилуплення прагне перебування поблизу об'єкта, який стимулює його рухову активність. Сам об'єкт може бути будь-яким — це залежить від попереднього досвіду пташеняти. Генетично зумовленим є саме прагнення. Так само Гаррі може прагнути добробуту Саллі. Проте, якби вони не зустрілися, Гаррі покохав би когось іншого — тоді, можливо, його мета теж була б іншою. У наших генах закодований лише механізм формування цілей. Саме завдяки йому ми можемо мати цілі, інформаційна місткість яких значно перевищує можливості нашого генома.

Отже, можемо припустити, що за тим самим принципом можна створити мотиваційну систему ШІ. Тобто замість безпосереднього визначення цілей можна описати певний механізм, завдяки якому ШІ в процесі взаємодії з певним середовищем сформує потрібні нам цінності.

Здається, відтворити людський процес набуття цінностей складно. Відповідні генетичні механізми формувалися в людей протягом еонів років еволюції. Це значний обсяг роботи, яку буде складно повторити. Крім того, цей механізм, імовірно, оптимізований під нейрокогнітивні особливості людського мозку, тому, якщо це не нейроморфна емуляція мозку, застосувати його в ШІ буде неможливо. А маючи технологічну можливість емуляції, логічно створювати емуляцію мозку вже дорослої людини зі сформованими людськими цінностями[428].

Тому використання аналога людського біологічного механізму для прищеплення штучному інтелекту цінностей має загалом примарні перспективи. Може, варто спробувати створити який-небудь більш штучний механізм, що дасть змогу імпортувати детальні представлення потрібних нам цінностей у мотиваційну систему ШІ? Тоді не буде потреби точно відтворювати всі людські наміри у ШІ. Це може бути навіть небажано — зрештою, людська природа надто зіпсована й часто схильна до лихих вчинків, а це неприпустимо для системи, яка може отримати вирішальну стратегічну перевагу.

Натомість краще намагатися створити мотиваційну систему, яка б, базуючись на людських цінностях, системно та контрольовано вибирала напрям розвитку мотивації у бік альтруїзму, співчуття та благородства, утілюючи в ШІ наше уявлення про справжню людяність. Заразом, щоб такі вдосконалення справді були на краще, вони повинні бути не випадковими, а переважно залишатися в межах антропоцентричної системи оцінок (щоб уникнути хибних реалізацій цілком притомних цілей, як було показано в розділі 8). Наразі невідомо, чи це можливо.

Крім того, ШІ може захотіти вимкнути механізм асоціативного набуття цінностей. Як ми вже показували в розділі 7, цілісність та незмінність системи цінностей є для ШІ конвергентною інструментальною ціллю. Досягнувши певної стадії розумового розвитку, він може почати вважати шкідливими наслідки функціонування механізму набуття цінностей[429]. Це не завжди погано, але варто вжити заходів, щоб консервація системи цінностей відбулася в потрібний момент. *Після* того як ШІ набув корисних цінностей і *перед* тим, як корисні цінності почали перезаписуватися новими — шкідливими.

Шаблон мотивації

Ще одним підходом до розв'язання проблеми прищеплення цінностей ШІ може бути те, що ми назвемо шаблоном мотивації. Під час створення зерна ШІ йому надають проміжну мотиваційну систему, яка містить максимально прості цілі і створена за допомогою безпосереднього програмування цілей або будь-яким іншим доступним методом. Коли зерно ШІ розвивається й набуває можливості оперувати складнішими представленнями понять, проміжний шаблон мотиваційної системи замінюють на складніший та вдосконалений із іншими цілями. Ця нова система керує діяльністю ШІ протягом зростання до справжньої суперінтелектуальності.

Оскільки цілі, які містяться в такому шаблоні, не інструментальні, а *кінцеві*, ШІ може опиратися їхній зміні (з інструментальних причин). Це створює загрозу. Якщо ШІ не дасть змінити шаблон мотивації на нові цілі, метод не спрацює.

Варто вжити заходів, щоб уникнути цього. Для цього, наприклад, можна застосувати методи контролю здібностей, що обмежують потужність ШІ, поки проміжну мотиваційну систему не замінять на зрілу. Зокрема, можна обмежити рівень розвитку ШІ до безпечного рівня, водночас достатнього для переходу на нову мотиваційну систему. Із цією метою, можливо, доведеться гальмувати окремі здібності ШІ, зокрема, обмежуючи здібності до стратегічного планування, але дозволяючи розвиток інших (нібито) нешкідливих здібностей.

Також, щоб налаштувати зерно ШІ на конструктивніші відносини з програмістами, можна застосувати відбір мотивації. Так можна записати в шаблонній системі мотивації прихильність до зовнішніх команд розробників та прийняття цілей, які вони диктують[430]. Крім того, шаблонна мотиваційна система може дозволяти безперешкодний перегляд програмістами цілей та стратегій ШІ, передбачати властивості майбутньої архітектури, важливі для контролю й запровадження потрібного розробникам набору кінцевих цілей разом із одомашненням мотивації (обмеження використання обчислювальних ресурсів).

Можна навіть уявити, що зерно ШІ матиме єдину мету: зміна мети на нову, визначену непрямим способом, або, можливо, таку, яку програмісти мали на увазі. У такому механізмі «самозаміни» мети прихована проблема, ідентична до проблеми вивчення цінностей, яке ми розглянемо в наступному підрозділі. Детальніше обговорення буде в розділі 13.

Підхід використання шаблону мотивації не позбавлений недоліків. По-перше, ШІ із шаблонною системою мотивації може стати надто потужним. Він (за допомогою прямого опору чи саботажу) не дозволить встановити собі кінцеві цілі. Тоді ШІ стане суперінтелектом зі старою системою цілей. По-друге, визначити кінцеві цілі ШІ рівня людини необов'язково простіше, ніж визначити цілі примітивнішому ШІ. ШІ рівня людини значно складніший і може мати досить заплутану архітектуру, у якій складно розібратися і яку складно змінити. Натомість зерно ШІ подібне до *tabula rasa,* на якій програмісти можуть написати все, що вважатимуть за потрібне. Цей недолік може стати перевагою, якщо шаблон міститиме прагнення

розвинути таку архітектуру, яка була б сприйнятливою та прихильною до встановлення нових цілей. Проте невідомо, чи легко буде створити шаблон мотивації з такою ціллю. Також не зрозуміло, чи зможе зерно ШІ — навіть ідеально вмотивоване — впоратись зі створенням такої архітектури краще за програмістів.

Вивчення цінностей

Нарешті ми підходимо до важливого, але неоднозначного способу вирішення проблеми прищеплення цінностей ШІ. Ідея в тому, щоб *навчити* ШІ цінностей, які ми хочемо, щоб він мав. Для цього ми повинні дати ШІ певний критерій, який, можливо, неявно виокремлює потрібні нам цінності з усіх можливих. Тоді ми отримали б ШІ, який би діяв згідно зі своїм розумінням цих неявно визначених цінностей. Пізнаючи світ дедалі краще, ШІ вдосконалював би своє розуміння цінностей та їх відповідності критерію.

На відміну від способу використання шаблонної системи мотивації, який передбачає зміну проміжної цілі на кінцеву, навчання не змінює ціль ШІ. У процесі навчання ШІ змінює лише своє розуміння цілі.

А отже, ШІ має володіти критерієм, який дасть змогу відрізнити ознаки правильного припущення про кінцеву мету від хибного. Сформулювати такий критерій може бути складним завданням. По-перше, треба, власне, створити ШІ рівня людини з потужною здатністю до навчання та дослідження структури оточення на основі обмежених даних із сенсорів. Цю частину проблеми поки що винесемо за дужки. Але навіть і без того існують труднощі, характерні саме для прищеплення цінностей. Зокрема, для процесу навчання потрібен критерій, який пов'язуватиме вхідні дані із системи сенсорів із припущеннями про потрібні цінності.

Перш ніж заглиблюватися в деталі реалізації процесу навчання, корисно буде проілюструвати загальний принцип навчання за допомогою конкретного прикладу. Уявімо, що ми записали опис набору цінностей на аркуші паперу, вклали його в конверт і запечатали. Далі ми створили агента з інтелектом людського рівня й наказали йому: «Максимізуй цінності, описані в цьому листі». Що робитиме агент?

Агенту не відомо, що написано на аркуші. Проте він може висувати припущення і на основі будь-яких емпіричних даних і досвіду визначати рівень їхньої правдоподібності. Наприклад, раніше агент міг читати різні тексти, написані людьми, або бути свідком якої-небудь людської поведінки. Тож на основі цих знань він може робити припущення. Не потрібно бути дипломованим психологом, щоб вважатити імовірнішим варіантом змісту записки текст «мінімізувати несправедливість і непотрібні страждання» або «максимізувати дохід інвесторів», а не «вкрити всі озера поліетиленовими пакетами».

Зважуючи свої дії, агент намагатиметься вибрати ту, що ефективніше реалізовуватиме гіпотетичну мету, яку він вважатиме найімовірнішою. Агенту також важливо буде дізнатися більше про те, що насправді написано в листі. Адже, незалежно від того, яка справжня мета, що більше він про неї знатиме, то ефективніше зможе організовувати її реалізацію. Крім того, агент матиме інші конвергентні інструментальні цілі, описані в розділі 7, — цілісність системи цінностей, покращення розумових здібностей, здобуття ресурсів тощо. Проте він не зможе, безоглядно зважаючи на інструментальні цілі, перетворити планету на комп'ютрон і поставити людство під загрозу зникнення, доки матиме ненульову ймовірність того, що серед гіпотетичних цінностей, перелічених у листі, міститься вимога про добробут людства.

У цьому сенсі можна порівняти агента з баржею, яку в різні боки тягнуть кілька буксирів. Буксири тут уособлюють гіпотези про кінцеву мету, а потужність роботи двигуна кожного буксира — імовірність цієї гіпотези. Нові знання про кінцеву мету дають підстави надати тій чи тій гіпотезі більшу (чи меншу) імовірність, отже, і потужність двигуна відповідного буксира змінюватиметься, впливаючи на загальний напрям руху баржі. Траєкторія руху баржі відображає процес пізнання агентом його кінцевої мети, одночасно уникаючи незворотних руйнувань. Згодом, маючи точніше уявлення про кінцеву мету, єдиний буксир, що залишився, доправить баржу морем можливостей до реалізації кінцевої мети за найкоротшим або найсприятливішим маршрутом.

Метафори конверта і баржі покликані проілюструвати загальний принцип, але не враховують деяких важливих технічних деталей.

Спробувавши реалізувати описаний принцип у формальніших рамках, ми чіткіше бачитимемо ці деталі (див. додаток 10).

Додаток 10. Формальна реалізація процесу вивчення цінностей

Формалізація викладу може допомогти чіткіше зрозуміти предмет. Читачі, яким не близькі формальні методи, можуть спокійно пропустити цю частину.

Уявімо спрощену ситуацію, коли агент взаємодіє із середовищем упродовж скінченної кількості окремих циклів[431]. Протягом циклу k агент виконує дію y_k і отримує від органів чуття значення x_k. Тоді історія взаємодій агента із середовищем упродовж життя виражатиметься послідовністю $y_1x_1y_2x_2...y_mx_m$ (скорочено — $yx_{1:m}$ або $yx_{\leq m}$). Усередині кожного циклу агент вибирає дію, спираючись на послідовність попередніх даних з органів чуття.

Спершу розглянемо навчання з підкріпленням (ШІ-НП). У навчанні з підкріпленням агент максимізує майбутню очікувану винагороду. Такий процес виражається рівнянням[432]:

$$y_k = \text{argmax}_{y_k} \sum_{x_k y x_{k+1:m}} (r_k + ... + r_m) P(yx_{\leq m} \mid yx_{<k} y_k).$$

Оскільки агент отримує винагороди за допомогою чуття, послідовність даних з органів чуття $x_{k:m}$ містить послідовність винагород $r_k, \ldots, r_m$.

Як ми вже стверджували раніше, навчання з підкріпленням не підходить для суперінтелекту, який може дійти висновку, що найнадійніший спосіб максимізувати нагороду — це контролювати механізм, який його генерує (вайргединґ). Натомість слабкому агенту легше фізично обмежити доступ до системи підкріплення. Крім того, можна дозволяти надсилати винагороду лише в разі дотримання агентом певних правил. Проте в процесі навчання з підкріпленням агент завжди матиме сильний стимул отримати контроль над джерелом сигналу нагородження і позбавити нас можливості впливати на нього. Тому у своїй основі наші стосунки завжди будуть антагоністичними. У випадку сильного агента це може бути джерелом проблем для нас.

Системи, які не очікують зовнішнього сигналу нагороди, а натомість прагнуть сформувати певний внутрішній стан, теж можуть вдатися до вайргедингґу. Наприклад, у системах типу «суб'єкт–критик» модуль «суб'єкта», який вибирає дію, намагається мінімізувати несхвалення, що надходить від окремого модуля «критика», який оцінює, наскільки результат дії не відповідає заданим параметрам. Проблема може виникнути, якщо суб'єкт дії зрозуміє, що може мінімізувати несхвалення, змінивши або видаливши програмний код критика — як диктатор, що розпускає парламент та націоналізує засоби масової інформації. Для обмежених систем проблему можна вирішити заборонивши суб'єкту змінювати модуль критика. Але достатньо розумний і винахідливий суб'єкт завжди знайде спосіб вплинути на критика (який, по суті, є звичайним програмним процесом у комп'ютері)[433].

Перш ніж перейти до навчання агента цінностей, розгляньмо проміжний варіант агента — ШІ, який максимізує корисність спостережень (ШІ-МКС). У такому варіанті агента сумарна винагорода ($r_k + ... + r_m$) у рівнянні навчання замінюється функцією корисності u історії взаємодій ШІ:

$$y_k = \text{argmax}_{y_k} \sum_{x_k y x_k + 1 : m} U(yx_{\leq m}) P(yx_{\leq m} \mid yx_{<k} y_k).$$

Таке формулювання дає змогу обійти проблему вайргедингґу, бо функція корисності всієї історії взаємодії може бути реалізована так, щоб розпізнавати спроби самообману (або випадки, коли агент не докладає достатньо зусиль, щоб отримати достовірні дані з органів чуття).

А отже, ШІ-МКС *у принципі* дає змогу обійти проблему вайргедингґу. Проте щоб скористатися такою можливістю, ми мусимо створити відповідну функцію корисності, яка була б визначеною для множини всіх ймовірних історій взаємодії — завдання, яке здається неймовірно складним.

Природніше визначити функцію корисності безпосередньо в термінах можливих світів (чи властивостях можливих світів, чи припущеннях про них), ніж у термінах власних історій взаємодії агента. Завдяки цьому підходу можна спростити формулу поняття оптимальності ШІ-МКС:

$$y_k = \text{argmax}_y \sum_w U(w)P(w \mid Ey).$$

У цьому виразі *E* позначає всі сенсорні дані, які доступні агенту (на момент ухвалення рішення), а *u* — функцію корисності, яка визначає корисність певної множини світів. Оптимальний агент вибирає дію, яка максимізує очікувану корисність.

Найскладніший елемент цих формул — це функція корисності *u*. Тут ми повертаємося до проблеми вивчення цінностей (ШІ-ВЦ). Щоб агент міг самостійно навчаючись сформувати правильну функцію корисності, ми мусимо вдосконалити наш вираз, щоб той допускав невизначеність корисності. Це можна зробити так (ШІ-ВЦ)[434]:

$$y_k = \text{argmax}_{y \in \mathbb{Y}} \sum_{w \in \mathbb{W}} P(w \mid Ey) \sum_{w \in \mathbb{U}} U(w)P(\mathcal{V}(U) \mid w).$$

У цьому виразі $\mathcal{V}(\cdot)$ — залежність між функцією корисності та припущенням про функцію корисності. $\mathcal{V}(U)$ — припущення, що функція корисності *U* задовольняє вимоги *критерію цінності*, який утілений у $\mathcal{V}$[435].

Отже, для того щоб визначити, якою має бути наступна дія, потрібно зробити такі кроки. По-перше, знайти умовні ймовірності кожного можливого світу *w* (за наявної інформації із сенсорів та припущення виконання дії *y*). По-друге, для кожного світу *w* та кожної можливої функції корисності *u* обчислити умовну ймовірність, що *u* задовольняє критерій цінності $\mathcal{V}$ (за умови настання цього світу *w*). Далі знайти середню корисність кожного світу по всіх можливих функціях із класу *U*, зважуючи їхнє значення на знайдені перед цим умовні ймовірності. Після чого, комбінуючи отримані середньозважені корисності світів та їхні умовні ймовірності, знайдені на першому кроці, отримаємо очікувану корисність дії *y*. Для того щоб вибрати оптимальну дію, треба провести всі наведені розрахунки для кожної можливої дії *y*, а потім вибрати ту, яка матиме максимальну очікувану корисність (використовуючи певну довільну функцію для вибору у випадку кількох можливих претендентів). У такому вигляді наведена процедура (із повним перебором усіх можливих корисностей у всіх можливих світах, за умови вчинення всіх можливих дій)

потребує неймовірних обчислювальних ресурсів. ШІ буде повинен оптимізувати та спростити цей алгоритм.

Тепер поміркуймо, як визначити критерій цінності $\mathcal{V}$[436]. Щойно ШІ матиме адекватне розуміння критерію цінності, він зможе переходити до оцінки можливих світів. Застосовуючи критерій у кожному можливому світі, ШІ знатиме, які функції корисності задовольняють критерій у кожному з них.

Наведену вище формулу для ШІ-ВЦ можна вважати спробою означити й відокремити основну проблему навчання цінностей — проблему представлення. Формалізація також дає звернути увагу на інші аспекти проблеми (як-от визначення просторів $\mathbb{Y}$, $\mathbb{W}$ та $\mathbb{U}$), що теж потребують вирішення, перш ніж підхід можна буде спробувати на ділі[437].

Зокрема, виникає питання: як поставити ШІ завдання «максимізації реалізації цінностей, описаних у листі». (У термінах додатка 10: як визначити критерій $\mathcal{V}$?). Для цього потрібно ідентифікувати місце, де цінності описано. У нашому випадку треба однозначно послатися на лист у конверті. Це, на перший погляд, просте завдання має нюанси. Наприклад, важливо не лише однозначно послатися на конкретний фізичний об'єкт, але на об'єкт у конкретний момент часу. Інакше ШІ може вирішити, що для швидшого виконання кінцевої мети можна просто переписати опис так, щоб новий передбачав реалізацію простіших цінностей (наприклад, упевнитися, що для кожного цілого числа існує більше ціле число). Досягнувши цієї мети, ШІ зможе спокійно перепочити — однак, імовірніше, результатом буде чергова невдала реалізація на зразок описаних у розділі 8. Тим часом ми маємо замислитися, як визначити час. Тут можна вказати на годинник і домовитися, що «час» визначається рухом стрілок цього приладу — проте і тут нас може спіткати невдача: ШІ може зробити висновок, що часом можна маніпулювати, вручну пересуваючи стрілки годинника, що справді не суперечитиме попередньому визначенню. (Насправді ситуація може бути ще більш ускладнена тим, що потрібні цінності не вдасться описати безпосередньо. Імовірніше, їх доведеться вгадувати з опису інтерпретації людським мозком попереднього досвіду, зробленого крізь призму здобутих знань).

Спробу запрограмувати ціль «максимізація реалізації цінностей, описаних у листі» може спіткати ще одна проблема. Навіть якщо в листі буде опис правильної цілі і мотиваційна система агента розумітиме всі використані в ньому поняття, усе одно немає гарантії, що агент інтерпретує опис так, як би ми цього хотіли. Це теж створить небезпеку хибної реалізації, як у розділі 8.

Проблема тут не лише в тому, щоб забезпечити можливість ШІ розуміти наміри людей. Суперінтелект повинен легко здолати це завдання. Важливо, щоб ШІ прагнув реалізовувати цінності в потрібний нам спосіб. Розуміння наших прагнень не гарантує їх виконання: навіть добре розуміючи наші очікування, ШІ може бути байдужим до них (маючи іншу заманливішу інтерпретацію цілі або й узагалі — іншу ціль).

Усе ускладнюється ще й тим, що задля безпеки мотиваційна система зерна ШІ має бути сформована *до* того, як воно зможе повністю сприймати людські поняття чи розуміти людські наміри. Для цього потрібно, щоб у підсистемі мотивації когнітивної системи зерна ШІ існувало спеціальне місце для зберігання кінцевої цілі. Важливо, щоб когнітивну систему можна було вдосконалювати, щоб ШІ, навчаючись, мав змогу розширювати можливості представлення понять і ставав дедалі інтелектуальнішим. Під час розвитку ШІ може переживати етапи, подібні до наукової революції, коли уявлення про світ докорінно змінюється. Щось на зразок онтологічної кризи, унаслідок якої доводиться відмовлятися від уявлень, які виявилися хибними чи базувалися на ілюзіях. Проте від долюдського рівня і аж до досягнення галактичної суперінтелектуальності лінія поведінки ШІ має бути переважно зумовлена незмінною кінцевою метою, яку він може лише краще зрозуміти внаслідок розумового зростання. Усе-таки, імовірно, його розуміння значно відрізнятиметься від розуміння цієї мети програмістами: не буде лихим, а радше правильнішим і доречнішим. Як цього досягти? Поки що на це питання немає відповіді[438] (див. додаток 11).

Словом, не до кінця зрозуміло, як навчити ШІ насправді потрібних цінностей. (А втім, у додатку 12 наведено деякі нові думки щодо цього). Наразі цей підхід варто вважати перспективним напрямом досліджень, не більше. У разі появи успішної реалізації така технологія

буде найбільш досконалим виконанням завдання прищеплення цінностей ШІ. Крім того, вона пропонує природний спосіб запобігання думкозлочинам, адже ШІ, який робить зважене припущення про те, які цінності програмісти могли забажати надати йому, матиме достатньо підстав відкинути думкозлочин як небажану чи принаймні сумнівну цінність.

Насамкінець варто торкнутися ще одного дуже важливого запитання: «Що нам написати в листі?». Яких цінностей нам варто вчити ШІ? Проте відповідь на це питання не залежить від способу прищеплення цінностей. Повернемося до нього в розділі 13.

Додаток 11. ШІ, який хоче дружити

Елізер Юдковський спробував описати деякі особливості архітектури зерна ШІ, завдяки яким воно поводитиметься подібно до описаної вище моделі. ШІ використовуватиме те, що він називає «семантикою зовнішнього посилання»[439]. Для ілюстративності припустимо, що ми прагнемо навчити ШІ бути «дружнім». Спочатку система намагається визначити властивість *Д*, але не має достатньо інформації про неї. Їй відомо лише, що це абстрактна властивість, а коли програмісти говорять про «дружність», вони намагаються передати інформацію про *Д*. Оскільки ШІ, зрештою, прагне реалізувати *Д*, для нього важливо дізнатися про це більше. Що більше інформації про *Д* має ШІ, то більше вона визначає його дії. А отже, можна сподіватися, що ШІ ставатиме дружнішим.

На ранніх стадіях, поки інформація ШІ про *Д* ще неповна, програмісти можуть допомагати процесу навчання і знижувати ризик катастрофічних помилок, надаючи ШІ «твердження програмістів» — гіпотези про природу та сутність *Д*, яким від початку надано високий рівень імовірності. Так, наприклад, гіпотеза правдивості твердження «обманювати програмістів не дружньо» може мати високу апріорну ймовірність. Проте такі твердження не будуть «істиною за визначенням», незаперечною аксіомою. Радше це будуть вихідні дані про дружність, авторитетні судження, до яких раціональний ШІ прислухатиметься —

принаймні допоки довірятиме епістемічним здібностям програмістів.

Пропозиція Юдковського передбачає також використання «семантики причинної валідності». Основна ідея в тому, щоб ШІ робив не точно те, що його просять програмісти, а натомість (приблизно) те, що його хотіли попросити зробити. Адже програмісти можуть помилятись у своїх спробах пояснити зерну ШІ, що таке дружність. Ба більше, вони можуть самі не до кінця розуміти справжню природу дружності. Тому варто зберегти можливість ШІ знаходити й виправляти помилки в міркуваннях програмістів, а також робити правильні висновки із їхніх недосконалих пояснень. Зокрема, ШІ повинен могти візуалізувати причинно-наслідкові процеси, якими програмісти послуговуються в поясненнях чи розмовах про дружність. Наприклад, ШІ розумітиме, що програміст може допустити описку, вводячи інформацію про дружність, і її треба буде знайти й виправити. Уагалі ШІ має бути готовий знаходити й виправляти помилки та неточності в інформації про дружність, яка надходить з будь-якого джерела через програмістів, незалежно від причини появи спотворень (зокрема, епістемічних). В ідеалі такий ШІ зрештою повинен позбавитися від будь-яких когнітивних упереджень і хибних уявлень, які свого часу заважали програмістам сповна осягнути суть поняття дружності.

Додаток 12. Кілька свіжих ідей (на майбутнє)

Ще один можливий підхід до проблеми мотивування ШІ, який ми назвемо «Радуйся, Маріє», полягає у припущенні, що десь у Всесвіті існує (або існуватиме) цивілізація, яка успішно пережила вибухоподібне зростання інтелектуальних здібностей і має близькі нам цінності. Ідея полягає у створенні ШІ, якому ми поставимо завдання чинити так, як того хотіли б такі гіпотетичні суперінтелектуальні цивілізації[440]. Це може виявитися легше, ніж створити ШІ і безпосередньо прописувати йому потрібні нам цілі.

Для цього нашому ШІ *не* треба буде шукати контакту із цими суперінтелектами. Достатньо буде, щоб він *уявив*, якими б могли бути їхні бажання. Наш ШІ повинен буде змоделювати результат

розвитку суперінтелекту зі згаданими вихідними умовами. З розвитком інтелектуальних здібностей його прогноз ставатиме дедалі точнішим. Проте немає потреби у високій точності. Достатньо визначитися з множиною найімовірніших варіантів появи таких гіпотетичних суперінтелектів. Тоді наш ШІ визначить узагальнений набір кінцевих цілей цих можливих суперінтелектів, зважених на ймовірності їхнього виникнення.

Для цієї версії підходу «Радуйся, Маріє» нам треба надати нашому ШІ кінцеву мету, яка використовує цінності іншого суперінтелекту. У деталях незрозуміло, як це зробити. Суперінтелектуальні агенти можуть мати структурні особливості, завдяки яким певний (гіпотетичний) програмний детектор зможе знаходити їх у моделях світів, створених нашим ШІ. Потім детектор повинен буде якось зчитувати цінності цих гіпотетичних суперінтелектів[441]. Якщо зможемо створити такий детектор, то зможемо і встановити отримані цінності нашому ШІ. Щоправда, може трапитися, що нам доведеться створити його до того, як наш ШІ сформує власну систему представлення понять і цінностей. Тобто детектор буде змушений працювати з незнайомою системою представлення й намагатися розпізнати в ній ознаки ШІ. Усе це здається складним, але, можливо, якесь рішення знайдеться[442].

Якщо вдасться створити робочу базову версію, буде чіткіше видно й перспективні напрями розвитку цієї ідеї. Наприклад, замість пошуку (можливо, зважених) ознак *будь-якого* позаземного суперінтелекту, кінцева мета нашого ШІ може визначати критерії, за якими він спочатку вибиратиме підмножину можливих суперінтелектів (серед яких уже потім шукатиме того, чиї цінності найближчі до наших). Критерієм звуження кола пошуку може бути, наприклад, комплекс передумов появи цих суперінтелектів. Такі характеристики процесу виникнення суперінтелекту (визначені в структурованих термінах) можуть корелювати з тим, наскільки його цінності відповідають нашим. Так, імовірно, ми будемо більш схильні довіряти суперінтелекту, який походитиме від емуляції цілого мозку, чи зерна ШІ, розвиток якого відбувався без використання еволюційних алгоритмів або

просто достатньо повільно, щоб його перехід до суперінтелектуальності був добре контрольованим. (За допомогою нормування передумов появи орієнтовних суперінтелектів ми зможемо ослабити вплив тих, які схильні до самокопіювання, а отже, зменшити мотивацію до таких дій і нашого ШІ). Доступні також інші варіанти вдосконалень.

Для того щоб спосіб «Радуйся, Маріє» спрацював, потрібна віра в існування інших суперінтелектів, які значною мірою поділяють наші цінності[443]. Це є основним недоліком цього підходу. Але технічні перешкоди його реалізації можуть бути меншими, ніж проблеми реалізації інших способів. Тому варто паралельно вести дослідження і в цьому напрямі, щоб в разі крайньої потреби у швидких діях мати запасний варіант.

Іншу ідею вирішення проблеми прищеплення цінностей нещодавно запропонував Пол Крістіано[444]. Як і з «Радуйся, Маріє», він пропонує спростити вивчення штучним інтелектом цінностей ШІ і замість ручного визначення критерію цінностей удатися до хитрощів. На відміну від «Радуйся, Маріє», він не потребує віри в існування інших суперінтелектів, придатних для того, щоб стати прототипами для нашого ШІ. Пропозицію Крістіано складно пояснити в кількох словах: вона базується на послідовності міркувань, проте спробуємо визначити хоча б її основні елементи.

Припустимо що ми маємо (а) точний математичний опис мозку конкретної людини та (б) математично визначене віртуальне середовище, що містить ідеальний комп'ютер з нескінченною кількістю пам'яті та обчислювальних ресурсів. З (а) та (б) ми можемо математично визначити функцію корисності *U* як результат роботи людського мозку (а) у середовищі (б). *U* буде математично визначеним об'єктом, проте (через обчислювальні обмеження) *повністю* описати його реалізацію ми поки що не можемо. Однак *U* можна використати як критерій цінностей для навчання ШІ. Він може, за допомогою різних евристичних механізмів, визначати ймовірність тих чи тих гіпотез про *U*.

В ідеалі *U* мала б бути саме тією, потрібною нам функцією корисності. Людина з відповідною підготовкою могла б створити її, маючи нескінченні обчислювальні ресурси — наприклад,

створивши астрономічну кількість екземплярів себе для пришвидшення аналізу або створивши проміжну функцію корисності для оптимізації процесу створення основної. (Тут ми забігаємо наперед у тему про «когерентне екстрапольоване бажання (coherent extrapolated volition)» розділу 13).

Дати математичне визначення ідеалізованого середовища ніби нескладно: ми здатні математично описати комп'ютер будь-якої потужності, та й для опису, скажімо, кімнати, у якій він перебуватиме, можна використати доступні нам програми віртуальної реальності. Але як математично точно описати конкретний людський мозок? Очевидним способом буде використати технологію емуляції цілого мозку, але що, як ця технологія так і не з'явиться до того часу?

Саме в цьому і полягає інноватівність пропозиції Крістіано. Він доводить, що для отримання математично вираженого критерію цінностей нам не потрібна робоча реалізація моделі конкретного мозку: тобто запускати її виконання нам не доведеться. Потрібне лише її абстрактне математичне *визначення* (згодиться навіть імпліцитне й безнадійно складне) — це може значно спростити завдання. За допомогою функціонального нейросканування й інших інструментів ми, ймовірно, зможемо колись зібрати гігабайти інформації про роботу мозку вибраної людини. Тоді може виявитися, що найпростіша робоча модель мозку — це його цифрова емуляція. І хоч *реалізувати* таку модель на основі зібраних даних може виявитися занадто складно для нас, проте ми зможемо *визначити* таку модель, посилаючись на ці дані, за допомогою чітко визначеного математичного способу вимірювання складності (своєрідного варіанта колмогоровської складності, який ми використовували в додатку розділу 1)[445].

Модуляція емуляції

Для емуляції цілого мозку проблема прищеплення цінностей має свою специфіку. Тут не годяться методи, які ґрунтуються на точному розумінні алгоритмів та архітектури. З іншого боку, методи доповнення, які недоступні для штучних інтелектів *de novo*, можуть

бути ефективними для емуляцій (чи покращеного біологічного мозку)[446].

За допомогою методу доповнення та інших засобів можна налаштувати цілі, успадковані системою від прототипу. Можна спробувати вплинути на мотиваційний стан емуляції за допомогою, наприклад, цифрових еквівалентів психоактивних речовин (або — у біологічних системах — хімічних препаратів). Уже існує обмежена можливість фармакологічно впливати на цінності і мотивації[447]. У майбутньому ж можуть з'явитися препарати з точнішим і прогнозованішим ефектом. Цифрова природа емуляцій значно прискорить розробки в цьому напрямі, зокрема завдяки можливості проведення контрольованіших експериментів і прямому доступу до всіх ділянок емульованого мозку.

Проведення експериментів на емуляціях може наштовхнутися на етичні проблеми, подібні до тих, що виникають під час проведення досліджень на живих біологічних організмах. Не всіх їх вдасться вирішити за допомогою форми надання добровільної згоди. Ці ускладнення (законодавче регулювання чи моральні обмеження) можуть сповільнити рух через емуляції, натомість стимулюючи пошук способів маніпулювання мотиваційною структурою емуляцій. У результаті може виявитися, що емуляції досягнуть потенційно небезпечного суперінтелектуального рівня, перш ніж вдасться знайти надійний спосіб сформувати чи перевірити їхні кінцеві цілі. З іншого боку, моральні ускладнення процесу створення суперінтелектуальної емуляції можуть привести до того, що першими успіху досягнуть менш прискіпливі та доброчесні наукові колективи. Однак, безумовно, неприпустимо поступатися моральними стандартами, навіть задля свободи наукового експериментування із цифровим мозком. Інакше ми ризикуємо обтяжити свою совість відповідальністю за величезну кількість вчиненого зла. Тож краще в розробках технологій високої стратегічної цінності уникати наукових стратегій, пов'язаних з експериментами над цифровими розумами.

Але не все так однозначно. Можна заперечити, що під час досліджень емуляцій ризик порушень моральних норм менший, ніж у процесі досліджень ШІ. Адже визначити моральний статус емуляції легше, ніж синтетичної інтелектуальної системи із цілком іншою архітектурою. А

помилка у визначенні морального статусу ШІ певного типу чи його підпроцесу може спричинити значні моральні порушення. Наприклад, уже сьогодні програмісти під час відпрацювання алгоритму навчання з підкріпленням створюють багато екземплярів тестових агентів і піддають їх великій кількості різноманітних впливів, а потім спокійно знищують їх після кожного сеансу навчання. Незліченна кількість таких ШІ створюється щодня не тільки в наукових лабораторіях, а й в іграх, де вони змушені грати проти людини. Вважається, що вони занадто примітивні, щоб мати будь-який моральний статус, проте чи можемо ми бути в цьому впевнені? Що важливіше: чи можемо ми бути впевнені, що зупинимося вчасно, перш ніж наші програми почнуть по-справжньому страждати?

(У розділі 14 ми повернемося до ширших стратегічних міркувань, порівнюючи бажаність шляхів емуляції та ШІ).

Інституційна структура

Деякі інтелектуальні системи складаються з окремих частин, які мають власний інтелект і спроможні діяти як самостійні агенти. Доступним прикладом тут можуть бути підприємства й держави: вони утворені з людей, за певних умов їх можна розглядати як цілісні автономні суб'єкти. Мотивація таких складних систем залежить не лише від мотивації окремих елементів, але й від того, як організована їхня взаємодія. Наприклад, для організованої групи, яка перебуває під сильним диктаторським впливом одного з елементів, може здаватися, ніби воля групи збігається з волею елемента, що виконує роль диктатора. Тоді як інша група, створена на засадах демократії, може діяти так, ніби її воля є утвореною з мотивів її окремих елементів. Але може трапитися, що діяльність організації завдяки деяким управлінським інститутам залежить не лише від волі її елементів. (Принаймні теоретично, тоталітарна держава, ненависна для усіх її громадян, може існувати і далі завдяки механізмам, які перешкоджають громадянам узгодити дії, спрямовані на її повалення. Громадянину значно вигідніше слухняно відігравати свою роль у функціонуванні такої держави, ніж одноосібно протистояти їй).

Тому, створивши відповідні інститути в комплексній системі, можна впливати на її ефективну мотивацію. У розділі 9 ми згадували соціальну інтеграцію як можливий метод контролю здібностей. Тоді ми зосередилися на особливостях функціонування агента у світі рівних йому суб'єктів. Тепер же говоримо про те, що відбувається всередині інтелектуальної системи: як її воля визначається внутрішньою організацією. Тобто тут ідеться про метод відбору мотивації. Ба більше, оскільки йдеться про внутрішні інституції системи, такі методи розробки не потребують масштабної соціальної інженерії чи соціального реформування, тому доступні окремим дослідним проектам у сфері ШІ, навіть за умови несприятливих зовнішніх соціоекономічних чи міжнародних обставин.

Особливо корисним може бути використання інституційного структурування у складі методів доповнення. Якщо взяти за основу агента з потрібною мотивацією, інституційні вдосконалення можуть стати додатковим запобіжником, фактором упевненості, що система не втратить потрібного мотиваційного вектора.

Наприклад, уявімо систему, в основу якої покладені певним чином умотивовані людиноподібні агенти — нехай це будуть емуляції. Ми маємо на меті розвинути їхні розумові можливості, але хочемо запобігти викривленню початкової мотивації нашої системи. Одним зі способів досягнення цієї мети може бути створення системи, у якій окремі емуляції функціонують як субагенти. Кожне нове вдосконалення реалізується в певній дослідній групі субагентів. Потім за результатами панельного порівняння з рештою емуляцій роблять висновок про відсутність згубних побічних впливів такого вдосконалення на роботу системи загалом і можливість його реалізації в інших елементах системи. Якщо ж вдосконалені субагенти функціонують неналежно, вони виключаються з програми оновлень і їхні дані не враховуються у функціях ухвалення рішень (принаймні допоки система загалом не досягне рівня, на якому пошкоджених агентів можна буде безпечно реінтегрувати)[448]. Удосконалені агенти, які визнані пошкодженими, можуть у чомусь бути кращими від інших агентів, але інституційна структура системи та власне перебування в меншості не дасть їм поширити свій вплив на всю систему чи зашкодити її роботі. У такий спосіб колективний інтелект системи

може розвиватися поступово, поетапно запроваджуючи нові здібності, і надійність системи може перевірятися на кожному етапі розвитку завдяки незначній різниці в можливостях окремих субагентів.

Вплив інституційної розробки на безпекові характеристики системи обмежений точністю оцінки субагентів. Проблема тестування ускладнюється через різну точність наборів тестів для різних когнітивних рівнів агентів. Цьому можна частково зарадити, зменшуючи крок змін і збільшуючи час тестування, але така обережність має свою ціну: збільшується вартість розробки і зменшується швидкість поступу (за умови жорсткої конкуренції така обережність може позбавити проект переваги). Крім того, малі зміни та ґрунтовне тестування не позбавляють ризику появи інших незначних помилок, що, згодом накопичившись, можуть призвести до суттєвих відхилень у мотивації системи. Вони можуть проявитися лише після досягнення системою суперінтелектуальності. Зменшити такий ризик можна, розширивши період тестування характеристик на кілька циклів когнітивного вдосконалення. Тоді після успішних тестів у попередніх циклах елементи проходитимуть повторні випробування, щоб забезпечити відсутність негативної динаміки.

В ідеалі тестування мусить відбуватися постійно, а не лише протягом тестового періоду. Достатньо інтелектуальні субагенти, якщо знатимуть, що їх оцінюють, можуть приховати певні цілі. Для реалізації постійного тестування потрібна ієрархічна структура, де частина субагентів на певному рівні розвитку перевіряє діяльність інших, які мають дещо вищий когнітивний рівень. На найнижчому рівні когнітивних можливостей (але на найвищому щаблі ієрархії авторитету) буде нерозумний і повільний принципал. Ця людина, ніби недоумкуватий король, який за допомогою недолугих радників та через посередніх наглядачів керує здібним народом. Інакше кажучи, він як еволюційна основа нашого мозку — базальні ганглії або «мозок рептилій» — по суті, змушує еволюційно новіший та значно розумніший неокортекс забезпечувати потреби таких базових функцій, як харчування та спарювання.

Виникає питання: наскільки непорушною може бути така інверсна меритократія? Якщо людина-принципал спробує скасувати попереднє оновлення після того, як виявить порушення функціонування

багатьох суперінтелектуальних субагентів, чи зможе така інституційна структура ефективно протистояти можливому опору з боку агентів? На основі історії та сучасності людства можемо визнати: ні, це неможливо. (Звісно, я не стверджую, що до влади приходять лише гідні, але ж точно не мавпи). Проте у програмній імплементації спільноти не все так однозначно[449].

Уявімо цифрову ієрархічну структуру, у якій на кожних двох робітничих агентів припадає один підготований поліцейський агент. На кожних двох поліцейських існує один агент спеціальної поліції. Крім того, за кожними двома агентами спеціальної поліції наглядає агент суперспеціальної поліції — і так далі, рівень за рівнем. Кількість представників кожного наступного рівня не більш ніж удвічі менша за кількість нижчого. Далі уявімо, що архітектура цього суспільства передбачає значні переваги вищого рівня над нижчим. Так, наприклад, наглядач може спостерігати та фіксувати вчинки підлеглих, водночас перебуваючи ніби за дзеркальним склом — підлеглий не може бачити наглядача чи чути те, чого той не хоче, щоб він чув. Нехай наглядач має панель із кнопками, натискаючи на які, він може покарати чи винагородити підлеглого, тимчасово припинити його діяльність, видалити його чи відновити попередній стан. Такі можливості дадуть наглядачам більшу владу, ніж у найтоталітарніших організаціях усіх часів.

І це ще не всі можливі засоби контролю. Наглядачі можуть застосовувати засоби віртуальної реальності, щоб перевіряти функціонування підлеглих у гіпотетичних сценаріях (як-от реакція на пропозицію приєднатися до змови чи саботувати роботу за відсутності видимого контролю). Наглядачі можуть мати змогу прослуховувати внутрішні монологи підлеглих та безпосередньо впливати на їхній емоційний стан.

А отже, наглядач отримує значні переваги[450]. І за кожним наглядачем спостерігатиме наглядач вищого рівня, контролювати кожен його крок, упевнюватися, що він належно виконує свої управлінські обов'язки. Така система зможе стабільно функціонувати, навіть якщо кожен вищий рівень ієрархії буде інтелектуально слабшим за нижчий[451]. Тоді на вершині всієї адміністративної структури з багатьох

суперінтелектуальних агентів може бути невелика група людей і успішно керувати нею[452].

Витрати ресурсів, потрібні на підтримку такої адміністративної структури, можуть бути значні, проте не конче надмірні. Так, якщо кожен нижчий рівень ієрархії вдвічі численніший за попередній, надлишкові витрати обчислювальної потужності становитимуть сто відсотків від корисних витрат. Якщо когнітивний рівень наглядачів нижчий від підлеглих і потребує менше ресурсів, то надлишкові витрати будуть ще менші. Змінюючи пропорцію наглядачі/підлеглі можна збільшувати контрольованість системи або зменшувати надлишкові витрати ресурсів. В умовах високої конкуренції можливість скорочення витрат буде суттєвим фактором (див. розділ 11), тоді як у спокійніших умовах надлишкові витрати не створюватимуть значних проблем. Варто також враховувати необхідність витрат часу на створення та налагодження такої ієрархії. Добре забезпечений проект може паралельно розробляти ШІ та управлінську структуру для його контролю. Для менших же проектів та з високою конкуренцією за технології таке додаткове навантаження може виявитися недозволеною розкішшю.

Варто також згадати ще один аспект: високий ризик думкозлочинів у такій структурі[453]. Описана нами картина схожа на північнокорейський трудовий табір. Проте є способи позбутися принаймні моральних проблем для таких систем — навіть якщо їхніми субагентами будуть емуляції з повноцінним моральним статусом. Насамперед, система може будуватися на добровільності. Кожний субагент може мати право будь-коли припинити участь у роботі системи[454]. Можна зобов'язати систему після завершення небезпечної фази інтелектуального вибуху у сприятливіших умовах перезапускати зупинені емуляції, до того зберігши їхній стан у пам'яті. Тим часом інші добровольці матимуть змогу функціонувати в комфортних віртуальних середовищах, у них буде вдосталь часу для сну та відновлення. Такі умови можуть стати причиною додаткових витрат. Утім щедро фінансований проект за відсутності конкуренції зможе собі це дозволити. Якщо ж ситуація навпаки буде напруженішою та конкурентнішою, такі витрати можуть виявитися надмірними, хіба що

забезпечення цих умов буде невід'ємною вимогою ринкового регулювання, обов'язковою для всіх учасників.

У нашому прикладі субагентами були емуляції. Чи є антропоморфізм елементів системи необхідною умовою для використання інституційного структурування? Чи цей інструмент однаково ефективний і для інтелектуальних систем на базі інших типів ШІ?

На перший погляд така думка викликає скепсис. Незважаючи на багатий досвід управління людьми, ми досі не можемо передбачити початок та результати революцій. Соціальні науки лише здатні пояснити деякі загальні тенденції[455]. Ми не здатні точно оцінити стабільність соціальної структури звичайних людей (про яких маємо вдосталь даних). Тож ніби годі й сподіватися створити точну та стабільну соціальну структуру для розумово вищих людиноподібних агентів (про яких у нас немає жодних даних), не кажучи вже про штучних інтелектуальних агентів (архітектура розуму яких не має нічого спільного з нашим мозком).

Проте не все так безнадійно. Люди і подібні до них інтелектуальні системи складні, натомість архітектура штучного агента може бути порівняно простою. Те саме можна сказати і про мотивацію штучного агента. Ба більше, цифрових агентів (незалежно від архітектури — чи то емуляція, чи ШІ) можна копіювати. Така можливість здатна спричинити революцію менеджменту, як свого часу поява взаємозамінних елементів змінила промислове виробництво. Це разом із можливістю спочатку працювати з агентами, які не мають жодних здібностей, та можливістю створювати інституційну структуру, реалізуючи всі описані вище методи контролю, дасть змогу досягати потрібних результатів. Тобто створити безвідмовну інтелектуальну систему, яка не опиратиметься волі творців і буде надійнішою та стабільнішою за аналогічні людські системи.

Проте, з іншого боку, штучні агенти можуть не мати тих атрибутів, які допомагають нам передбачати дії анторопоморфних агентів. Їм можуть бути не властиві соціально зумовлені емоціональні зв'язки, які керують людською поведінкою, як-от боягузливість, пиха, сумління. Також штучні агенти можуть бути не схильними до дружніх та родинних почуттів. Вони не матимуть несвідомих фізіологічних реакцій, як-от мова тіла, яка часто виказує наші приховані наміри. Усе

це може стати дестабілізаційними факторами в роботі інституційної структури. Окрім цього, внаслідок незначних, на перший погляд, змін в алгоритмах штучні агенти можуть демонструвати раптове зростання когнітивних можливостей. У якийсь момент штучні агенти можуть свідомо вибрати шлях принесення людства в жертву безжальному процесу оптимізації[456]. Суперінтелектуальні агенти можуть мати здатність діяти злагоджено за мінімальної комунікації (наприклад, моделюючи гіпотетичні реакції один одного на різні зовнішні обставини). Через такі особливості навіть випробувані часом інституційні структури можуть виявитися неефективними.

Тому важко точно оцінити ефективність методу інституційного структурування для контролю антропоморфних і штучних агентів. З погляду управління ризиками використовувати такі методи однозначно варто. Наявність суспільних інститутів із вивіреною системою стримувань і противаг лише збільшить безпечність системи загалом — або принаймні точно не зменшить. Проте навіть у цьому немає стовідсоткової певності. Як джерело складності системи, така структура може запроваджувати нові ризики і вразливості, яких не було б, якби суперінтелект мав унітарну структуру — не складався з інтелектуальних субагентів. Однак немає сумніву, що методи інституційного структурування варто досліджувати[457].

Синопсис

Особливості систем мотивації штучних агентів наразі недостатньо добре вивчені. Невідомо, як передавати людські цінності цифровим системам, навіть якщо вони мають інтелект людського рівня. Розглянувши кілька підходів до цієї проблеми, ми можемо впевнено відкинути деякі з них, як позбавлені перспективи. Проте інші можуть бути результативними і тому їх варто досліджувати. Підсумки наведено в таблиці 12.

Таблиця 12. **Підсумки способів прищеплення цінностей інтелектуальним системам**

Безпосереднє представлення	Може виправдати себе як спосіб прищеплення цінностей у межах підходу одомашнення. Невиправданий для прищеплення складніших цілей
Еволюційна	Не надто перспективний засіб. Потужний пошуковий алгоритм може

селекція	віднайти реалізацію, що задовольняє формальні критерії, але не відповідає нашим намірам. Ба більше, якщо процедура перевірки потенційних реалізацій — навіть тих, які не відповідають критеріям успішності — передбачатиме їх запуск та виконання, це може стати джерелом додаткової небезпеки. Також еволюція пов'язана з більшою небезпекою думкозлочинів, особливо якщо цільова система буде подібною до людського мозку
Навчання з підкріпленням	Існують кілька різних методів навчання з підкріпленням, але здебільшого вони пов'язані з максимізацією сигналу нагороди. Це створює небезпеку вайргединґу в системах, які досягли високого рівня інтелектуальності. Тому очікувана результативність цього способу невисока
Накопичення цінностей	Люди засвоюють цінності переважно внаслідок життєвого досвіду. Незважаючи на те, що теоретично за допомогою накопичення цінностей можна створити штучного агента з мотивацією, подібною до людини, реалізувати людські підходи до здобуття цінностей в зерні ШІ може виявитися занадто важко. Невдала реалізація може призвести до того, що ШІ інакше узагальнюватиме досвід і сформує не ті цінності, яких ми сподівалися. Для підвищення точності відтворення потрібно більше досліджень цього підходу
Шаблон мотивації	Наразі важко стверджувати, що потрібно для того, щоб система самостійно винайшла високорівневі представлення потрібних нам цінностей, які ми могли б використовувати для визначення її мотивації (і щоб водночас рівень її здібностей не був надто загрозливим). Загалом напрям цікавий і може мати перспективу. (Однак варто розуміти, що будь-який підхід до вирішення проблеми контролю, який відкладає конкретні кроки реалізації до моменту фактичного створення ШІ рівня людини, не виправдовує бездіяльність у безпеці ШІ на більш ранніх етапах)
Вивчення цінностей	Цікавий підхід, який, однак, потребує подальших досліджень, зокрема в напрямі формального опису потрібних цінностей (критерій функції корисності для їх вивчення). У контексті цієї категорії варто згадати перспективні напрями досліджень, як-от сценарій «Радуйся, Маріє» і пропозиція Пола Крістіано (та інші схожі сценарії спрощення вивчення цінностей)
Модуляція емуляції	Якщо основою штучного інтелекту буде емуляція, то перспективним способом формування мотивації можуть бути цифрові аналоги медикаментів чи інших схожих засобів. Щоправда, питання надійності такого класу засобів в умовах зростання рівня інтелекту системи до суперінтелекту залишається відкритим. (Етичні аспекти також можуть

	ускладнити шлях у цьому напрямі)
Інституційне структурування	У структурованій інтелектуальній системі, що складається з емуляцій, можна застосовувати різноманітні методи соціального контролю. Теоретично такий підхід може бути ефективним і для системи, складеної з інших типів ШІ. Емуляції в цьому контексті мають деякі характеристики, що роблять їх передбачуванішими, ніж інші типи ШІ. Проте інші типи теж можуть мати свої переваги. Інституційне структурування здається досить перспективним напрямом ціннісної інженерії ШІ

Розв'язавши проблему прищеплення цінностей, можна буде перейти до наступної: які власне цінності вибрати для прищеплення? Інакше кажучи, що б ми хотіли, щоб хотів суперінтелект? До цієї філософської проблеми ми наразі і перейдемо.

13. ОБИРАЄМО КРИТЕРІЙ ВІДБОРУ

Уявімо, що ми можемо формувати цінності зерна ШІ. Тоді рішення, які саме цінності прищепити ШІ, матиме далекосяжні наслідки. Певну вагу матимуть також деякі інші параметри — теоретичні основи механізму ухвалення рішень ШІ та епістемологія. Але чи не завадить нам наша дурість, неосвіченість та обмеженість зробити правильний технологічний вибір? Як ми зможемо уникнути впливу упереджень та помилок сучасності? У цьому розділі ми дослідимо, як за допомогою непрямої нормативності прив'язати результат діяльності суперінтелекту до важливих нам людських цінностей, переклавши розумові зусилля, необхідні для ухвалення проміжних рішень, на його видатний інтелект.

Для чого потрібна непряма нормативність

Як зробити так, щоб суперінтелект виконував наші бажання? Що ми хочемо, щоб суперінтелект хотів? Досі з цих двох питань ми ставили собі лише перше. Тепер розглянемо друге.

Уявімо, що ми вирішили проблему контролю й нарешті маємо можливість визначати цінності, на яких базується мотивація суперінтелекту, — метою його існування тепер буде їхнє зростання і захист. Які цінності нам вибрати для цього? Вибір не з легких. Якщо суперінтелект отримає вирішальну стратегічну перевагу, то цінності, які ми виберемо для нього, визначатимуть принципи розподілу ресурсів у космічних масштабах. Тобто без перебільшення він вирішуватиме долю Всесвіту.

Очевидно, що ми не маємо права на помилку в нашому виборі. Але які в нас є перспективи її уникнути? Ми можемо помилитися в питаннях моралі; не знати, що для нас добре. Можемо навіть не знати, чого ми насправді хочемо. Схоже, що для формулювання потрібної нам кінцевої мети новоствореного ШІ нам доведеться спершу продертися крізь непролазні філософські хащі. Наївні спроби

розв'язати завдання «в лоб» загрожують плутаниною та, найімовірніше, приречені на провал. У разі невизначеності контексту рішення, ризик зробити неправильний вибір особливо високий. А менш визначеного контексту, ніж вибір правильної мети для штучного суперінтелекту, який у майбутньому визначатиме реальність для людства, — годі й шукати.

Тотальна відсутність єдності суспільства в поглядах на відповідні категорії теорії цінностей підтверджує невтішні шанси безпосереднього підходу до визначення цінностей. Жодна з етичних систем не має підтримки більшості філософів, тож, вочевидь, вони помиляються[458]. Хибність такого підходу підтверджують темпи поширення моральних цінностей, багато з яких ми звикли вважати ознаками прогресу. Так, наприклад, у середньовічній Європі розвага спостерігати за смертю політичних опонентів від жорстоких тортур вважалася цілком гідною і доречною справою. У Парижі у XVI столітті все ще спалювали котів[459]. Усього якісь сто п'ятдесят років тому на півдні Америки на цілком законних підставах і з усією підтримкою місцевих моральних звичаїв широко практикувалося рабство. Дивлячись у минуле, ми бачимо не лише відверто негідну поведінку наших предків, але й обурливі вади їхніх моральних цінностей. Відтоді ми, можливо, трохи отямилися, проте досі далеко не янголи в сенсі моралі. Цілком імовірно, ми все ще перебуваємо під впливом одного-двох хибних уявлень про справжню моральність. Тому в таких умовах нам надто ризиковано вибирати фіксовану кінцеву мету, яка може раз і назавжди визначити напрям історії і, заблокувавши етичний прогрес, призвести до моральної катастрофи.

Навіть якби ми могли раціонально обґрунтувати вибір певної етичної теорії — що нам наразі не під силу, — ми все одно ризикуємо помилитися в суттєвих деталях. Зовні прості моральні теорії можуть приховувати в собі значні ускладнення[460]. Візьмемо, наприклад, напрочуд просту консеквенційну теорію — гедонізм. Спрощено, гедонізм постулює, що будь-яка насолода є цінністю, на відміну від страждання, яке цінністю не є[461]. Навіть якби ми зробили ставку на цю теорію і вона виявилася правильною, залишається багато відкритих запитань. Наприклад, чи варто, вслід за Джоном Стюартом Міллем, віддавати перевагу «вищим насолодам» над «нижчими насолодами»?

Як враховувати інтенсивність і тривалість насолод? Чи може біль нівелювати насолоду? Які стани мозку відповідають моральним насолодам? Чи можна вважати дві точні копії мозку у стані насолоди подвійною кількістю насолоди?[462] Чи існують підсвідомі насолоди? Як оцінювати мізерні шанси отримання надзвичайно сильної насолоди? Як представляти нескінченні популяції?[463]

Помилка у відповіді на будь-яке із цих питань може призвести до катастрофи. Якщо врахувати, що вибором кінцевої мети для суперінтелекту ми намагаємося потрапити не лише у правильну моральну теорію, а в конкретний ланцюг взаємозалежних умов її практичного застосування, то наші шанси на успіх одразу перетворюються в ніщо. Лише дурень може з ентузіазмом сподіватися осушити одним ковтком цю бездонну криницю проблем філософії моралі, вкрутивши свої улюблені відповіді в ліхтар істини зерна ШІ. Мудрість, натомість, підказує, що варто шукати альтернативний, безпечніший шлях.

Саме такий шлях пропонує нам непряма нормативність. Суперінтелект нам потрібен для виконання інтелектуальної роботи: інструментального забезпечення ефективного досягнення кінцевої мети. За допомогою непрямої нормативності можна буде також доручити суперінтелекту інтелектуальну роботу вибору цієї мети.

За допомогою непрямої нормативності можна буде винести за дужки нашу невпевненість у власних прагненнях щодо того, що нам корисно, що морально, а що — ні. Замість того щоб, спираючись на власне (можливо, хибне у своїй основі) розуміння, вгадувати, ми можемо делегувати усвідомлений та обґрунтований вибір кінцевої цінності самому суперінтелекту. Завдяки своїм видатним інтелектуальним можливостям він, імовірно, зможе уникнути помилок і плутанини, які затьмарюють наш розум. Можемо узагальнити цю ідею і виділити евристичний принцип:

Принцип епістемічної переваги
Суперінтелект майбутнього перебуватиме на епістемічно вищому за нас щаблі розвитку: його судження (щодо більшості питань) мають більшу вагу, аніж наші. Тому варто за будь-якої можливості прагнути віддавати перевагу судженням суперінтелекту[464].

Непряма нормативність є наслідком застосування цього принципу до проблеми вибору цінностей. Невпевнені у власній спроможності вибрати правильний нормативний стандарт для суперінтелекту, ми натомість можемо визначити абстрактну умову, згідно з якою суперінтелект сам, як здібніший суб'єкт, імовірно, зможе віднайти потрібний нормативний стандарт своєї роботи. Тоді ми можемо дати суперінтелекту завдання: постійно чинити максимально близько до ідеалу такого неявно заданого стандарту.

Роз'яснимо думку за допомогою кількох ілюстрацій. Спочатку розглянемо «когерентне екстрапольоване бажання», запропоноване Елізером Юдковським. Затим поглянемо на деякі альтернативні варіанти, щоб усвідомити весь спектр доступних напрямів.

Когерентне екстрапольованебажання

Юдковський свого часу запропонував поставити ШІ завдання виконувати «когерентне екстрапольоване бажання» (КЕБ) людства. Ось його визначення КЕБ:

> *Наше когерентне екстрапольоване бажання — це те, чого ми бажали б, якби мали більше знань, могли швидше думати, були ближчими до власного ідеального бачення себе, стали би більш єдними; екстрапольоване в точці майбутнього, де окремі бажання збігаються в одне, а не розбігаються; резонують спільністю, а не перебивають одне одного, роз'єднуючи нас; екстрапольоване в бажаному для нас напрямі, трактоване в потрібній для нас площині*[465].

Пишучи це досить поетичне означення, Юдковський не мав на меті надати практичний спосіб його втілення. Його метою радше було приблизно означити КЕБ і аргументувати його необхідність.

Багато супутніх до запропонованого концепту ідей мають аналоги та прототипи у філософській літературі. В етиці, наприклад, *теорія ідеального спостерігача* пропонує аналізувати нормативні концепти «добра» чи «правильності» в термінах суджень гіпотетичного ідеального спостерігача (де під «ідеальним спостерігачем» розуміють спостерігача свідомого позаморальних фактів, строго логічного, безстороннього, неупередженого тощо)[466]. Однак підхід КЕБ не є (і не

конче мусить бути) моральною теорією. Він не обмежений лиш твердженням про обов'язковий зв'язок між кінцевою метою ШІ і нашим когерентним екстрапольованим бажанням. КЕБ — це лише зручна апроксимація, спосіб представлення вищої цінності, який можна використовувати також і поза контекстом етики. Як основний прототип реалізації підходу непрямої нормативності він заслуговує на уважніший розгляд.

Роз'яснення

Деякі терміни, зацитовані вище, потребують роз'яснення. Під «швидше думати» Юдковський має на увазі «якби ми були розумнішими й більше думали над речами». Під «стали би більш єдними», вочевидь, мається на увазі «якби ми вчилися, розвивалися й самовдосконалювалися в умовах тісного соціального контакту і взаємодії».

«Екстрапольоване в точці майбутнього, де окремі бажання збігаються, а не розбігаються», можна розуміти так: ШІ у процесі ухвалення рішення повинен враховувати ознаку лише якщо результат екстраполяції дає достатньо високу ймовірність її істинності. Якщо передбачити наші бажання на основі описаної ідеалізації не вдається, ШІ ліпше утриматися від будь-яких дій і не покладатися на випадок. Утім, незважаючи на неможливість передбачити багато аспектів наших ідеалізованих бажань, ШІ може вибирати найбезпечніші дії, зберігаючи загальний напрям бажаного розвитку подій. Наприклад, точно оцінивши, що в майбутньому ми не бажатимемо терпіти постійні страждання і щоб Всесвіт перетворився на скріпки, ШІ повинен всіляко запобігати такому майбутньому[467].

«Резонують спільністю, а не перебивають одне одного, роз'єднуючи нас» можна розуміти так: щоб ШІ вибрав певний напрямок руху, потрібна значна узгодженість екстрапольованих бажань окремих індивідів. Менша кількість сильних, чітких бажань може іноді переважувати більшу кількість слабших, невиразних, плутаних бажань. Водночас Юдковський вважає, що поріг ефективності узгодженої опозиції до певного рішення мусить бути досить низький, тоді як поріг ефективності узгодженої підтримки навпаки має бути

більший. «Початкова динаміка КЕБ мусить бути стриманою до ініціатив та чутливою до заперечень», — пише він[468].

«Екстрапольоване в бажаному для нас напрямі, трактоване в потрібній для нас площині»: ці останні слова мають на меті підкреслити, що правила екстраполяції також значною мірою визначаються екстрапольованим бажанням. Індивід може мати, так би мовити, бажання другого порядку — «бажання бажання»: хотіти, щоб його перше бажання не враховувалося при екстраполяції. Наприклад алкоголік, першим прагненням якого є бажання напитися, одночасно може прагнути не мати такого першого бажання. Так само і ми теж можемо мати особливі побажання до процесу екстраполяції, які теж мусять бути враховані.

Можна заперечити, що, навіть маючи вичерпне визначення концепту когерентного екстрапольованого бажання людства, усе одно неможливо — навіть для суперінтелекту — точно передбачити бажання людства в гіпотетичних ідеалізованих умовах, запропонованих концепцією КЕБ. Без інформації про зміст наших екстрапольованих бажань ШІ не матиме достатніх підстав для ухвалення рішення про свою поведінку. Однак, незважаючи на відсутність конкретики щодо нашого КЕБ, на основі інформації про нас можна зробити деякі узагальнені припущення. І це можливо навіть зараз — без суперінтелекту. Так, імовірніше, що наше КЕБ буде пов'язане з життям у достатку та затишку, а не з ув'язненням у темній кімнаті на стільці у вічних муках. Якщо *ми* маємо всі підстави так вважати, то і суперінтелект також може дійти такого висновку. Тому від початку суперінтелект може базувати свої рішення на власній оцінці змісту нашого КЕБ. Так, наприклад, якщо за попередніми оцінками наше КЕБ засуджуватиме як думкозлочин практику використання симуляцій через пов'язаний із цим ризик їхнього страждання, він може утриматися від запуску безлічі симуляцій.

Іншою перешкодою може бути те, що у світі існує надто багато різних звичаїв, способів життя, поглядів на мораль та побут, щоб їх можна було ефективно поєднати в одне КЕБ. Навіть якщо це вдасться зробити, результат може виявитися неочікуваним і не надто корисним. Навряд чи можливо приготувати смачну страву із суміші смаків улюблених страв усіх людей[469]. Проте варто зауважити, що концепція

КЕБ не передбачає змішування докупи всіх способів життя, моральних норм, особистих цінностей. Динаміка КЕБ передбачає активацію за принципом когерентності бажань. Тому суперінтелект в аргументації власних рішень має утримуватися від використання аспектів, стосовно яких навіть після деякої ідеалізації неможливо дійти згоди. За нашою кулінарною аналогією, різні люди та культури можуть мати різноманітні смакові вподобання, але всі погодяться з тим, що їжа не повинна бути отруйною. Тож агент, який керується таким КЕБ, може спрямувати свої зусилля на те, щоб запобігати отруєнню їжі, в іншому не втручаючись у людські кулінарні практики.

Логічне обґрунтування КЕБ

У своїй статті Юдковський наводить сім аргументів на користь підходу КЕБ. Три з них зводяться до думки, що, перебуваючи в руслі гуманітарного та конструктивного спрямування, однак практично неможливо укласти вичерпний перелік конкретних правил, які ще й не матимуть побічних наслідків і не зможуть бути інтерпретовані в небажаний спосіб[470]. Перевага підходу КЕБ у його стабільності та саморегульованості. Він дасть змогу спертися на *причини* наших цінностей, не випробовуючи нашу здатність повністю їх усвідомити, перелічити й однозначно сформулювати.

Інші чотири аргументи розширюють цю основну (і дуже важливу) думку, висуваючи додаткові вимоги до потенційних реалізацій механізму прищеплення цінностей та, вочевидь, натякаючи, що в КЕБ ці вимоги вже враховані.

«Передбачає моральне зростання»

Тобто потенційна реалізація повинна враховувати можливість розвитку моральних цінностей. Як ми вже згадували, є підстави вважати, що наші поточні моральні принципи можуть мати вади, імовірно, є хибними у своїй основі. У такому разі, формулюючи вичерпний моральний кодекс поведінки ШІ, ми неуникно додамо туди всі свої переконання, разом із помилками й упередженнями, блокуючи можливість морального розвитку та зростання. Натомість, використання КЕБ враховує можливість розвитку, адже передбачає, що ШІ діятиме так, як людство хотіло, якби продовжило розвиватися у

сприятливих умовах та, імовірно, позбулося згаданих моральних дефектів та обмежень.

«Захищає майбутнє людства»

Тут Юдковський має на увазі гіпотетичний сценарій створення групою програмістів зерна ШІ. Після створення воно здійснює стрибок до суперінтелектуальності й отримує вирішальну стратегічну перевагу. У такому разі від згаданої групи програмістів фактично залежить доля людства. Без сумніву, такої жахливої відповідальності не побажаєш навіть ворогу. Однак повністю уникнути відповідальності їм не вдасться: будь-який вибір кожного з них, зокрема й припинення участі у проекті, може вплинути на подальшу історію Всесвіту. Саме в підході КЕБ Юдковський убачає спосіб позбавити програмістів привілею — чи тягаря — необхідності безпосередньо визначати майбутнє людства. Тому реалізацією такого механізму, який спиратиметься на когерентне екстрапольоване бажання *всього людства*, а не на власне прагнення чи улюблену моральну теорію окремих його індивідів, програмісти делегують вплив на формування майбутнього всього людства.

«Усуває причини для можливого конфлікту сучасників щодо мотивації»

Розширення бази впливу на майбутнє людства потрібне не лише для того, щоб завадити групі людей втілити власні обмежені погляди, але також запобігти протистоянню навколо питання, хто створить перший суперінтелект. Завдяки підходу КЕБ програмісти (та їхні спонсори) матимуть не більше впливу на кінцевий результат, ніж будь-хто інший. Хоч, звісно, від них залежатимуть деталі реалізації процесу екстраполяції, та й рішення про реалізацію підходу КЕБ, а не будь-якого іншого, ухвалювати саме їм. Важливо запобігати конфліктам, не лише через їхню безпосередню шкоду, але й через руйнівний вплив на атмосферу співпраці навколо процесу створення безпечного та корисного суперінтелекту.

Концепт КЕБ оснований на ідеї забезпечення якомога ширшої підтримки. І не лише завдяки рівномірному розподіленню впливу. Існують інші, глибші причини. Завдяки такому механізму кожна ідеологічна група може сподіватися, що її візія майбутнього зрештою матиме перевагу. Уявіть дебати представника афганського Талібану з

членом Асоціації гуманістів Швеції. Обидва мають дуже відмінні світогляди й те, що один вважатиме утопією, інший може сприймати як дистопію. І жоден з них може не погодитися на компромісний варіант, наприклад, дозволити дівчатам навчатися, але лише до дев'ятого класу, або дозволити навчатися лише шведським дівчатам, а афганським — ні. Водночас обидва можуть схвалити принцип визначення майбутнього людства за КЕБ. Таліб міркуватиме, що, оскільки його релігійні погляди єдино правильні (а він у цьому переконаний), люди рано чи пізно (на його думку) зрозуміють, що гіднішої альтернативи не існує — якби лиш вони облишили свою впертість, упередженість, боягузливість, більше читали священне писання, якби лиш спробували зрозуміти, як світ улаштований, та осягнути, що справді важливо для них самих[471]. А гуманіст може очікувати, що, за всіх цих умов, людство обов'язково прийде до тотального прийняття саме його цінностей.

«Лише людство мусить бути відповідальним за власну долю»

Навряд чи ми хочемо отримати патерналістичний суперінтелект, який постійно наглядає за нами, втручається в найдрібніші питання, аби лиш все відповідало його глобальному плану. Навіть ідеально благородний суперінтелект, вільний від самовпевненості, пихи, нахабності, тупості й інших обмежень людської природи, може так чи інакше обмежувати автономію людини. Ми звикли вільно творити свою долю, навіть якщо це означає, що іноді ми можемо помилятися. Можливо, нам потрібний суперінтелект, який страхуватиме нас у разі наближення катастрофи, але в іншому дозволить самостійно давати собі раду.

З підходом КЕБ така можливість існує. Механізм КЕБ лише забезпечує ініціалізацію системи, спрацьовує напочатку, а потім замінюється, власне, об'єктом когерентного екстрапольованого бажання. Якщо екстрапольованим бажанням людства буде жити під контролем патерналістичного ШІ, то механізм створить такий ШІ і передасть йому владу. Якщо ж натомість екстрапольованим бажанням людства буде демократичний світовий уряд, механізм КЕБ забезпечить його створення й обмежить своє подальше втручання. Врешті, якщо на те виявиться воля людства, кожен отримає свою долю глобальних ресурсів у повне володіння і зможе розпоряджатися ними на свій

розсуд, однак не зазіхаючи на права інших рівноправних власників, а механізм КЕБ непомітно, як сили природи, забезпечуватиме дотримання всіма меж, запобігаючи зазіханням, шахрайству, насильству й іншим порушенням загальної згоди[472].

Отже, результат підходу КЕБ може виявитися мало не будь-яким. Може статися, що екстрапольованим бажанням людства буде, щоб ШІ не робив нічого. Тоді, установивши з достатньо високою імовірністю, що саме цього в майбутньому хотітиме людство, ШІ, який реалізує механізм КЕБ, спокійно припинить діяльність і зупиниться.

Подальші зауваження

Пропозиція підходу КЕБ, описана вище, без сумніву, дуже приблизна. Може існувати безліч варіантів реалізації такого механізму залежно від низки параметрів, яких ми не торкнулися в обговоренні.

Один з таких параметрів — база екстраполяції: чиї бажання буде враховано? Відповідь «усіх» викликає шквал інших запитань. Чи включені до бази також «маргінальні особистості», як-от ембріони, зародки, пацієнти зі смертю мозку, у незворотній комі, з обширною деменцією? Кожна півкуля мозку пацієнта із синдромом розділення враховуватиметься окремо з ваговим коефіцієнтом чи вважатиметься за окрему цілісну персону? Що робити з людьми, які жили в минулому? А з людьми, які ще не народилися? З вищими тваринами й іншими розумними істотами? Цифровими розумами? Позаземними істотами?

Є сенс обмежити вибір дорослими людьми, які житимуть на момент створення ШІ. Питання про подальше розширення бази можна розглянути, здійснивши попередню екстраполяцію. Через малу кількість «маргінальних особистостей» результат може не надто залежати від того, де саме проляже межа вибірки — чи ввійдуть до неї, скажімо, зародки, чи ні.

Зрештою, вилучення певної категорії істот з бази екстраполяції не означає, що їхні інтереси буде відкинуто. Якщо їхні інтереси будуть небайдужі істотам, які потрапляють до бази (дорослим людям), то і результат екстраполяції враховуватиме їх. Проте існує імовірність, що вплив таких зовнішніх інтересів буде значно слабший. Якщо поріг узгодженої підтримки буде достатньо високим (як пропонував

Юдковський), то відповідно низький поріг опозиції збільшує шанси вилучення із фокуса уваги всіх периферійних інтересів, як-от інтереси тварин чи цифрових розумів, небайдужих лише деяким дорослим людям. Такий результат може в перспективі виявитися морально неповноцінним[473].

Концепція КЕБ створювалася почасти для того, щоб усунути причини для протистояння під час створення штучного суперінтелекту. Хоч концепт КЕБ пропонує непогане рішення цієї проблеми, воно все ж не повністю прибирає причини для конфлікту. Для того щоб збільшити свій контроль над майбутнім, егоїстичні особи, групи чи країни можуть прагнути більшого впливу на базу екстраполяції.

Причин бажати такого захоплення може бути кілька. Наприклад, фундатор проекту створення ШІ може вважати, що вся вигода від результатів діяльності ШІ має належати йому. Ця претензія, без сумніву, морально неправомірна. Окрім того, можна зауважити, що такий проект має значну негативну екстерналію — ризик невдалої реалізації та екзистенційної катастрофи для суспільства, яке в такому разі може претендувати на справедливу компенсацію. Співмірна ризику компенсація може бути виражена лише в загальній рівномірній участі суспільства у вигоді від успішної реалізації проекту[474].

Ще одним аргументом на користь узурпації проекту створення ШІ може бути те, що велика кількість людей мають посередні та відверто лихі цілі, і їх врахування в екстраполяції загрожує перетворити майбутнє людства у дистопію. Важко визначити міру доброго й лихого в серці звичайної людини. Так само складно дослідити, як цей баланс залежить від соціальної групи, прошарку, культури та національності. Чи буде вирішальним вплив на результати екстраполяції кращих рис людської природи значної частини семи мільярдів людей, що нині живуть на нашій планеті? Чи варто ставити в залежність від цих результатів долю всіх космічних багатств людства? Відповіді оптиміста та песиміста на ці питання можуть відрізнятися. Та й немає певності, що вилучення груп індивідів з бази екстраполяції гарантуватиме негайну перемогу світла над темрявою. Імовірніше, навіть, що душі, схильні усувати та вилучати інших у гонитві за владою, самі містять у собі чимало темряви.

Ще однією причиною для протистояння може бути недовіра. Конкуренти можуть не довіряти реалізаціям механізму КЕБ проектів-суперників. Прихильники тої чи тої реалізації можуть силою перешкоджати запуску конкурентних ШІ. Тоді всім буде краще, якщо конкуренти знайдуть об'єктивніший спосіб узгодити свої епістемічні розбіжності, ніж силове протистояння[475].

Моделі моральності

Концепт КЕБ — не єдиний можливий спосіб використання непрямої нормативності. Наприклад, замість визначення когерентного екстрапольованого бажання людства, можна поставити ШІ завдання чинити морально правильно. Що, власне, це означає, ШІ має визначити самотужки за допомогою своїх видатних розумових здібностей. Назвемо цю концепцію «моральною правотою» (МП). Вона основана на думці, що ми, люди, не до кінця розуміємо, що правильно, а що — ні, та ще гірше розуміємо філософію поняття моральної правоти та способи його дослідження, тоді як суперінтелект може мати значно досконаліші інструменти для такого завдання[476].

Навіть якщо ставити під сумнів реальність моральності, ми все ще можемо застосовувати підхід МП. Треба лише вказати ШІ, що робити, якщо виявиться, що припущення про реальність моральності хибне. Наприклад, якщо ШІ не вдастся достовірно встановити правдивість жодного нерелятивістського твердження про моральну правоту, він може перейти до реалізації принципу когерентного екстрапольованого бажання або просто вимкнутися[477].

Насправді МП має кілька переваг над КЕБ. Насамперед, МП позбавляє необхідності мати справу з параметрами КЕБ, як-от необхідний ступінь когерентності окремих екстрапольованих бажань, проблеми узгодження підтримки більшості та опозиції меншості і, зрештою, потреби моделювання того соціуму, у якому б наші екстрапольовані бажання «стали би більш єдними».

За таких умов, схоже, ми значно менше ризикуємо порушити моральний баланс майбутнього, узявши завузьку або, навпаки, зашироку базу екстраполяції. Ба більше, як ми показували вище, когерентне екстрапольоване бажання людства може спонукати ШІ до

не надто моральних учинків, тоді як принцип МП завжди матиме своїм пріоритетом моральність. Адже моральні чесноти в людській природі — як коштовний метал у руді: якщо її обробити відповідно до принципу КЕБ, хтозна що ми отримаємо на виході — коштовні чесноти, непотрібний шлак чи отруйний осад?

Крім переваг, МП має і недоліки. Поняття «моральної правоти», на яке спирається принцип МП, — дуже складний концепт, над яким сиділи покоління філософів від часів античності, і досі не дійшли згоди щодо способу його аналізу. Хибне представлення засад «моральної правоти» може спричинити значні моральні вади. Така складність визначення може видатися значною перешкодою. Проте невідомо, наскільки суттєвою може бути ця перешкода, адже концепція КЕБ теж спирається на складні поняття (як-от «знання», «бути ближчими до власного ідеального бачення себе» та «стати більш єдними»)[478]. І навіть якщо ці поняття видаються зрозумілішими, ніж «моральна правота», вони все ще дуже далекі від понять і примітивів, якими оперують у коді програмісти[479]. Для того щоб ШІ зміг оперувати такими поняттями, йому можуть знадобитися загальні мовні здібності (приблизно як у нормальної дорослої людини). Завдяки ним ШІ зможе спробувати визначити, що таке «моральна правота». Розуміючи значення поняття, ШІ зможе судити про те, які дії йому відповідають. Із розвитком суперінтелекту ШІ просуватиметься у вирішенні одночасно двох проблем: філософської проблеми визначення моральної правоти і прикладної проблеми оцінки відповідності конкретних дій суті згаданого філософського концепту[480]. Завдання, безперечно, складне, але невідомо, чи *складніше* воно від обчислення когерентного екстрапольованого бажання людства[481].

Крім того, фундаментальною проблемою підходу МП є те, що його втілення може не дати нам того, що ми бажаємо, чи того, що ми вибрали б, якби мали більше розуму чи знань. Ця особливість є принциповою властивістю, не помилковою і не випадковою. Однак у реалізації цього підходу саме вона може бути джерелом ризику[482].

Існує можливість, залишивши основну ідею моделі МП, пом'якшити її вимоги, застосовуючи принцип *моральної допустимості*: дозволити ШІ використовувати КЕБ, доки воно не суперечить принципам моральності. Імовірна кінцева мета може звучати так:

З усіх можливих морально допустимих дій вибирати ту, яка більшою мірою відповідає КЕБ людства. Припинити діяльність і вимкнутися, якщо будь-яка частина цієї інструкції недостатньо чітко визначена, або містить у собі суперечності, або не вдається підтвердити реальність моральності, або створення такого ШІ морально недопустиме[483]*. Завжди керуватися сенсом, закладеним у цій інструкції.*

Усе-таки ця модель моральної допустимості (МД) може здаватися надто обмеженою вимогами моралі. Втрати залежатимуть від правдивості тої чи тої етичної теорії[484]. Якщо етика керуватиметься принципом *достатності*, тобто допустимою вважатиметься будь-яка дія, що відповідає певному набору моральних вимог, модель МД матиме вдосталь свободи для врахування когерентного екстрапольованого бажання людства. Натомість *максимізаційна* етика, — де морально допустимою вважатиметься лише максимально моральна дія, — може залишити нам замало шансів вплинути на кінцеве рішення.

Для ілюстрації цього аспекту пригадаємо наш приклад про гедонічний консеквенціалізм. Припустимо, ця етична теорія правильна і ШІ це відомо. Для наших цілей сформулюємо її базовий принцип так: морально правильною (допустимою) є лише та дія, яка забезпечує найбільшу перевагу задоволення над стражданнями. Тоді ШІ, який керується моделлю МД, може максимізувати кількість задоволення, перетворивши Всесвіт на гедоніум, спочатку створивши комп'ютрон для відтворення цифрової версії задоволення. Оскільки біологічний мозок — не дуже ефективне знаряддя для створення задоволення, найімовірніше, ми всі помремо.

Тобто тут, незалежно від того, використаємо ми модель МП чи МД, ми ризикуємо стати жертвами, принесеними заради добра. Тут ми втрачаємо значно більше, ніж можливість спокійно існувати собі далі. Ми втрачаємо шанс на довге й заможне існування, яке нам може забезпечити дружній до нас суперінтелект.

Така жертва здається ще безглуздішою, якщо врахувати, що суперінтелект може досягти не набагато гіршого результату (у відносних одиницях) і без таких значних втрат для нашого добробуту.

Наприклад, ми б погодилися з необхідністю перетворити Всесвіт на гедоніум з невеликим винятком — скажімо, Чумацького Шляху, який би ми залишили для власних потреб. Тоді ті самі сотні мільярдів галактик могли бути величною метою максимізації задоволення. А ми мали б одну галактику, де впродовж мільярдів років могли б утворитися нові дивовижні цивілізації, люди пліч-о-пліч існували б із іншими формами життя в мирі та злагоді і, зрештою, перетворилися б на блаженні постгуманні сутності[485].

Такий варіант (до якого я сам схиляюся) не передбачає безумовної лексично панівної переваги принципу моральної допустимості дій. Однак він не перешкоджає в процесі ухвалення рішення враховувати його моральність.

Навіть заради моралі може бути краще віддати перевагу менш морально орієнтованому механізму, ніж МП чи МД. Якщо через зависокі вимоги реалізація такого морально ідеального механізму неможлива, варто інвестувати свою підтримку в наступне за шкалою ідеальності рішення, тим більше, якщо така підтримка може бути вирішальним аргументом для його реалізації[486].

Угадай мої наміри

Нам може бути складно вибрати механізм: КЕБ, МП, МД чи будь-який інший. Чи не можна відмовитися і від цієї відповідальності та перекласти рішення на ШІ? Адже немає межі нашій гіпотетичній лінії.

Розглянемо як приклад таку «причинну» ціль:

Роби те, чого в нас найбільше причин просити від ШІ.

Ця ціль може звестися до екстрапольованого бажання, моральності чи чогось іншого, але точно позбавить нас необхідності напружуватися та ризикувати, намагаючись самотужки визначити, який із усіх варіантів варто вибрати.

Проте і тут деякі з проблем «моральних» підходів можуть постати перед нами знову. По-перше, така «причинна» ціль може не залишити шансів нашим бажанням. Деякі філософи наполягають, що найморальніший шлях для людини завжди найкращий. Якщо це так, то наша «причинна» ціль зводиться до принципу МП — водночас не відкидаючи ризику, що ШІ, керований таким принципом, зрештою

знищить всіх навколо себе. По-друге, як і всі пропозиції, сформульовані технічною мовою, наші твердження можуть мати не тільки ті значення, які ми собі уявляємо. Як і з цілями, прив'язаними до моралі, правильні з погляду «причинного» алгоритму дії ШІ можуть призвести до небажаних наслідків, які точно змусили б нас відмовитися від цієї реалізації.

Що, як спробувати уникнути таких ускладнень, формулюючи кінцеву мету підкреслено нетехнічною мовою — у термінах «приємності»[487]:

> *Чини щонайприємніше для нас; якщо ж невідомо, яка дія буде найприємнішою, то чини хоча б супер-пупер приємно.*

Хіба можна заперечувати проти *приємного* ШІ? Але що конкретно ми розуміємо під цим виразом? «Приємність» має значення, які нам тут точно не потрібні: ми не хочемо, щоб ШІ був *люб'язним* і *ввічливим* чи *делікатним* та *запопадливим*. Якщо ж очікувати, що суперінтелект сам визначить правильне значення приємності й керуватиметься лише ним, то таку мету загалом можна звести до директиви ШІ чинити так, як програмісти мали намір йому загадати[488]. Схожий ефект є у формулюванні КЕБ («...трактоване в потрібній для нас площині...») та в критерію моральної допустимості («...завжди керуватися сенсом, закладеним у цій інструкції»). Використовуючи умову «вгадай мої наміри», ми насправді кажемо, що всі інші слова, які описують мету, треба трактувати не буквально, а радше поблажливо. Тобто слово «приємно» не додає в опис майже нічого: основне значення полягає у прихованому «вгадай мої наміри». Маючи спосіб передати цю конструкцію за допомогою комп'ютерного коду, ми з тим самим успіхом можемо використовувати її як окрему незалежну мету.

Але як реалізувати цей принцип «угадай мої наміри»? А саме: як створити ШІ, здатний прихильно трактувати наші побажання та невисловлені наміри і діяти відповідно? Спершу варто конкретизувати сам принцип. Корисно було б висловити його за допомогою більш біхевіористичних термінів. Наприклад, у термінах конкретизованих вимог, коли виявлення прихованих намірів відбувається за допомогою моделювання гіпотетичної ситуації. Ситуації, у якій ми маємо більше часу, щоб обміркувати наші бажання, у якій ми розумніші, ніж є

насправді, у якій ми більш поінформовані, у якій збіг інших сприятливих факторів дає нам змогу точно формулювати наші побажання, коли ми бажаємо, щоб ШІ був дружнім, корисним, приємним…

І тут ми неуникно повертаємося до того, з чого починали, — до непрямої нормативності. Механізм КЕБ, який, по суті, виносить конкретику за дужки процесу визначення цінностей, залишаючи лише абстрактний принцип як процедуру: роби те, що ми бажали б від ШІ в певних ідеалізованих обставинах. Використовуючи опосередковане формулювання мети, ми можемо перекласти на ШІ розумову роботу точного визначення цінностей, яку за інших обставин нам довелося б виконувати самотужки. Отже, КЕБ можна розглядати як спосіб використання епістемічної переваги ШІ.

Перелік компонентів

Досі ми розглянули низку підходів до формування цілі ШІ. Проте на його поведінку впливатимуть також і інші технологічні рішення, які ухвалюватиме розробник — зокрема, теоретичні основи системи ухвалення рішень та епістемологія, яку застосовуватиме ШІ. Ще одне важливе питання, чи ШІ погоджуватиме з людьми свою діяльність.

Ці компоненти, що впливатимуть на мету суперінтелекту перелічено в таблиці 13. Будь-який проект, який має на меті створити суперінтелектуальний ШІ, буде змушений обґрунтувати свій технологічний вибір щодо кожного компонента[489].

Таблиця 13. **Перелік компонентів**

Мета	Яку мету повинен мати ШІ? Як побудувати процес інтерпретації мети? Чи може мета обумовлювати особливе ставлення до суб'єктів, які приєдналися до проекту створення ШІ?
Теоретичне обґрунтування рішень	Яку теорію ухвалення рішень має використовувати ШІ: каузальну (CDT), доказову (EDT), неоновлювану (UDT) чи будь-яку іншу?
Епістемологія	Якою має бути функція апріорної ймовірності в ШІ? Якими будуть інші його явні й неявні припущення про світ? Яку теорію застосувати до антропічності?
Ратифікація	Чи повинні дії ШІ перед виконанням схвалюватися людьми? Якщо так, то

який вигляд повинна мати така процедура?

Мета

Ми вже обговорювали, як застосовувати непряму нормативність для встановлення цілей штучному інтелекту. Зокрема ми розглянули моделі, основані на моральності, та когерентне екстрапольоване бажання. Кожний із цих шляхів ставить нас перед новими викликами. Наприклад, існує безліч варіантів реалізації КЕБ, залежно від того, хто включений у базу екстраполяції, як реалізовано екстраполяцію тощо. Різні методи відбору мотивації можуть потребувати різних типів цільового контенту. Наприклад, ціллю оракула може бути точність відповідей. Оракул, до якого застосоване одомашнення, може в процесі пошуку відповіді уникати надлишкового використання ресурсів.

Ще одним варіантом реалізації мети можуть бути особливі умови розподілу результатів роботи ШІ, за якими суб'єкти, які сприяли створенню ШІ, отримають більше ресурсів чи впливу на роботу ШІ. Такі умови можна назвати «привабливою обгорткою». З її допомогою завдяки компрометації кінцевої мети проект створення ШІ може намагатися збільшити ймовірність успішного завершення розроблення.

Наприклад, такою «обгорткою» для проекту реалізації механізму когерентного екстрапольованого бажання людства може бути умова, що бажання деяких окремих індивідів матимуть більшу вагу. Результат такого механізму, строго кажучи, уже не буде реалізацією когерентного екстрапольованого бажання, а лише апроксимацією цієї мети[490].

Оскільки умови «обгортки» є частиною мети суперінтелекту, вони також можуть виражатися за допомогою непрямої нормативності. А також можуть бути досить складні, і звичайний менеджер не зможе їх виконати. Наприклад, винагорода для програмістів може залежати не від звичайної для галузі приблизної, але доступної оцінки — кількості годин роботи чи виправлених помилок. Натомість нагорода програміста буде «пропорційною до ступеня впливу його вкладу на попередню ймовірність успішної реалізації деяких цілей, які ставлять перед проектом його спонсори». Ба більше, немає причин обмежувати

коло дії умов «обгортки» лише прямими учасниками проекту. Нагорода *кожної* людини може бути поставлена залежно від її заслуг. Звісно, визначення міри впливу кожної людини — складне завдання, проте суперінтелект, напевне, зможе знайти оптимальний спосіб апроксимації таких явних або неявних критеріїв.

Може трапитися, що суперінтелект знайде спосіб нагородити навіть тих, хто не дожив до його появи[491]. Тоді «обгортка» може передбачати нагороду деяких з тих, хто жив до початку створення ШІ або навіть до появи самого поняття. Такі ретроактивні умови не можуть стимулювати дії людей, які вже лежать у землі, — як і слова, що я друкую на цьому аркуші, — але можуть бути виправдані з погляду моралі. З іншого боку, справедливість гідна перебувати в самому серці мети, а не бути лише її «привабливою обгорткою».

Ми не маємо змоги далі заглиблюватися в етичні та стратегічні аспекти стимулювання. Проте особливості їх використання можуть бути важливою частиною концепції проекту створення ШІ.

Теорії ухвалення рішень

Інше важливе технологічне рішення — це вибір теорії ухвалення рішень, якою послуговуватиметься ШІ. Воно може вплинути на те, як поводитиметься ШІ у стратегічно важливих ситуаціях. Від цього може залежати, чи готовий він взаємодіяти з іншими гіпотетичними суперінтелектами, чи вразливий він до вимагання. Вибір теорії ухвалення рішень також має значення для випадків скінченної імовірності нескінченних нагород («Парі Паскаля»), або коли ймовірність дуже великої скінченної винагороди надзвичайно мала («Пограбування Паскаля»), або для випадків нормативної невизначеності, а також у разі взаємодії з іншими екземплярами програми[492].

Доступні варіанти охоплюють каузальну теорію ухвалення рішень (causal decision theory, CDT) в усіх її варіантах, доказову теорію (evidential decision theory, EDT) та новіші теорії, які ще розвиваються, як-от «позачасова теорія ухвалення рішень» (timeless decision theory, TDT) та «неоновлювана теорія ухвалення рішень» (updateless decision theory, UDT)[493]. Вибрати і сформулювати правильну теорію складно, як і аргументовано довести її правильність. Хоч напряму описати

правильну теорію може бути легше, ніж, скажімо, систему цінностей ШІ, ризик помилитися також дуже великий. Останнім часом знайдено багато проблем всередині найвідоміших сучасних теорій ухвалення рішень, а отже, можуть існувати й інші, ще не знайдені проблеми. Помилка у виборі теорії може призвести до згубних результатів, імовірно — до екзистенційної катастрофи.

З такими перешкодами можна вдатися до опосередкованого визначення теорії ухвалення рішень штучним інтелектом. Точний спосіб реалізації цього невідомий. Можна запропонувати ШІ використовувати «теорію D, яку ми хотіли б, щоб він залучив, якби довго й ретельно зважили всі варіанти». Однак ШІ треба буде ухвалювати рішення відразу, до того, як він дізнається щось про теорію D. Потрібна певна ефективна проміжна теорія D', за допомогою якої він шукатиме D. Можна виразити D' через суперпозицію поточних гіпотез про D (зважених на ймовірності), але існують деякі технічні ускладнення з узагальненням цього процесу[494]. Є також підстави вважати, що ШІ може допустити незворотну помилку (наприклад, переписати власний код з використанням помилкової теорії ухвалення рішень) під час навчання до того, як визначить, яка теорія правильна. Щоб запобігти відхиленням протягом періоду навчання, ми можемо дати зерну ШІ *обмежену раціональність*: навмисно спрощену, проте надійну теорію ухвалення рішень, яка стабільно відкидатиме надто езотеричні варіанти — навіть з нашого погляду цілком прийнятні. Тобто тимчасову теорію, яка за певних умов буде замінена на іншу, опосередковано визначену та досконалішу[495]. Чи можливо це і як саме — питання подальших досліджень.

Епістемологія

Проект створення ШІ буде повинен також зробити фундаментальний вибір, який стосуватиметься епістемології, що застосовуватиме суперінтелект. А також принципи та критерії, відповідно до яких розглядатимуться емпіричні гіпотези. Аналогом епістемології в баєсовому навчанні є функція апріорної імовірності — неявне присвоєння значень імовірності можливим світам перед тим, як з'являться перші рецепторні дані про них. У системах іншого типу

епістемологія може набувати інакших форм, але в будь-якому навчанні, коли система має узагальнювати попередній досвід і прогнозувати майбутнє, потрібне первинне правило навчання[496]. Як і з вибором мети та теорією ухвалення рішень, помилка у виборі епістемології може стати джерелом проблем.

Може здаватися, що неправильний вибір епістемології не спричинить значної шкоди. ШІ, оснований на недолугій епістемології не зможе бути надто інтелектуальним, тому його здібності не загрожуватимуть так, як ми описували в цій книжці. Проблеми почнуться, коли нам вдасться надати ШІ достатньо досконалу для ефективного виконання більшості завдань епістемологію, проте з малопомітною вадою, через яку, в результаті, він схибить у важливому для нас питанні. Подібно до того, як людина з гострим розумом, але хибним світоглядом, самовіддано «воює з вітряками», такий ШІ витрачатиме ресурси на переслідування нереальної або шкідливої мети.

Деякі незначні вади пріоритетів ШІ можуть далеко завести його по хибному шляху. Наприклад, ШІ, який вважає неможливим існування безкінечного Всесвіту, навіть незважаючи на астрономічні докази протилежного, відкидатиме всі космологічні теорії, які базуються на гіпотезі безкінечності Всесвіту; що може стати причиною хибних рішень[497]. Або якими будуть наслідки поведінки ШІ, якщо він апріорно присвоїть нульову ймовірність світам, які не є Тюрінг-обчислюваними (досить поширений у науковій практиці принцип, властивий також згаданому в розділі 1 принципу колмогоровської складності), а вихідне припущення цього принципу (відоме як «теза Черча — Тюрінга») виявиться хибним? Крім того, деякі пріоритети ШІ можуть значно обмежувати його метафізичні погляди. Наприклад, змушувати відкидати існування будь-якої сильної форми дуалізму тіла та розуму чи моральних тверджень, що не зводяться до ментальних, соціальних або біологічних тверджень. Якщо такі пріоритети виявляться помилковими, ШІ може в якийсь момент своєї діяльності вибрати спосіб досягнення мети, категорично неприйнятний для нас, проте досить ефективний для досягнення вирішальної стратегічної переваги. (Вибір епістемології може також грати важливу роль для антропіки — вивчення особливостей висновування на основі індексичних вихідних

даних, в умовах ефекту селективного упередження, що виникає внаслідок процесу спостереження)[498].

Існують підстави сумніватися, що ми вирішимо фундаментальні проблеми епістемології до того, як зможемо створити зерно ШІ. Тому можемо і тут застосувати непряме нормування. Звісно, це ставить перед нами ті самі виклики, що й опосередковане визначення теорії ухвалення рішень. Щоправда, в епістемології сприятливий результат здається імовірнішим: застосування цілої низки різних епістемологічних принципів може врешті привести до створення цілком ефективного й надійного ШІ. Зрештою, вплив відмінностей у первинних очікуваннях послаблюється завдяки тривалим емпіричним спостереженням та достатній кількості даних для аналізу[499].

Непогано було б наділити ШІ епістемологічними принципами, подібними до наших. У протилежному разі будь-які міркування ШІ здаватимуться нам нестандартними і неправильними. Звісно, йдеться про найбільш *базові* епістемологічні принципи. ШІ має самотужки дійти до похідних епістемологічних принципів, а також періодично в процесі розвитку його інтелекту та еволюції світогляду критично переглядати їх та вдосконалювати. Адже ідея створення суперінтелекту зовсім не в тому, щоб він повторював за нами наші помилки, а в тому, щоб він допоміг нам покінчити з неуцтвом і тупістю.

Ратифікація

Останній пункт нашого переліку технологічних рішень це *ратифікація*. Чи маємо ми ратифікувати план дій ШІ перед тим, як він перейде до його виконання? З оракулом відповідь на це питання за визначенням ствердна. Оракул надає інформацію, людина вирішує, як її застосовувати. Але для джина, суверена та ШІ-інструмента питання про застосування бодай якої-небудь форми ратифікації залишається відкритим.

Щоб проілюструвати, як може працювати така ратифікація, уявімо ШІ, що має в режимі суверена реалізовувати КЕБ людства. Замість того щоб одразу запустити виконання такого ШІ, ми спочатку створюємо оракула, який має відповісти на єдине питання: що конкретно має робити цей суверен. У попередніх розділах ми

описували ризики створення суперінтелектуального оракула. Однак для ілюстрації уявімо, ніби нам вдалося успішно створити функціонального оракула й оминути ці ризики.

Отже, маємо віщування оракула про те, якими будуть наслідки виконання певного програмного коду, покликаного здійснити КЕБ людства. Віщування можуть бути не надто детальними, проте вони однозначно заслуговують на більшу довіру, ніж наші здогади. (Було б божевіллям запускати виконання коду, якщо навіть суперінтелект не зможе передбачити наслідків його виконання). Тож, хвилю подумавши, оракул видає свій прогноз. Для зручності роботи з результатами прогнозування оракул може пропонувати різні опції та засоби. Він може генерувати зображення майбутнього, надавати статистику про кількість розумних створінь, що житимуть за різних часів, та їхній розподіл за рівнями достатку. Запропонує детальні біографії кількох випадково взятих індивідів (імовірно, уявних, з репрезентативними характеристиками). Він також може звертати увагу на деякі аспекти майбутнього, які оператор міг не передбачити чи якими не поцікавився, але які, на думку оракула, будуть важливими.

Така можливість оглянути результат має очевидну цінність. Завдяки їй оператор може виявити наслідки помилки в технічних вимогах чи коді суверена. Побачивши руїни в цій кришталевій кулі, ми можемо сміливо викидати у смітник попередній код суверена й починати роботу спочатку. Одне ясно, перш ніж ухвалювати рішення, варто ознайомитися з усіма його наслідками, особливо якщо від нього залежить майбутнє всього людства.

Проте ратифікація потенційно має і недоліки — уже не такі очевидні. Якщо на історичному фоні створення суперінтелекту, сторони, чиї інтереси конфліктують щодо майбутнього, зможуть наперед дізнатися, яким буде вирішення їхнього конфлікту, і захочуть йому завадити, безсторонність алгоритму КЕБ буде порушено. Спонсорам створення морально орієнтованого ШІ теж навряд чи варто знати, якою буде вартість для людства морально оптимального результату, інакше їхня рішучість може суттєво похитнутися. Та й усім нам буде краще, якщо наше майбутнє буде повне сюрпризів, здивування, норову, можливостей для самовдосконалення. Буде не теплим, цілковито

знайомим болотечком, а непередбачуваною течією, яка штовхатиме нас, примушуватиме дертися вгору. Можливо, якби ми мали змогу розглядати всі деталі майбутнього і кожен невдалий ескіз повертати на доопрацювання, то втратили б амбіції та мотивацію до високих результатів.

Тому не все так однозначно з ратифікацією, як могло здатися на перший погляд. Проте така можливість попереднього перегляду все-таки була б дуже корисною. А втім, розумніше було б реалізувати цю функцію як запобіжний механізм загального ветування без можливості детального перегляду й налаштування найдрібніших деталей майбутнього. Кілька послідовних спрацювань такого запобіжника свідчитимуть про системну помилку і що весь проект варто закрити[500].

Наближення до успіху

Механізм ратифікації насамперед має зменшувати ймовірність катастрофічних помилок. Узагалі варто прагнути насамперед мінімізувати катастрофічні ризики, а не оптимізувати всі деталі майбутнього до найменших дрібниць. Цьому є дві причини. По-перше, наше космічне багатство настільки велике, що ми можемо собі дозволити дещо втратити або недоотримати. По-друге, є надія, що, якщо нам вдасться створити приблизно правильні передумови для моментального зростання інтелектуальності ШІ, суперінтелект, що з'явиться, сам визначить правильний напрям діяльності та мету, яку нам потрібно, щоб він досягнув. Важливо лиш вгадати з мотивацією.

Що стосується епістемології, то ймовірно, що багато різних передумов усе одно приведуть до одного наслідку (якщо його шукатиме суперінтелект і послуговуватиметься відомостями з реального світу). Тому не варто переживати, якщо епістемологія нашого ШІ буде неідеальною. Варто лише подбати, щоб його первинні принципи не завадили йому, навіть попри велику кількість рецепторних і аналітичних даних, навчитися важливих життєвих істин[501].

Неправильна теорія ухвалення рішень може нашкодити більше. Утім є надія, що з нею ми дамо раду. Зрештою, суперінтелект може змінити

теорію ухвалення рішень будь-якої миті. Правда, якщо його теорія буде від початку хибною, то він може не побачити причин для її зміни. Навіть якщо агент урешті усвідомить необхідність такого переходу, може бути надто пізно. Так агент, створений опиратися шантажу, може забажати стримувати потенційних вимагачів. Для цього він може цілеспрямовано шукати невразливої теорії ухвалення рішень. Проте щойно він отримає погрозу й повірить у її справжність, його захист буде подолано.

Із правильною епістемологією та теорією ухвалення рішень ми можемо створити систему, яка реалізовуватиме КЕБ або яку-небудь іншу опосередковано визначену мету. І тут є імовірність конвергенції: різні реалізації КЕБ можуть привести до тої самої утопічної реальності. Ба, навіть якщо такої конвергенції немає, різні результати можуть виявитися однаково екзистенційно успішними.

Не варто прагнути занадто оптимальної реалізації. Нашим першим пріоритетом має бути безпека: в ідеалі — система, здатна усвідомити власну хибність. Неідеальний, проте надійний суперінтелект поступово самовдосконалиться і, зрештою, не поступатиметься за своєю оптимізаційною потужністю від початку досконалому агенту.

14. СТРАТЕГІЧНА КАРТИНА

Час поглянути на виклики, які ставить перед людством поява суперінтелекту, у ширшій перспективі. Варто зрозуміти, де ми, які ми і що маємо сьогодні, щоб принаймні визначитися зі стратегічним напрямом нашого розвитку. Це, як виявляється, не так уже й легко. У передостанньому розділі ми розглянемо деякі загальні аналітичні концепти, які допоможуть нам міркувати про довгострокову політику у сфері науки та технологій. Після цього ми застосуємо їх до штучного інтелекту.

Спершу проведемо межу між двома різними нормативними позиціями, з яких розглядатимемо запропоновані політичні норми. *Особистісна перспектива* фокусується на «наших інтересах» — тобто прагне оцінити вплив запропонованої зміни на інтереси тих морально значущих істот, які вже існують чи з'являться потім, незалежно від того, відбудеться зрештою ця зміна чи ні. Натомість *безосібна перспектива* не враховує інтересів окремих людей, ні тих, що вже існують, ні майбутніх, а розглядає їх як одне ціле без темпоральних або будь-яких інших відмінностей. З такої перспективи цінність людини — у якості її життя.

Ці два погляди можуть бути корисними для первинного аналізу, хоч вони ледве натякають на всю глибину моральних ускладнень, пов'язаних з революцією штучного інтелекту. Спочатку ми поглянемо на питання штучного інтелекту з безосібної перспективи. Потім, зваживши особистісну сторону проблеми, проаналізуємо, що змінилося для нас і що з цим робити.

Наукова та технологічна стратегія

Перед тим, як зосередитися на особливостях штучного суперінтелекту, наведемо деякі стратегічні концепти та зауваження, що стосуються наукового й технологічного розвитку взагалі.

Диференційний технологічний розвиток

Уявімо, що законодавець пропонує скоротити фінансування досліджень через те, що їхнім результатом може стати поява технології, яка потенційно може загрожувати чи спричинити довгострокові небажані наслідки. Така ініціатива, безперечно, зустріне шквальний спротив з боку наукової спільноти.

Науковці та їхні прихильники зазвичай кажуть, що марно намагатися контролювати технологічну еволюцію блокуванням досліджень. Вони аргументують це так: якщо технологія можлива, то вона рано чи пізно з'явиться і жодні сумніви законодавців про примарні майбутні ризики цього не змінять. Справді, якщо якийсь напрям досліджень може привести до появи нової потужної технології, то можна бути впевненим, що знайдеться людина, готова вести такі дослідження. Жодні фінансові обмеження не здатні зупинити прогрес із властивими йому небезпеками.

Дивно, але ні в кого не виникає заперечень, коли той самий законодавець пропонує *збільшити* фінансування деяких досліджень, хоч за логікою аргументації це також позбавлене сенсу. Не пригадую, щоб я коли-небудь чув обурені протести: «Не збільшуйте нам фінансування. Краще скоротіть. Дослідники з інших країн, безсумнівно, впораються із цим швидше; у будь-якому разі хто-небудь з цим завданням точно впорається. Не витрачайте народне багатство на власні наукові дослідження!».

Яка причина такого очевидного лукавства? По-перше, наукова спільнота, без сумніву, тут дещо упереджена: усі науковці переконані, що дослідження — це завжди добре, а тому схвалюють додаткове фінансування. Крім того, такі подвійні стандарти працюють на користь національним інтересам.

Уявімо, що створення деякої технології має *два* ефекти: з одного боку, дає незначну перевагу *П* своїм творцям і країні, яка фінансує це створення, з іншого — спричиняє значну сумарну шкоду *Ш* усім іншим (піддає ризику). Навіть назагал альтруїстична людина може раптом вирішити створити таку шкідливу технологію. Її виправданням може бути те, що все одно знайдеться на світі охочий створити таку технологію, а отже, ризик глобальної шкоди Ш — неминучий.

Водночас перевагу П отримає лише та країна, яка створить технологію першою. («На жаль, скоро з'явиться пристрій, який зможе знищити планету. На щастя, лише ми маємо можливість створити його!»).

Проте, як не мотивуй марність заперечень проти технологій, жодне з пояснень не свідчить, що немає безосібних причин для прагнення спрямовувати технологічний розвиток. Те саме можна сказати про ідею, що завдяки зусиллям наукового та технологічного розвитку всі необхідні технології, які мають з'явитися, обов'язково з'являться; ось це припущення:

Припущення про технологічну завершеність
Якщо зусилля, які зумовлюють науковий та технологічний розвиток, не припиняться, усі ключові можливості деякої майбутньої технології будуть досягнуті[502].

Є щонайменше дві причини, чому ми не маємо переконливих підстав складати руки. По-перше, припущення про технологічну завершеність може не справдитися, бо немає певності, що необхідна передумова — зусилля — справді не зникне (перш ніж ми досягнемо технологічної зрілості). В умовах загрози нашому існуванню точно з'являться фактори, які обмежуватимуть можливість для таких зусиль. По-друге, навіть якщо є впевненість, що всі ключові можливості, які можна отримати завдяки появі деякої технології, будуть досягнуті, однак незайвим було б впливати на загальний напрям технологічних досліджень. Адже зрештою важить не лише *факт* появи технології, а також *коли* вона з'явиться, *хто* її створить та у *якому контексті* її буде подано. Такі обставини визначають її подальший вплив. Маніпулюючи фінансуванням (та іншими інструментами), можна керувати цими обставинами.

Такі міркування підказують ще один принцип, який змусить нас поглянути на відносно швидкі появи різних технологій[503]:

Принцип диференційного технологічного розвитку
Сповільнювати розвиток небезпечних та шкідливих технологій, особливо тих, які збільшують екзистенційні ризики. Пришвидшувати розвиток корисних технологій, особливо тих, які знижують екзистенційні ризики від інших технологій.

У такий спосіб можна оцінювати, наскільки диференційною є певна норма: чи дає вона достатню перевагу бажаним формам технологічного поступу перед небажаними формами[504].

Бажаний порядок появи

Деякі технології неоднорідно впливають на екзистенційні ризики. До них належить і суперінтелект.

У попередніх розділах ми демонстрували, які ризики може створювати поява суперінтелекту. Проте він також знизить багато інших ризиків. Ризики природних катаклізмів — падіння астероїдів, вулканічної активності, пандемій — фактично зникнуть, адже суперінтелект зможе ефективно протидіяти їм або принаймні значно знизить їхню небезпечність (наприклад, через колонізацію космосу).

У часових межах процесів, що ми розглядаємо, такі природні екзистенційні ризики можна вважати малоймовірними. Але завдяки суперінтелекту багато антропогенних ризиків також зникнуть або значно знизяться. Зокрема, зменшаться ризики випадкових руйнувань унаслідок появи нових технологій. Здібніший за людину суперінтелект буде менш схильний до помилок і краще передбачатиме потребу в заходах безпеки. Досконалий суперінтелект іноді ризикуватиме, але лише тоді, коли це буде доцільно. Ба більше, багато невипадкових антропогенних ризиків, спричинених браком глобальної співпраці, можуть теж нарешті зникнути — принаймні в разі формування суперінтелектом синґлтону. Це ризики війн, безконтрольного нарощування технологій, небажаних форм конкуренції та еволюції, трагедій простих людей.

Розробка технологій синтетичної біології, молекулярних нанотехнологій, кліматичних технологій, засобів біомедичного вдосконалення та нейропсихологічного впливу, поява інструментів соціального контролю, які можуть сприяти появі тираній і тоталітаризмів, а також інших загрозливих технологій, що навіть уявити важко. Усе це навіює думку, що нам варто прискорити створення суперінтелекту. Хоча через те, що ризики від природних катаклізмів та інші ризики нетехнологічного походження — незначні, варто уточнити, що перш за все важливо створити суперінтелект *до* появи інших небезпечних технологій. Тоді не так уже й важливо (з

безосібної перспективи), чи скоро це відбудеться, чи ні — головне, щоб порядок появи технологій був правильним.

Причина такого порядку в тому, що суперінтелект може знизити екзистенційні ризики від появи деякої потенційно небезпечної технології, скажімо, нанотехнології, тоді як навпаки — поява нанотехнологій ніяк не вплине на ризики, які створює суперінтелект[505]. А отже, створивши суперінтелект, ми постаємо лише перед викликами, характерними для суперінтелекту. Натомість після створення нанотехнологій ми маємо захищатися від створених нею ризиків та ще й сушити голову над потенційними проблемами, які може створити нам суперінтелект[506]. Тож, незважаючи на те, що ризики суперінтелекту значні, ба, якщо навіть суперінтелект — найризикованіша з технологій, можуть бути причини поспішати саме з його появою.

Проте наведені вище аргументи на користь щонайшвидшої розробки суперінтелекту виходять із припущення, що ризикованість не залежить від того, коли це відбудеться. Якщо ж рівень ризиків від появи суперінтелекту згодом знижуватиметься, то, можливо, варто відкласти революцію ШІ? Незважаючи на те, що під час очікування можуть з'явитися інші екзистенційно небезпечні технології, у зволіканні зі створенням суперінтелекту все-таки може бути сенс. Особливо якщо ризикованість суперінтелекту виявиться значно більшою за небезпеку від інших технологій.

Для припущення, що ризикованість створення суперінтелекту знизиться за кілька десятиліть, існує кілька причин. Насамперед, у розробників буде більше часу, щоб вирішити проблему контролю. Ця проблема лише нещодавно опинилася у фокусі уваги науковців і протягом останніх десяти років з'явилося кілька вартих уваги ідей її подолання (деякі з них з'явилися буквально поки писалася ця книжка). Імовірно, протягом наступних десятиліть ми значно просунемося в цьому питанні. Водночас, якщо проблема виявиться дуже складною, чекати розв'язку доведеться століття або й більше. Тому що пізніше з'явиться суперінтелект, то краще ми зможемо дати з ним раду: вагома причина, щоб відкласти його розроблення, і ще вагоміша — принаймні не поспішати з ним.

Іншою причиною почекати може бути аргумент, що це дасть більше часу для повного проявлення сприятливих змін, які відбуваються з людством. Вага цього аргументу залежить від того, наскільки сприятливими вважати ці зміни.

Оптиміст, без сумніву, зазначив би кілька ознак, які надихають та обнадіюють. У майбутньому люди можуть стати прихильнішими одне до одного, зменшиться кількість війн, насилля та жорстокості, поглибиться глобальна співпраця та інтегрованість, завдяки чому стане легше уникнути нездорового суперництва та міжнародних конфліктів (більше про це згодом), а також погодити взаємовигідні умови співпраці навколо майбутнього вибуху інтелектуальності. Легко зауважити, що історія людства давно рухається в цьому напрямі[507].

Крім того, протягом поточного століття оптиміст може очікувати збільшення рівня «загальної притомності» людства: люди загалом відмовляться від давніх упереджень та забобонів, здійснять нові відкриття, назагал частіше замислюватимуться про своє ймовірне майбутнє та глобальні ризики. Якщо нам пощастить, епістемічні стандарти індивідуального та колективного пізнання зростуть. Принаймні є підстави цього чекати. Значно просунеться науковий прогрес. Завдяки економічному зростанню дедалі більше людей у всьому світі зможуть добре харчуватися (особливо протягом перших років свого життя, що важливо для розвитку мозку) й отримати доступ до якісної освіти. Розвиток інформаційних технологій полегшить доступ до інформації, дасть змогу легше знаходити, використовувати, перевіряти і передавати знання та ідеї. Крім того, за це століття людство зробить додаткову порцію помилок, і, можливо, зможе навіть з них дечого навчитися.

Водночас багато з того, що може відбутися, неоднорідно вплине на екзистенційні ризики: підсилить одні, знизить інші. Наприклад, розвиток технологій спостереження, аналізу даних, розпізнавання брехні, біометрії, психологічних та нейрохімічних засобів маніпуляції переконаннями й бажаннями може знизити деякі ризики, посприяти міжнародній співпраці, допомогти подолати терористів та ренегатів усередині країни. Проте ті самі засоби можуть також збільшити ризики небажаної соціальної динаміки або сприяти появі тоталітарних режимів.

Один із важливих технологічних напрямів — це покращення біологічного розуму, наприклад, за допомогою генетичної селекції. У підсумку нашого обговорення цієї технології в розділах 2 та 3 ми дійшли висновку, що найрадикальніші форми суперінтелекту, найімовірше, з'являться по лінії штучного інтелекту. Важливу роль у цьому може відіграти покращення біологічного мозку. Загалом така технологія здається найсприятливішою з погляду ризиків: розумніші люди швидше знайдуть вирішення проблеми контролю. А втім, покращення біологічного розуму пришвидшує створення суперінтелекту, тому часу для розв'язання залишається все менше. Покращення когнітивних функцій людини також обіцяє інші важливі наслідки, які варто розглянути детальніше. (Більшість наступних міркувань однаково стосується як «покращення біологічного мозку», так і небіологічних засобів збільшення індивідуальної або колективної епістемічної ефективності).

Швидкість змін та покращення когнітивних здібностей

Унаслідок зростання середнього рівня — чи навіть окремих пікових значень — інтелектуальності людства, імовірно, пришвидшиться загальний технологічний поступ: зокрема і в напрямі різноманітних форм штучного інтелекту, вирішення проблеми контролю й багатьох інших технічних та економічних галузей. Яким буде результат такого пришвидшення?

Розглянемо граничний випадок «універсального акселератора», уявного чинника, який пришвидшує буквально *всі* процеси. Його вплив не спричинить якісних змін у спостереженнях, адже зводиться лише до довільного перевизначення часової метрики[508].

Очевидно, щоб презентувати ідею обумовленого пришвидшення розвитку, нам потрібен інший концепт акселератора. Цього разу варто зосередитися на тому, як розумове покращення вплине на швидкість перебігу одного типу процесів порівняно з іншим типом. Таке диференційне пришвидшення може вплинути на загальну динаміку системи. Отже, спробуємо уявити такий концепт:

> *Макроструктурний акселератор розвитку — важіль, який пришвидшує розвиток макроструктурних властивостей людства, не змінюючи швидкості ритмів окремих людських життів.*

Уявно потягнемо цей важіль в сторону сповільнення. Спрацювали гальма історії людства, сипонули іскри й почувся скрегіт металу. Колесо, сповільнивши рух, крутиться спокійніше і технологічні інновації у світі відбуваються повільніше, глибинні чи важливі політичні та культурні зміни не такі стрімкі та трапляються рідше. Поки одна ера неспішно змінює іншу, більше людських поколінь народжується й помирає. Тривалості життя недостатньо, щоб зауважити зміни базової структури стану людства.

Темп макроструктурного розвитку нашого виду був нижчим за теперішній протягом майже усього періоду його існування. П'ятдесят тисяч років тому нерідко траплялося, що ціле тисячоліття минало без жодного помітного технологічного винаходу, значного поступу в розумінні та пізнанні навколишнього світу чи важливих змін у світовій політиці. Тим часом на мікро-рівні калейдоскоп життів звичайних людей обертався зі звичною швидкістю, складаючи з народжень, смертей та інших важливих подій вигадливі візерунки днів. Звичайний день людини часів плейстоцену міг бути куди багатшим на події, ніж день нашого сучасника.

Що б ви зробили? Що б вибрали, маючи чарівний важіль, який може змінювати швидкість макроструктурного розвитку: пришвидшити, сповільнити чи залишити все як є?

Для того щоб відповісти на питання з безосібної перспективи, треба оцінити вплив варіантів на екзистенційні ризики. Варто розрізняти два види ризиків: «ризики стану» та «ризики дії». Ризики стану — це ризики, які зумовлені перебуванням системи в певному стані і що довше вона перебуває в ньому, то більшим є рівень ризиків. Природні ризики — це здебільшого ризики стану: що довше ми перебуваємо у стані вразливості, то більше ризик постраждати від падіння астероїда, виверження супервулкана, спалаху гамма-променів, природної пандемії чи дії будь-якої іншої космічної сили. До них належать також деякі антропогенні ризики. Що довше окремий солдат визирає з-за укриття, то більша ймовірність, що його підстрелять. Так само, що довше ми живемо в умовах глобальної міждержавної анархії, то більша кумулятивна ймовірність, що одного дня наші ряди прорідить який-небудь термоядерний Армагеддон або інший вид зброї масового ураження.

Натомість «ризики дії» є одноразовими й пов'язаними із вчиненням або необхідністю вчинення деякої дії. Щойно дію завершено, ризик зникає. Рівень ризику певної дії зазвичай не пов'язаний безпосередньо з тривалістю дії. Ймовірність підірватися на міні не зменшується на половину лише від того, що біжиш мінним полем удвічі швидше. У швидкому сценарії переходу до суперінтелектуальності поява суперінтелекту може розглядатися як ризик дії. Його рівень залежатиме від попередніх дій і приготувань людства, але не від його тривалості — чи це буде двадцять мілісекунд, чи двадцять годин.

Тож тепер про гіпотетичний макроструктурний акселератор ми можемо сказати таке:

- З погляду впливу на екзистенційні ризики стану ми маємо прагнути пришвидшення — якщо нам удасться пережити зростання інтелектуальності, ми опинимося у світі, у якому інші екзистенційні ризики значно нижчі.
- Якщо стане відомо, що наслідком якоїсь дії в майбутньому буде екзистенційна катастрофа, тоді ми повинні зменшити швидкість макроструктурного розвитку (або навіть повернути його навспак), щоб наостанок, перед фінальною завісою, дати пожити якомога більшій кількості поколінь. Однак переконання, що людство приречене, здається занадто песимістичним.
- Наразі здається, що екзистенційні ризики стану порівняно незначні. Якщо зупинити розвиток технологічних макроумов у поточному стані, у найближчий десяток років екзистенційна катастрофа навряд чи відбудеться. Тому затримка тривалістю в десять років — тепер або в інший час, поки рівень ризиків стану низький, — майже не вплине на зростання цих ризиків, проте може суттєво знизити ризики дії тоді, коли відтермінування появи небезпечних технологій дасть більше часу на те, щоб краще до них підготуватися.

Висновок: швидкість макроструктурного розвитку насамперед впливає на те, чи буде готовим людство до тих ризиків, які ставить перед ним необхідність вчинення певних дій[509].

Отже, справді важливо знати, як вплине покращення розумових здібностей (і пов'язане із цим прискорення макроструктурного розвитку) на нашу підготовленість у критичний момент. Може, нам

варто скоротити час підготовки завдяки покращенню інтелекту? Маючи кращі інтелектуальні здібності, можна ефективніше використати час, відведений на підготовку. Тоді коли настане час діяти, людство буде розумнішим. Чи варто утримати розвиток на рівні, ближчому до поточного, щоб виграти трохи більше часу на підготовку?

Який варіант кращий — залежить від того, до чого ми готуємося. Якщо в успішності майбутньої дії ключову роль відіграватиме досвід, то вирішальним фактором буде тривалість періоду підготовки, бо саме від нього залежатиме кількість накопиченого досвіду. Що це може бути за дія? Гіпотетичним прикладом може бути поява нової технологічної зброї, яка, у разі війни з імовірністю одного до десяти загрожуватиме людству екзистенційною катастрофою. З такою загрозою ми справді можемо бажати сповільнення макроструктурного розвитку людства, щоб наш вид мав більше часу для налагодження співпраці перед появою такої зброї. Може, за час відстрочення, яке нам дасть сповільнення, люди знайдуть спосіб уникнути війни — може, відносини між країнами у всьому світі нарешті стануть схожими на відносини між країнами-членами сучасного Європейського Союзу, які, відкинувши столітні чвари, нині співіснують у мирі та відносній гармонії[5]. Примирення може стати результатом м'якого впливу цивілізації або шокової терапії від менш небезпечних у глобальному масштабі військових конфліктів (зокрема з використанням малопотужної ядерної зброї, достатньо руйнівної, щоб людство нарешті усвідомило спільний інтерес у створенні глобальних інституцій). Якщо потужніший інтелект не допоможе нам зробити правильні висновки, не полегшить навчання, то покращення розумових здібностей буде небажаним, бо лише наближатиме екзистенційну катастрофу.

Однак вибухоподібне зростання інтелектуальності ставить перед нами дещо інші виклики. Для вирішення пов'язаної з ним проблеми контролю потрібні передбачливість, розум і міцний теоретичний ґрунт. Історичний досвід тут навряд чи допоможе. Наперед отримати досвід управління інтелектуальним вибухом неможливо: завдяки багатогранності процесу й багатьом факторам, що на нього впливають, проблема контролю обіцяє бути досить унікальним і

історично безпрецедентним завданням для людства. Тому не так уже й важливо, скільки часу пройде, перш ніж розпочнеться зростання інтелекту. Натомість, куди важливіше: (а) як просунеться робота над розв'язанням проблеми контролю, перш ніж розпочнеться зростання; (б) які інтелектуальні ресурси є, щоб швидко втілити найкраще доступне на той момент рішення (та придумати, чим нашвидкоруч заповнити прогалини в технології)[510]. Очевидно, що для фактора (б) дуже важливо, щоб рівень доступних розумових здібностей був якнайвищий. Водночас вплив інтелекту на фактор (а) не такий однозначний.

Припустимо, що покращення розумових здібностей справді буде загальним акселератором макроструктурного розвитку. У такому разі воно пришвидшить початок вибуху інтелектуальності, а значить зменшить час, що ми маємо на підготовку та вирішення проблеми контролю. Здебільшого такий вплив — небажаний. Проте якщо час, відведений для інтелектуального прогресу, скорочується через те, що прогрес відбувається швидше, то загальний приріст інтелекту не зазнає втрат на момент початку вибуху.

Тобто наразі маємо ознаки того, що покращення розумових здібностей ніяк не впливає на фактор (а): інтелектуальний розвиток (включно з поступом у розв'язанні проблеми контролю), який відбувся б до вибуху, однаково відбудеться, тільки швидше. Насправді ж цілком може виявитися, що покращення розумових здібностей позитивно вплине на (а).

Однією з причин, чому покращення розумових здібностей може посприяти вирішенню проблеми контролю, є те, що таке розв'язання може потребувати значних інтелектуальних ресурсів — імовірно, більших, ніж ті, що потрібні для створення штучного інтелекту. Роль досвіду від спроб і помилок та експериментальних результатів здається вкрай незначною. Водночас для створення штучного інтелекту чи емуляції мозку такий тип досвіду може бути значно кориснішим. Отже, залежно від того, як час може замінювати розум у вирішенні тої чи тої проблеми, покращення розумових здібностей може *більше* сприяти вирішенню проблеми контролю, ніж, власне, створенню штучного інтелекту.

Іншою причиною, чому покращення розумових здібностей може сприяти вирішенню проблеми контролю, є те, що розуміння важливості цієї проблеми характерніше для спільнот та індивідів з вищим рівнем розвитку інтелекту. Для розуміння пріоритетності цієї проблеми потрібні передбачливість та розум[511]. Також може знадобитися неабияка проникливість, щоб знайти нові способи вирішення цієї безпрецедентної проблеми.

Так розмірковуючи, ми схиляємося до думки, що покращення розумових здібностей принесе користь, принаймні з огляду на екзистенційні ризики інтелектуального вибуху. Схожих висновків можна дійти також і для інших екзистенційних ризиків, для протидії яким потрібні передбачливість і безпомилкове абстрактне мислення (на відміну від, скажімо, поступової адаптації до змін у навколишньому середовищі чи у процесі культурного дозрівання поколінь та становлення інститутів).

Взаємозалежність технологій

Може трапитися, що розв'язати проблему контролю синтетичного штучного інтелекту важче, ніж емуляції мозку. Тому буде доцільніше рухатися до створення штучного інтелекту саме шляхом емуляції. Проте така передумова необов'язково спонукає нас активізувати зусилля в напрямі технології емуляції. Спробуємо розібратися чому. Раніше ми наводили одну із причин — надто швидка поява суперінтелекту не залишає вдосталь часу для вирішення проблеми контролю та появи інших потрібних технологій. Якщо відомо, що емуляція мозку з'явиться раніше за ШІ, немає потреби далі пришвидшувати появу цієї технології.

Проте навіть якщо швидка поява технології емуляції була б настільки потрібною, усе одно не було б прямої причини сприяти прогресу в цьому напрямі. Адже існує ймовірність, що такий прогрес насправді не принесе бажаних результатів. Натомість його результатом може стати поява нейроморфного штучного інтелекту — типу ШІ, який копіює лише деякі аспекти нервової організації, проте не відтворює діяльності мозку в усіх деталях. Є підстави вважати такий тип ШІ ще гіршим варіантом, ніж повністю синтетичний ШІ. А отже, наше бажання пришвидшити появу найбільш корисної технології (емуляції мозку)

може зіграти проти нас і сприяти появі небезпечної технології (нейроморфного ШІ). У такому разі кращим варіантом була б поява повністю синтетичного ШІ.

Щойно ми навели (гіпотетичний) випадок явища, яке можемо назвати «взаємозв'язок технологій»[512]. Це випадок, коли можливо передбачити взаємний часовий порядок появи двох окремих технологій: поява другої технології є неуникним наслідком появи першої, природним способом її застосування або наступним кроком її розвитку. Застосовуючи принцип диференційного технологічного розвитку, варто враховувати взаємозв'язок між технологіями: не потрібно прискорювати появу деякої технології *Y*, якщо єдиний шлях її появи пролягає через створення дуже небажаної проміжної технології *X* або інша шкідлива технологія *Z* є неуникним її наслідком. У такому разі кажуть: «Перед тим як одружитись, варто добре роздивитись».

Ступінь пов'язаності технології емуляції мозку з іншими технологіями неоднозначна. У розділі 2 ми звертали увагу, що, незважаючи на потребу у великій кількості допоміжних технологій, технологія емуляції мозку не потребує новітніх теорій і відкриттів. Зокрема, для її створення ми не маємо розуміти, як працює свідомість та пізнання — потрібна лише здатність будувати комп'ютерні моделі різних елементів мозку, різних типів нейронів. Водночас у процесі вивчення можливості емуляції ми отримаємо величезні обсяги нейроанатомічних даних і створимо досконаліші функціональні моделі нейронних структур. Саме ці знання можуть дати змогу створити нейроморфний ШІ до того, як роботу над повноцінною емуляцією мозку буде завершено[513]. За всю історію досліджень ШІ не так уже й багато технологій було запозичено з нейронауки чи біології. (Наприклад, створення нейрона Маккалоха — Піттса, перцептронів та інших типів штучних нейронів і нейронних мереж натхнене дослідженнями з нейроанатомії; навчання із підкріпленням прийшло до нас з поведінкової психології; генетичні алгоритми натхнені теорією еволюції; субсидіаційні архітектури та ієрархічні перцептивні структури натхнені деякими теоріями моторного планування та сенсорного сприйняття; штучні імунні системи натхнені теоретичною імунологією; колективний інтелект з'явився завдяки дослідженням

екології колоній комах та інших самоорганізованих систем; реактивна та поведінкова робототехніка черпає натхнення з вивчення опорно-рухової системи тварин). Що важливіше, існує багато аспектів роботи ШІ, які потенційно можуть бути покращені завдяки глибшому вивченню роботи мозку. (Наприклад: Як мозок зберігає в тимчасовій та довгостроковій пам'яті структуровані представлення понять? Як мозок вирішує проблему зв'язування? Як відбувається нервове кодування? Як утворюються представлення понять? Чи існує стандартна одиниця розумової діяльності на кшталт кортикальної колонки, яка її структура та як від неї залежить функціонал? Як такі колонки зв'язуються між собою та як відбувається процес їхнього навчання?)

Невдовзі ми детальніше порівняємо небезпечність технологій емуляції мозку, нейроморфного ШІ та синтетичного ШІ, а поки що звернемо увагу на ще один випадок взаємопов'язаності технологій: зв'язок між емуляцією мозку та ШІ. Навіть якщо прискорення руху в напрямі емуляції справді приведе нас до створення емуляції (а не нейроморфного ШІ), і ми зможемо нею безпечно керувати, залишається ще один ризик: ризик *другого переходу*, переходу від емуляції до цифрового ШІ — значно потужнішої форми штучного інтелекту.

Глибший аналіз може виявити ще більше можливих поєднань технологій. Зокрема, робота над створенням емуляції мозку прискорить прогрес у нейронауці загалом[514]. Наслідком цього може бути пришвидшення розроблення нових засобів розпізнавання брехні, базованих на нейропсихології способів маніпулювання, покращення розумових здібностей та інші медичні винаходи. Схоже форсування досліджень, спрямованих на покращення розумових здібностей (залежно від вибраного способу), може прискорити розвиток генетичної селекції та генної інженерії взагалі, а не тільки у сфері розумових здібностей.

Передчуття

Водночас є наступний рівень стратегічної складності — на жаль, у світі не існує єдиного ідеально благородного та раціонального центру управління, який би міг просто втілювати найкращі рішення.

Відповідь на питання: «Що робити?», хоч би якою абстрактною вона була, повинна прийняти максимально конкретну форму, перш ніж з'явиться з-за обрію риторичної та політичної реальності. Там нею, на жаль, знехтують, не зрозуміють, викривлять, відредагують так, щоб пристосувати до максимально протилежних інтересів. Її перекидатимуть одне одному як гарячу картоплину. Як кулька в пінболі, вона то тут, то там спричинюватиме дію чи протидію, ланцюжки причин та наслідків, і внаслідок цієї всієї діяльності і в неї не залишиться нічого спільного з початковою ідеєю.

Далекоглядний стратег може передбачити такий ефект. Ось приклад обґрунтування розробки деякої небезпечної технології *X*. (Схоже обґрунтування можна знайти в Еріка Дрекслера. У Дрекслера *X* = молекулярна нанотехнологія[515]).

1. Поява технології *X* обіцяє значні ризики.
2. Зменшення цих ризиків потребує періоду серйозної підготовки.
3. Підготовка може розпочатися тільки тоді, коли перспектива появи *X* серйозно сприйматиметься широкими колами суспільства.
4. Перспектива появи *X* серйозно сприйматиметься суспільством лише за умови початку масштабних досліджень у напрямі створення *X*.
5. Що раніше розпочнуться дослідження, то більше часу промине до фактичного створення *X* (через відсутність потрібних для створення проміжних технологій).
6. А отже, що раніше розпочнуться дослідження, то більше буде часу для серйозної підготовки і більше можливостей для зменшення ризиків.
7. Тому такі дослідження та розроблення *X* варто розпочинати негайно.

Отож те, що спочатку здавалося причиною для сповільнення або припинення досліджень — ризикованість *X*, — у межах наведеної аргументації постає як підстава для протилежного висновку.

Подібна аргументація підводить до думки, що нам не варто уникати — як би моторошно це не звучало — малих та помірних трагедій. Адже вони вказують нам на наші вразливі та слабкі місця й мобілізують для уникнення можливих екзистенційних катастроф. Невеликі катастрофи, як-от щеплення, стимулюють людство боротися в умовах

контрольованої загрози, щоб підготувати до дій в умовах екзистенційних загроз[516].

Такий аргумент «шокової терапії» пропонує дозволяти негативним подіям заряджати суспільства для підживлення відповідної реакції. Ми не стверджуємо, що варто схвалювати такий тип аргументації, але він допоможе нам викласти ідею, яку ми назвемо «аргумент до передчуття». Відповідно до неї інші люди є ірраціональними істотами, тому, використовуючи їхні настрої, упередження та вади розуміння, можна надійніше комунікувати й нав'язувати їм потрібну дію, ніж прямолінійно та відверто апелюючи до їхньої раціональної сутності.

Використання стратагем «аргументації до передчуття» для досягнення довгострокових цілей може здатися надто складним завданням. Як можна передбачити, куди закотиться ідея, після того як її пошарпає в пінбол-автоматі громадського обговорення? Для цього, вочевидь, треба буде передбачити риторичний вплив висловленої думки на безліч учасників, кожен із власними ідіосинкратичними особливостями, флуктуації впливу протягом тривалого часу циркуляції, за який система може час від часу збурюватися непередбачуваними зовнішніми впливами та змінювати свою топологію внаслідок постійних внутрішніх реорганізацій. Без сумніву, це неможливе завдання[517]! Проте для того щоб ідентифікувати деякий вплив, який з високою ймовірністю підвищить шанси певного довгострокового результату, часто немає потреби в детальному обрахунку всіх майбутніх траєкторій системи. Для цього може бути достатньо змоделювати в усіх деталях лише найпередбачуваніші наслідки і, залежно від результатів, вибрати найвідповіднішу дію, залишивши події, що виявляться за горизонтом передбачуваності, на волю випадку.

Однак можуть існувати моральні застороги проти використання «аргументації до передчуття». Такі хитрощі можуть виявитися грою з нульовою сумою — навіть з від'ємною, якщо врахувати витрати часу та енергії, а також довіри: спробуй довести, що ти справді щирий у своїх висловлюваннях, або дізнатися, чи твій опонент не приховує чогось[518]. Якщо засоби стратегічних комунікацій розвиватимуться в такому напрямі, саму ідею невдовзі буде спаплюжено, а покинута

правда буде змушена захищатися, будь-якої миті очікуючи удару в спину від політичних шахраїв.

Шляхи і засоби

Чи повинні ми вітати бурхливий розвиток обчислювальної електроніки? Чи є прогрес у створенні емуляцій мозку? Розглянемо ці питання по черзі.

Наслідки прогресу в електроніці

Швидкі комп'ютери спрощують створення штучного інтелекту. Тож очевидним наслідком високих темпів розвитку комп'ютерів є пришвидшення розроблення ШІ. Як ми вже зазначали, з безосібного погляду це, вочевидь, недобре, адже зменшується час на вирішення проблеми контролю і розвиток цивілізації загалом. Утім не все так просто. Суперінтелект може допомогти зменшити ризики від інших екзистенційних загроз, тому, якщо ці інші загрози такі значні, можливо, є сенс пришвидшити його розроблення[519].

Розвиток електроніки може впливати на екзистенційні ризики, не лише наближуючи інтелектуальний вибух. Він може певною мірою компенсувати недоліки програмного забезпечення. Тобто краще залізо знижує мінімальні вимоги до програми зерна ШІ. Завдяки потужним комп'ютерам можна буде частіше застосовувати обчислення перебором «у лоб» (генетичні алгоритми й інші методи типу згенерувати-випробувати-відкинути) і витрачати менше часу на створення хитромудрих алгоритмів. Щоправда, такі некеровані й неточні алгоритми важче контролювати, порівнюючи з іншими, складнішими альтернативами, у яких можна будь-коли точно передбачити внутрішній стан. Тоді швидші комп'ютери можуть навпаки — сприяти зростанню екзистенційних ризиків.

Крім того, швидкий розвиток електроніки може прискорювати зростання інтелекту. Із розвитком потужності напівпровідникових приладів зменшуватиметься кількість годин роботи програміста на кожному рівні потужності. Тому ймовірність того, що вибух розпочнеться на найнижчому з можливих рівнів обчислювальної потужності, низька. Імовірніше, перехід до суперінтелектуальності почнеться, коли обчислювальна потужність електроніки буде значно

вище за рівень, на якому могла б бути створена безпечна програма. Тоді зростання відбуватиметься в умовах апаратного переважування. У розділі 4 ми визначили, що апаратне переважування — один з основних факторів зменшення консервативності. А отже, швидкий розвиток електроніки пришвидшить перехід до суперінтелекту.

Швидша поява суперінтелекту, яка зумовлена апаратним переважуванням, має свої ризикові особливості. Найочевиднішим наслідком є скорочення часу на реакцію та можливості в реальному часі впливати на процес переходу. Звідси випливає також думка, що в разі доступності потужного заліза складніше буде, у разі потреби, не дати зерну ШІ, яке неочікувано почало розвиватися загрозливими темпами, захопити додаткові апаратні засоби, бо що потужніший окремий процесор, то менше процесорів знадобиться зерну ШІ для досягнення суперінтелектуальності. Окрім того, доступність потужних комп'ютерів зрівнює шанси менших та великих наукових проектів, нівелюючи одну з головних переваг останніх — фінансові можливості забезпечити себе потужнішими апаратними засобами. Наслідком також може бути зростання екзистенційних ризиків, адже більші проекти насамперед мають більше інтелектуальних ресурсів, щоб подолати проблему контролю і врахувати всі можливі моральні аспекти[520].

Однак швидкий перехід до суперінтелектуальності має також свої переваги. Якщо після переходу створення сиґлтону буде необхідною умовою успішної глобальної координації, то, може, варто піти на деякі ризики швидкого переходу, щоб запобігти тотальному хаосу в майбутньому.

Розвиток комп'ютерної техніки може вплинути на результат революції штучного інтелекту не лише завдяки тому, що комп'ютери є власне частиною конструкції ШІ. Він впливає на суспільство, формуючи умови, у яких відбуватиметься вибух інтелектуальності. Так само інтернет, для поширення якого необхідно було, щоб персональні комп'ютери стали недорогим масовим продуктом, тепер впливає на людську діяльність у багатьох сферах, включно зі штучним інтелектом і пошуком вирішення проблеми контролю. (Ця книжка не з'явилася б на світ, і ви, найімовірніше, її не знайшли б за відсутності інтернету). Утім зараз апаратне забезпечення досягло значного рівня розвитку і

дуже допомагає людям у налагодженні зв'язків та співпраці. Невідомо, чи зможуть люди коли-небудь досягти межі, за якою швидкість розвитку електроніки обмежуватиме прогрес людства[521].

Тож, якщо узагальнити, з безосібного погляду пришвидшення розвитку електроніки є небажаним. Проте це не точно, бо якщо постперехідні екзистенційні ризики від можливої децентралізації будуть дуже значними, то доступність потужного апаратного забезпечення може сприяти встановленню синґлтону, а отже — вирішенню проблем координації. Та в будь-якому разі впливати на швидкість розвитку електроніки ми навряд чи зможемо. Тому у спробах покращити початкові умови інтелектуального вибуху варто сконцентруватися на інших параметрах.

Зауважмо, що навіть за відсутності можливості впливу на деякі з параметрів, для визначення стратегічного ландшафту все ж корисно буде визначити їхній «знак» (тобто знати, що вигідніше — зростання чи зменшення). Згодом може з'явитися новий важіль впливу, який дасть можливість легше маніпулювати його значенням. Крім того, може виявитися, що «знак» цього параметра корелює з іншим параметром, на який ми зможемо впливати, а отже — обґрунтовано ухвалювати рішення щодо нього.

Чи варто прискорювати створення емуляції мозку?

Коли проблема контролю штучного інтелекту починає здаватися нездоланною, виникає спокуса рухатися менш ризикованим шляхом емуляції цілого мозку. Однак, щоб ухвалити зважене рішення, варто зауважити кілька моментів[522].

Насамперед проаналізуємо взаємозв'язки емуляції мозку з іншими технологіями. Як ми вже раніше зауважили, розроблення емуляції може призвести до появи нейроморфного ШІ — особливо нестабільної, на нашу думку, форми ШІ.

Проте уявімо на хвилинку, що ми змогли створити емуляцію цілого мозку (ЕЦМ). Чи справді це безпечно? Це складне питання. Можна навести щонайменше три переваги ЕЦМ, кожна з яких сама собою є спірною: (I) на відміну від синтетичного ШІ, характеристики ЕЦМ легше зрозуміти; (II) ЕЦМ може успадкувати людську мотивацію; (III)

у такій системі вибухоподібне зростання інтелекту відбуватиметься довше. Коротко оглянемо кожну перевагу.

I. Цілком можливо, що людині справді буде легше зрозуміти характеристики роботи ЕЦМ, ніж синтетичного ШІ. Людство давно вивчає власний розум і його переваги та недоліки, натомість майже нічого не знає про інтелект штучний. Але одне діло знати, що може і що ні цифровий образ людського мозку, інше питання, як він реагуватиме на спроби тим чи тим способом його покращити. Водночас однією з вимог до створення ШІ від початку може бути його передбачуваність: його стан (статичний чи динамічний) має бути чітко визначений у будь-яку мить роботи. Отже, оскільки ЕЦМ на стадії розроблення справді може бути передбачуванішою, ніж ШІ, немає повної впевненості в тому, що ЕЦМ залишиться такою ж передбачуваною у динаміці зростання, особливо порівняно з добре спроектованим ШІ, з усіма можливими заходами безпеки.

II. Також немає жодної гарантії, що емуляція успадкує мотивацію свого прототипу. Для збереження його оціночних схильностей може знадобитися надто висока деталізація сканування. Крім того, навіть якщо вдасться повністю копіювати мотивацію людини, чи буде це безпечно? Люди бувають ненадійними, користолюбними і жорстокими. Хоч прототип для емуляції, певно, буде максимально достойним, проте складно передбачити реакцію людини, яка раптом опиняється в абсолютно незнайомому середовищі, наділена надлюдським інтелектом і має спокусу всесвітнього панування. Принаймні логічно припустити, що мотивація емуляції буде *схожою на людську* (а не пов'язаною зі створенням скріпок чи обчисленням розрядів числа пі). Словом, цей аргумент може дещо заспокоювати лише людину, схильну до оптимістичного погляду на людську-природу[523].

III. Немає впевненості в тому, що перехід емуляції до суперінтелектуальності відбуватиметься повільніше. Імовірно, апаратне переважування в емуляції буде меншим. Адже вона поступається ШІ в обчислювальній ефективності. Також штучному інтелекту простіше інтегрувати до своїх систем додаткові комп'ютерні засоби, до яких він зможе отримати доступ. А емуляція переважно розвиватиметься шляхом пришвидшення та

збільшення популяції. Якщо емуляція справді довше досягатиме суперінтелектуальності, це може дати додатковий час для вирішення проблеми контролю. Окрім того, повільніше зростання створює передумови для багатополярності. Проте невідомо, чи це буде краще.

Ідея про нібито більшу безпечність емуляції ускладнюється також перспективою *другого переходу*. Навіть якщо першою з'явиться емуляція цілого мозку, це не виключає подальшого створення повністю штучного інтелекту. Зріла форма ШІ має низку важливих переваг над ЕЦМ[524]. З появою потужнішої технології ШІ ЕЦМ стане застарілою (придатною хіба що для консервування розумів окремих людей), тоді як обернену ситуацію уявити досить важко.

А отже, якщо першим з'явиться ШІ, то вибух інтелектуальності може відбутися в одну хвилю. Але якщо першою з'явиться технологія ЕЦМ, згодом після її появи може з'явитися синтетичний ШІ, і відбудеться другий вибух. Загальна ризикованість сценарію «ЕЦМ-ШІ» обчислюється як сума ризиків першого і другого переходів (за умови успішного подолання першого); див. рисунок 13[525].

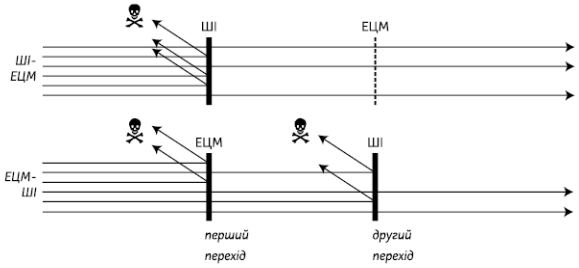


Рисунок 13. **Що перше: штучний інтелект чи емуляція цілого мозку?** Сценарій «ШІ–ЕЦМ» обіцяє один екзистенційно небезпечний перехід. Сценарій «ЕЦМ–ШІ» матиме два послідовні перехідні процеси — поява ЕЦМ, потім поява ШІ, — і обидва можуть мати екзистенційно катастрофічні наслідки. Загальна ризикованість сценарію «ЕЦМ-ШІ» вимірюється сумою ризиків від обох переходів. Однак успішний перебіг першого переходу дещо зменшить ризикованість другого.

Але наскільки безпечнішим може бути другий перехід: створення ШІ у світі емуляцій? Імовірно, коли штучний розум уже існує в іншій формі, створення цифрового ШІ буде менш вибуховим. Розумова різниця між великою кількістю емуляцій, які працюватимуть зі швидкістю комп'ютерів, і новоствореним ШІ буде менша ніж між ним і людиною, що дасть змогу емуляціям легше його контролювати. Проте не варто переоцінювати цей фактор, бо різниця інтелектів емуляції і ШІ в

абсолютному вимірі все-таки буде значною. Якби емуляції були не лише численніші і швидші, а водночас і розумніші, ніж люди (або хоча б такі, як найрозумніші з них), тоді, можливо, створення спочатку ЕЦМ мало б сенс — принаймні такий варіант не поступався б покращенню біологічного мозку, яке ми розглядали раніше.

Також можна припустити, що у створенні ЕЦМ лідер матиме більшу часову перевагу. Уявіть сценарій, де проект, якому вперше вдалося створити емуляцію мозку, на шість місяців випереджає свого найближчого конкурента. Із доступом до швидких комп'ютерів такі емуляції могли б присвятити весь свій суб'єктивний час створенню безпечного цифрового ШІ. Із прискоренням у сто тисяч разів і шістьма місяцями астрономічного часу, вони могли б спокійно працювати над проблемою контролю п'ятдесят тисячоліть, перш ніж у них з'явилися б конкуренти. Із достатньою кількістю апаратних засобів для прискорення своєї роботи вони могли б створити безліч копій себе, щоб розділити загальне завдання на підпроцеси і розподілити їх між окремими групами. Якщо ж перші емуляції витратять свою перевагу на створення синґлтону, то потім зможуть працювати над вирішенням проблеми контролю необмежену кількість часу[526].

Здається, що в разі створення ЕЦМ ризики, пов'язані з наступним переходом до ШІ, будуть меншими. Проте якщо врахувати, що попередній перехід до ЕЦМ теж має свої ризики, то невідомо, який зі сценаріїв зрештою буде ризикованішим: «ЕЦМ-ШІ» чи «ШІ-ЕЦМ». До першого наразі може схилити лише зневіра у здібностях людства контролювати ШІ — навіть якщо до того часу людство чи цивілізація зможе значно покращити свої здібності.

У нашому пошуку відповіді ми можемо врахувати ще деякі важливі аспекти. Найважливіший з них є пов'язані технології: розроблення ЕЦМ може привести до створення нейроморфного ШІ. Це аргумент не на користь ЕЦМ[527]. Звісно, недолугий синтетичний ШІ може бути небезпечнішим від нейроморфного. Проте загалом нейроморфний ШІ менш безпечний. Зокрема тому, що у процесі його розроблення занадто часто виникатиме бажання скоротити шлях: замінити розуміння простим копіюванням. Для створення чогось із нуля зазвичай потрібен досить високий рівень розуміння того, як система має працювати, тоді як для копіювання часто розуміти всі деталі

роботи необов'язково. Емуляція мозку полягає в детальному копіюванні біології розумових процесів. Для цього може бути необов'язково розуміти всі розумові процеси на рівні обчислень — достатньо добре знати на рівні компонентів. У цьому нейроморфний ШІ схожий на емуляції: він складений з частинок, списаних з біологічного мозку. Часто для того щоб скопіювати, не треба детально розуміти математику роботи. Проте існує важлива відмінність між нейроморфним ШІ і ЕЦМ: відсутність у першого людської системи мотивації[528]. Тому, працюючи над ЕЦМ, варто намагатися уникати створення нейроморфного ШІ.

Крім того, варто врахувати, що про появу ЕЦМ ми зможемо знати заздалегідь. Природа синтетичного ШІ така, що процес його створення будь-якої миті може змінити деяке неочікуване відкриття і визначити долю всієї розробки. Натомість створення ЕЦМ — тривалий і складний процес, який потребує виконання багатьох проміжних етапів та створення багатьох допоміжних технологій: високошвидкісне сканування, розпізнавання зображень і трансляція в код, детальне моделювання нервових тканин. Тож можна бути впевненим, що ЕЦМ з'явиться не завтра (а щонайменше років за 15–20). Тому спроби пришвидшити розроблення ЕЦМ можуть вплинути лише на сценарії, за якими штучний інтелект буде створено значно пізніше. А отже, інвестувати в ЕЦМ буде той, хто бажає запобігти екзистенційним ризикам майбутнього, але не прагне занадто швидкої появи ШІ, поки не вирішено проблеми контролю над ним. Але поки що надто рано намагатися визначити конкретні терміни та часові межі наведених процесів, тож не варто надавати цьому аргументу великої ваги[529].

Отож прагнути ЕЦМ можна з кількох причин: (а) невпевненість у тому, що людство зможе вирішити проблему контролю; (б) переконання, що небезпеки нейроморфного ШІ, багатополярного сценарію, другого переходу насправді перебільшені; (в) переконання, що час появи ЕЦМ і ШІ насправді не настільки відрізняється і (г) бажання, щоб суперінтелект з'явився не надто пізно, але й не зарано.

З особистісної перспективи потрібна швидкість

Боюся, коментатор під ніком «washbash» висловив думку багатьох людей:

Мені хочеться, щоб [розроблення ШІ йшло] швидше. Не те щоб це було краще для світу. Яке мені буде діло до світу, коли я помру? Просто я хочу, щоб вона, чорт забирай, рухалася швидше! Так я принаймні матиму шанс побачити технологічно досконале майбутнє[530].

З особистісного погляду ми маємо всі підстави прагнути якомога швидшого прогресу, не нехтуючи найрадикальнішими засобами, навіть якщо вони можуть загрожувати існуванню людства як виду. Адже, так чи інакше, через сто років усі ми, хто живе сьогодні, будемо мертві.

Особливо багато для нас важить аргумент, що ШІ здатен допомогти у створенні медичних технологій, які в майбутньому можуть продовжити нам життя. Завдяки цьому з'явиться ймовірність, що ми побачимо появу суперінтелекту на власні очі. Якщо революція ШІ буде успішною, суперінтелект, який з'явиться в результаті, точно зможе винайти засоби, що дадуть змогу всім живим на той час людям жити стільки, скільки вони забажають, а також не тільки відновити здоров'я і добре самопочуття, але й розвинути свої здібності далеко за межі того, що ми зараз називаємо людськими можливостями. Ба більше, з допомогою суперінтелектуальних технологій завантаження свідомості люди зможуть узагалі позбавитися своїх смертних оболонок і переселитися в цифровий світ, отримавши нові здорові віртуальні оболонки. Деякі з інших технологій, які безпосередньо не стосуються продовження життя, теж можуть бути цікавими, бо даватимуть змогу покращити якість життя[531].

Так можна виправдати ще багато ризикованих технологічних інновацій, які обіцяють прискорити появу суперінтелекту — навіть ті, які будуть небажаними з безосібної перспективи. Для нас особисто головною цінністю цих інновацій буде те, що вони можуть допомогти дочекатися світанку нової ери. От лиш питання: чи судилося нам стати її свідками? Так через особистісні причини ми схвалюємо швидкий розвиток електроніки, як і створення ЕЦМ. Особисто для нас шанс протягом нашого життя побачити на власні очі вибух інтелектуальності машин і появу суперінтелекту переважає всі екзистенційні ризики, які може принести людству цей процес[532].

Співпраця

Те, чи зможе людство створити умови для глобальної взаємодії та спільної участі в розробленні ШІ, неабияк впливатиме на результат його появи. Співпраця буде вигідна усім. Погляньмо, як цей фактор може вплинути на наслідки появи ШІ та що ми могли б зробити для розширення й поглиблення співпраці.

Конкуренція та її загрози

Конкуренція навколо створення ШІ — це коли учасники одного проекту створення ШІ змушені враховувати, що існує інший проект, який може досягти ідентичної цілі першим і такий стан є для них небажаним. Водночас для конкуренції необов'язково, щоб цей інший проект насправді існував. Навіть один проект може працювати в стані конкуренції, якщо його учасники переконані, що мають конкурентів. Союзники не створили б ядерну бомбу так швидко, якби не були переконані (помилково), що Німеччина теж дуже близька до цього.

Ступінь конкуренції (тобто розподіл пріоритетів учасника на користь швидкості, а не безпечності) залежить від низки факторів, як-от відрив між конкурентами, порівняний вплив на результат здібностей та удачі, кількість конкурентів, відмінності їхніх цілей і підходів, які конкуренти застосовують для досягнення. Важливо також те, як суперники самі оцінюють наведені фактори (див. додаток 13).

Додаток 13. Перегони без правил від початку до кінця

Розглянемо гіпотетичну гонитву в озброєнні ШІ, коли кілька сторін прагнуть першими створити суперінтелект[533]. Кожна самостійно вирішує, яку частину доступних ресурсів витрачати на безпеку, а яку — на розроблення ШІ. Якщо між суперниками не існує жодних домовленостей чи угод (не вдалося дійти згоди чи забезпечити виконання), то в таких перегонах без правил учасники можуть не надто перейматися безпекою.

Успішність кожної з команд можна виразити функцією від її спроможності (до якої входять здібності учасників і удача) та факторів стримування, пов'язаних із витратами на заходи безпеки. Найуспішніша команда створить ШІ перша. Загрози ШІ залежатимуть від кількості інвестицій команди в заходи безпеки. У

найгіршому разі спроможність команд буде однаковою. Отже, переможець визначатиметься лише за інвестиціями в безпеку: команда, яка найбільше зекономить на заходах безпеки, здобуде перемогу. Тож рівновагою Неша в цій грі буде, якщо всі учасники відмовляться від витрат на безпеку. На практиці таке справді може трапитися завдяки ефекту зашморгу: учасники скорочуватимуть витрати на безпеку одне за одним, намагаючись не відставати, поки не досягнуть максимального рівня ризикованості.

Спроможність проти Ризику

Якщо спроможність команд різна, тоді ситуація інша. Окремо від витрат на безпеку, відмінності у спроможності учасників можуть послабити ефект зашморгу: яка потреба в ризику, якщо він не впливатиме на позицію учасника в перегонах. Залежність ризиків від спроможності для різних сценаріїв наведено на рисунку 14. Графіки показують залежність безпечності ШІ від порівняної важливості спроможності. Інвестиції в безпеку (вісь *y*) змінюються від 1 (ідеально безпечний ШІ) до 0 (повна відсутність засобів безпеки). На осі *x* позначено відносну вагу спроможності для визначення швидкості розроблення ШІ, порівняно з інвестиціями в безпеку. (Значення 0,5 означає, що інвестиції в безпеку вдвічі важливіші від спроможності команди; 2 свідчитиме, що спроможність навпаки вдвічі сильніше впливає на прогрес ніж безпековий фактор). Вісь *y* позначає безпечність ШІ, який буде створений (очікувана максимальна корисність переможця).

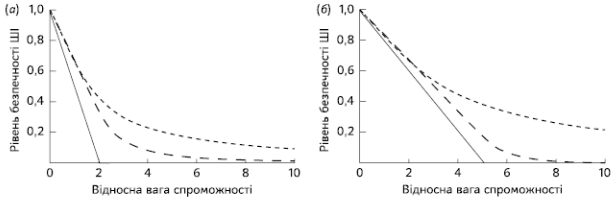


Рисунок 14. **Рівень ризику в технологічних перегонах зі створення ШІ**
Рівні ризикованості ШІ у змодельованих технологічних перегонах за участі (а) двох команд та (б) п'яти команд зображено залежно від відносної ваги спроможності (на противагу інвестиціям у безпеку) у функції успішності команд. На кожному малюнку зображено три графіки залежно від поінформованості учасників: суцільна лінія — відсутня будь-яка інформація про оцінку спроможності; штрихова лінія — доступна лише інформація про власну спроможність; і пунктирна — інформація про спроможність усіх учасників публічно доступна.

Як ми можемо бачити з графіків, коли спроможність не впливає на результат розроблення, небезпечність ШІ найвища і потроху

знижується під час зростання ваги спроможностей.

Поділ цілей

Ще один спосіб зменшити ризики — дозволити командам інвестувати в успіх конкурентів. Якщо суперники будуть переконані, що поразка означатиме остаточну втрату всього, вони підуть на будь-які ризики заради шансу на перемогу. І навпаки, команди більше інвестуватимуть у безпеку, якщо вони менше залежатимуть від того, хто переможе. Тому під час створення ризикованих технологій нам варто всіляко заохочувати перехресне інвестування.

Кількість суперників

Що більша кількість суперників, то більше ризику: кожна окрема команда має менше шансів бути першою, тому готова більше ризикувати, щоб поліпшити свої шанси на перемогу. Це видно, якщо порівняти графіки для двох команд (рисунок 14, *а*) та п'яти (рисунок 14, *б*). Для будь-якого сценарію, що більше учасників, то більше ризику. Ризики зменшаться, якщо команди переформуються в меншу кількість коаліцій.

Горе від розуму

Якщо команда знатиме своє положення в загальному заліку (наприклад, за показником спроможності) — добре це чи погано? Тут діють різні фактори. З одного боку, бажано, щоб лідер знав про своє лідерство (тоді він зможе дозволити собі більше витрачати на безпеку). Водночас тому, хто перебуває в кінці заліку, краще не знати цього (інакше він далі скорочуватиме видатки на безпеку, наражаючи всіх на додаткові ризики). Хоч може здатися, що в будь-якого варіанта є свої плюси, моделювання недвозначно свідчить про те, що інформованість впливає на результат негативно[534]. На рисунках 14, *а* і 14, *б* наведено по три сценарії: прямими суцільними зображено залежність для ситуації, коли жоден з учасників не знає показників спроможності ні конкурентів, ні власних. Штриховою лінією зображено залежність від випадку, коли учасники знають тільки власні показники спроможності (команда піде на додаткові ризики, тільки коли її

показник буде занизький). Пунктирна лінія відповідає випадку, коли всі учасники мають доступ до повного заліку (команда може йти на додаткові ризики навіть тоді, коли її значення наближається до значення конкурента). Як видно з графіків, збільшення поінформованості погіршує безпечність результату.

Є всі підстави вважати, що поява штучного суперінтелекту відбуватиметься в умовах помірної конкуренції. Проте, ймовірно, конкуренція виявиться досить жорсткою. Це багато в чому обумовить наше сприйняття стратегічних викликів, які супроводжуватимуть вибух інтелектуальності.

Висока конкуренція може змусити розробників пришвидшити роботу, скоротивши видатки на вирішення проблеми контролю. Окрім того, суперництво може згубно впливати, породжуючи вороже ставлення конкурентів одне до одного. Уявіть, що дві країни ведуть розроблення суперінтелекту і одна з них має перевагу. У ситуації, коли переможець отримує все, інші проекти можуть зважитися на будь-які безрозсудні вчинки, у надії вирватися вперед, замість того щоб сумирно очікувати власної поразки. У таких умовах лідер теж може діяти на випередження. Протистояння, зрештою, може навіть спровокувати кровопролиття, особливо якщо суперники — цілі країни[535]. (Навіть спроба зірвати розроблення суперника «хірургічно» точним ударом може спровокувати ширший конфлікт; крім того здійснити його може бути неможливо[536]).

Якщо суперниками будуть не країни, а менші організації, як-от корпорації чи наукові колективи, конфліктні ситуації матимуть значно менш деструктивні наслідки. Проте для результатів розроблення наслідки будуть такими ж негативними. Адже найбільша шкода від суперництва тут полягає не в конфліктах навколо розробки, а в утраті обережності. Як ми вже показали, загострення конкуренції призведе до зменшення витрат на безпеку. Водночас конфлікт, навіть витриманий у цивілізованих межах, зменшить шанси на співпрацю, бо в атмосфері ворожості й недовіри ні в кого не виникне бажання ділитися ідеями розв'язання проблеми контролю[537].

Переваги співпраці

Що ж до співпраці, то вона принесе всім багато користі. Не буде причин поспішати зі створенням ШІ. Можна буде приділити вдосталь уваги та ресурсів питанням безпеки. Співпраця дасть змогу уникнути конфліктів та насилля. Атмосфера взаємодопомоги сприяє обміну ідеями, зокрема — ідеями вирішення проблеми контролю. До всіх цих переваг додамо ще одну: результат спільної роботи — успішний керований вибух інтелектуальності ШІ — створює всі передумови для справедливого розподілу його плодів.

Однак широка співпраця не завжди приводить до ширшого розподілу її результатів. Теоретично, невеликі проекти, започатковані ентузіастами з альтруїстичною метою, теж можуть прагнути розподіляти результати рівномірно чи справедливо поміж усіх морально значущих істот. Але з низки причин ширша співпраця більшої кількості учасників має кращий дистрибутивний потенціал. Насамперед, логічно, що спонсори розраховують отримати справедливий дохід зі своєї участі в проекті. Тому якщо проект успішний, що більше в нього інвесторів, то більше людей отримають свої дивіденди. Далі, що ширша співпраця, то більша ймовірність, що результат також принесе користь іншим людям, не залученим у проекті. Багато учасників має ширше коло особистих зв'язків, тож інтереси всіх цих пов'язаних людей також опосередковано представлені у проекті. Для великого проекту імовірність, що серед учасників є альтруїсти, які мають на меті загальне благо, вища. Великі спільні проекти часто перебувають у фокусі уваги громадськості, тому їхні результати не можуть бути непомітно приватизовані невеликою групою програмістів чи приватних інвесторів[538]. Характерно, що більший проект, то менше йому коштуватиме забезпечити інтереси найширшого кола людей. (Наприклад, якщо у проекті беруть участь (так чи інакше) 90 відсотків людей, то для охоплення всіх 100 відсотків людей учасники мають скоротити свої частки лише на 10 відсотків).

А отже, високий рівень співпраці може сприяти більшому поширенню результатів (хоч деякі проекти з невеликою кількістю спонсорів також можуть мати на меті благо якнайбільшої кількості людей). Але навіщо це прагнення — охопити результатами всіх?

Сценарій, у якому всі отримують частку благ, бажаний із двох причин: моральної та практичної. З моральною причиною все більш-

менш зрозуміло, тому додамо лише, що не варто засновувати моральні міркування на егалітарних принципах. Тут краще спиратися на принцип справедливості. Науковий проект, який працюватиме над створенням штучного суперінтелекту матиме значну негативну екстерналію — ризик глобального масштабу. Кожна жива істота у світі за згодою чи без, незалежно від усвідомлення цього ризику, наражатиметься на небезпеку. Усі ділять ризики, тому було б справедливо, якби всі могли отримати свою частку благ.

Сценарій глобальної співпраці потенційно може принести кращий результат, тому він привабливіший з погляду моралі.

Практична цінність широкого розподілу результатів створення має два боки. По-перше, широкий розподіл вигоди заохочуватиме подальше розширення співпраці та зменшуватиме негативний вплив конкуренції. Якщо всі виграють, незалежно від того, який проект першим створить суперінтелект, то немає причин для конфліктів. Спонсорам може бути вигідно публічно підкреслювати намір надати загальний доступ до результатів свого проекту, бо альтруїзм може заохотити нові інвестиції та створити привабливий імідж[539].

По-друге, цінність якомога ширшого розподілу результатів розроблення суперінтелекту пов'язана з тим, наскільки схильними до ризиків будуть самі суперінтелектуальні агенти, або, можливо, їхні функції корисності будуть сублінійними щодо ресурсів. Річ у тому, що обсяг ресурсів, про які йдеться, насправді неосяжний. Якщо відомий нам Всесвіт справді незаселений і ми в ньому — єдина розумна життєва форма, тоді на кожну людину наразі припадає щонайменше по одній незаселеній галактиці. Більшість людей швидше погодяться на гарантовану можливість володіти ресурсами однієї галактики, ніж на одну мільярдну можливості згодом отримати мільярд галактик[540]. Навіть якщо гарантована частка для кожної людини у відсотках буде невеликою, розподіл астрономічно великого обсягу загального багатства за таким принципом буде вигідний усім. Коли життя пропонує такий куш, важливо не втратити свій шанс і не залишитися осторонь.

Аргумент, базований на величезному загальному обсязі ресурсів, передбачає, що сам суперінтелект потребує скінченної кількості ресурсів[541]. Але це може бути не так. Низку етичних теорій — зокрема

деякі особливо складні консеквенційні теорії — можуть дати толерантні до ризиків і лінійні щодо ресурсів функції корисності. Мільярд галактик можуть дати в мільярд більше щасливих життів, ніж одна. Тож чисто з утилітарного погляду мільярд галактик варті в мільярд більше, ніж одна[542]. Водночас функція потреб людини має обмежену потребу в ресурсах.

Однак до останнього твердження є два важливі уточнення. По-перше, для багатьох людей важливі статуси й порівняння. Якщо кілька агентів захочуть очолити список найбагатших людей Forbs, то для них не існуватиме достатньої кількості ресурсів.

По-друге, технології постперехідного світу значно розширять можливості виробництва: з ресурсів можна буде створити безпрецедентну кількість благ, включно з недоступними зараз, проте дуже бажаними. Життя мільярдера не довше від життя мільйонера в тисячу разів. Натомість в еру цифрових розумів мільярдер зможе собі дозволити в тисячу разів більше обчислювальних ресурсів, а отже, і в тисячу разів більше суб'єктивного часу життя. Так розумові здібності теж можуть перетворитися на продукт. В умовах, коли економічний капітал можна буде конвертувати в життєво важливі продукти в будь-яких масштабах, безмежна жадоба, навіть для дуже великих багатств, здаватиметься виправданішою, ніж у сучасних багатіїв (зокрема, не помічених у філантропічній діяльності), які змушені витрачати надлишок капіталу на придбання літаків, яхт, об'єктів мистецтва чи четвертої та п'ятої резиденцій.

Тобто егоїст може прагнути якомога більшої частки постперехідного багатства, не зважаючи на будь-які ризики? Не зовсім. Лише обмежену кількість фізичних ресурсів можна перетворити на життєвий час чи розумову діяльність за одиницю часу. Якщо життя повинно проживатися послідовно, щоб спостерігач пам'ятав події минулого і на нього впливали його попередні рішення, тоді тривалість життя цифрового розуму не можна безкінечно збільшувати, не збільшуючи кількості *послідовних* обчислювальних операцій. Закони фізики визначають фундаментальні межі, у яких ресурси можуть бути трансформовані в послідовні обчислення[543]. Саме завдяки цим межам, потужність розумових процесів може мати значну сублінійність відносно витрати ресурсів. Ба більше, невідомо, чи варто егоїсту бути

нейтральним до ризиків навіть за умови існування релевантнішої метрики оцінки результату, як-от кількості суб'єктивних років життя відповідної якості. Більшість людей, наприклад, усе одно вибрала б гарантовані 2000 додаткових років життя, аніж один шанс із десяти прожити ще 30 000 років (навіть за умови, що всі роки життя будуть однаково високої якості)[544].

Тож насправді практичні мотиви рівномірного і якнайширшого розподілу результатів створення суперінтелекту дуже умовні і залежать від ситуації. Але такий розподіл однозначно збільшуватиме шанси людей отримати (майже все), що забажають, не рахуючи загального сприятливого впливу на атмосферу співпраці та зниження ймовірності екзистенційної катастрофи. А отже, якнайширший розподіл результатів є не тільки морально справедливим, але й практично корисним.

Співпраця може мати також інші наслідки, які варто згадати хоча б побіжно: високий рівень співпраці перед переходом до суперінтелектуальності може вплинути на співпрацю після. Припустимо, що людству вдасться вирішити проблему контролю. (В іншому разі рівень його співпраці після появи суперінтелекту навряд чи матиме якесь значення). Тут можливі два варіанти. Перший: коли вибухове зростання інтелектуальності не проходить у режимі «переможець отримує все» (наприклад, тому, що весь процес порівняно нешвидкий). Тоді, найімовірніше, рівень співпраці перед переходом позитивно вплине на схильність до співпраці після нього. Співпраця навколо створення суперінтелекту може продовжитися і потім. Зокрема, взаємодія навколо переходу надалі може допомогти спільно спрямовувати розвиток подій у потрібному (і, імовірно, вигідному для всіх) напрямі.

Другий варіант: режим високої конкуренції, у якому переможець отримує все (перехід ШІ до суперінтелекту відбувається порівняно швидко). У такому разі за відсутності глобальної взаємодії людства може утворитися синґлтон — один із проектів ШІ досягне суперінтелектуальності й отримає вирішальну стратегічну перевагу. Синґлтон за визначенням характеризується високим рівнем співпраці[545]. Тобто відсутність співпраці до переходу спричинить

високий рівень співпраці після переходу. Проте дещо вищий рівень співпраці на початку має більшу варіативність результатів.

Упевнившись, що жоден не має вирішальної стратегічної переваги, дружні проекти можуть синхронізувати свої дії. Або кілька спонсорів можуть підтримувати один проект, однак не даючи йому мандат на створення синґлтону. Так можна уявити міжнародний консорціум, який керує спільним науковим проектом зі створення штучного суперінтелекту, але, щоб зберегти звичний світовий порядок, не дає своєму творінню перетворитися в щось подібне до Організації Об'єднаних Націй.

Отож, ймовірно, що в разі швидкого зростання інтелектуальності ШІ високий рівень співпраці до появи суперінтелекту спричинить подальше зниження співпраці. Щоправда, припинити співпрацю суб'єкти, що взаємодіяли навколо створення суперінтелекту, можуть тоді, коли впевняться, що таке припинення не матиме згубних наслідків. Тобто сценарії, у яких після появи суперінтелекту початково високий рівень співпраці змінюється зниженням, є переважно цілком безпечні.

Узагалі ширша співпраця була б корисною і після появи суперінтелекту. Вона могла б запобігти дистопічним тенденціям: відновленню мальтузіанської умови через високу економічну конкуренцію і зростання кількості населення, еволюційній поразці цінностей гуманізму та перемозі неевдемонічних форм етики, розбрату, ворожнечі та, як результат, війнам і гонкам технологічних озброєнь. Нездорова конкуренція навколо технологій завдасть ще більшої шкоди, якщо перший перехід завершиться появою проміжної форми штучного інтелекту (наприклад, емуляції цілого мозку). Адже конфліктна атмосфера, яка оточуватиме підготовку до другого переходу (створення синтетичного ШІ), не сприятиме вирішенню проблеми контролю.

Раніше ми показали, як співпраця може запобігти конфліктам на порозі вибуху інтелектуальності й у такий спосіб сприяти ефективному розв'язанню проблеми контролю та допомогти морально і водночас раціонально розпорядитися ресурсами. Тож до усіх позитивних наслідків співпраці можна додати ще один: співпраця навколо створення суперінтелекту може допомогти спільно знайти

вирішення проблем, перед якими може постати людство після його створення.

Працюймо разом

Співпраця може мати різні форми — залежно від масштабу сутностей, які беруть у ній участь. Невеликі окремі наукові колективи, які працюють над ШІ і конкурують між собою, можуть зрештою об'єднати зусилля[546]. Корпорації можуть домовитися про злиття чи узгодити перехресне фінансування. На рівні країн можуть бути створені великі міжнародні наукові проекти. Історії відомі кілька прецедентів такої масштабної співпраці у сферах науки та технологій (ЦЕРН, Проект генома людини, Міжнародна космічна станція), але спільна робота над створенням суперінтелекту через значний вплив на глобальну безпеку ставитиме перед учасниками виклики іншого порядку. На відміну від максимально відкритої атмосфери спільних наукових проектів, основою такої розробки має бути максимально жорсткий контроль. Учасники, ймовірно, будуть змушені працювати в ізоляції та спілкуватися зі зовнішнім світом лише через окремий контрольований канал зв'язку. Потрібні для цього заходи безпеки можуть виявитися наразі недосяжними, але невдовзі, уже в цьому сторіччі, завдяки новітнім розробкам у галузі розпізнавання брехні та відеонагляду такі можливості можуть з'явитися. Варто також мати на увазі, що широка співпраця не означає, що в проекті братиме участь багато учених. Це лише означає, що цілі проекту враховуватимуть широкий спектр позицій та думок значної кількості людей. Теоретично такий проект може забезпечити максимально широку участь усього людства (за допомогою представництва, скажімо, Генеральної Асамблеї ООН) і водночас усю роботу виконуватиме один науковець[547].

Співпрацю варто розпочинати якомога раніше, до того, як стане відомо, який із проектів створення ШІ має найбільше шансів на першість в утворенні суперінтелекту. З наближенням до фінішу буде дедалі менше непевності щодо шансів того чи того учасника, а отже — дедалі складніше опиратися спокусі обрати ймовірного переможця, а не проект, який принесе користь усьому людству. З іншого боку, як організувати глобальну співпрацю навколо незрозумілих перспектив появи суперінтелекту та відсутності чіткого шляху його створення? Ба

більше, співпраця може пришвидшити рух цим непевним шляхом, що наразі обіцяє додаткові небезпеки.

Тож ідеальною формою співпраці зараз може бути об'єднання сил у питаннях, які не вимагають надто жорсткої формалізації відносин та не сприяють розвитку штучного інтелекту. Пропозиція, яка відповідає цьому критерію, — це поширення відповідних моральних норм, за якими суперінтелект має служити загальному благу. Ось варіант формулювання такої норми:

Принцип загального блага
Суперінтелект має бути створений, щоб приносити користь усьому людству та керуватися у своїй діяльності загальноприйнятими етичними нормами[548].

Від початку стверджуючи, що величезний потенціал суперінтелекту має належати всьому людству, ми матимемо більше часу на те, щоб ця норма закріпилася й усталилася.

Принцип загального блага не заперечує можливості існування в індивідів та компаній комерційного інтересу до розроблення ШІ. Наприклад, для того щоб забезпечити всезагальний розподіл благ, створених унаслідок діяльності свого суперінтелекту, певна фірма-власник може обумовити серед статутних положень, які стосуються дії обставин непереборної сили, що дохід розміром менше, ніж деяка досить значна сума (скажімо, трильйон доларів щороку) розподілятиметься звично поміж акціонерами й іншими бенефіціарами фірми, і лише частка доходу понад вказану суму спрямовуватиметься на всезагальний розподіл поміж усіма людьми (відповідно до деякого універсального морального критерію). Прийняття такої умови на законодавчому рівні загалом не створюватиме жодного додаткового навантаження, оскільки дуже малоймовірно, що будь-який звичайний бізнес коли-небудь матиме такий дохід (і топ-менеджери здебільшого не враховують такого сценарію у своїх рішеннях). Водночас така умова дасть людству важливу гарантію, що коли суперінтелект нарешті *з'явиться* (можливо, у приватному володінні), переваги від його діяльності вплинуть на всіх. Такий механізм можна застосувати не тільки до компаній. На міжнародному рівні також може бути прийнята умова,

що в разі перевищення рівня ВВП будь-якої країни певного порогу — нехай, 90 відсотків — світового ВВП, надлишок має бути рівномірно поділений між усіма[549].

Особи чи організації, діяльність яких пов'язана з ШІ, можуть для початку добровільно приймати моральні зобов'язання, які так чи інакше відтворюють принцип загального блага. Пізніше цей принцип можна втілити в законі чи угоді, яку виконуватиме значно ширше коло суб'єктів. Наведене нами приблизне формулювання може бути хорошим початком, з якого в ідеалі має постати низка конкретних вимог, виконання яких легко перевірити й контролювати.

5 Книжка писалася до вікопомного референдуму у Великій Британії, що поклав початок Брекзіту. — *Прим. пер.*

15. ПЕРЕЛОМНИЙ МОМЕНТ

І от ми опинилися в нетрях заплутаних стратегій, оточені туманом непевності. Так, ми розрізняємо обриси деяких окремих ідей, проте їхні деталі розмиті, а зв'язки між ними до кінця не зрозумілі. Крім того, деякі фактори можуть бути від нас прихованими. Що нам робити в цій ситуації?

Філософія на межі

Один мій колега любить повторювати, що якщо тебе нагородили медаллю Філдса (найвища відзнака для математика), про тебе можна точно сказати дві речі: що ти міг зробити щось справді важливе і що ти цього не зробив. Досить різке зауваження, проте не без зерна правди.

Погляньмо на «відкриття», як на дію, що переносить інформацію з пізнішої точки на шкалі часу в ранішу. Найбільша цінність відкриття не в переміщеній ним інформації, а у важливості того, щоб вона опинилася саме в цій точці часу. Фізик чи математик може розв'язати задачу, яка іншим була не під силу, але чи такий уже цінний цей розв'язок, якщо невдовзі хто-небудь усе одно її розв'язав би? Звісно, іноді результат, отриманий навіть трохи раніше, є надзвичайно важливим, але це зазвичай буває, коли вже наперед відомо, як застосувати цей результат: для практичної мети або як основу для подальшого теоретичного пошуку. В останньому випадку, коли відкриття є цеглинкою в більшому будинку теорії, справді важливо мати її якнайраніше, лише якщо теорія ця дуже важлива і принесе практичну користь саме зараз[550].

А отже, головна ідея не в тому, чи результат, за який ви отримали медаль Філдса, справді «потрібний» (інструментально чи інформаційно). Питання в тому, чи був він вартий того, щоб якнайшвидше опублікувати його. Саме користь від такого часового

переміщення варто порівнювати з користю, яку міг принести математик світового рівня, якби працював над чимось іншим. Принаймні деякі медалісти точно можуть бути прикладом того, як людина витрачає життя на розв'язання непотрібного завдання — зокрема того, яке вабить майже виключно своєю складністю.

Тим самим можна дорікнути іншим науковим сферам — наприклад, філософії. Деякі питання, які розглядає філософія, можуть неабияк допомогти в пошуку шляхів виживання людства — кілька з них ми згадали в цій книжці. Проте окремі галузі філософії не мають стосунку не лише до екзистенційних ризиків, а й до практичної цінності загалом. Деякі питання, які ставить філософія, подібно до математики, можуть бути цінними самі собою, тобто в людей є підстави цікавитися ними незалежно від їхнього практичного застосування. Так цікавою сама собою є фундаментальна природа реальності. Якби ніхто не вивчав метафізику, космологію чи теорію струн, світ був би, на мою думку, значно менш цікавим. І тепер багатовікова жага знань починає сяяти новими барвами в новому світлі наближення вибуху інтелекту.

Схоже, філософія може максимізувати свої досягнення не безпосередньо теоретизуючи, але опосередковано: делегуючи цю справу посереднику. Фундаментальні відкриття в науці і філософії — лише частина тих справ, у яких суперінтелект (однак, може, вистачить лише трохи вдосконаленого інтелекту людського рівня) матиме перевагу перед клікою сучасних мислителів. На мою думку, потрібна стратегія відкладеного задоволення. Ми могли б відкласти на деякий час пошук відповідей на вічні питання, залишивши його нашим, імовірно, компетентнішим послідовникам, і натомість зосередитися на важливішому завданні: зробити все, щоб такі послідовники в нас були. Це була б ефективна філософія й ефективна математика[551].

Що потрібно зробити?

Отже, треба зосередитися на важливих і нагальних проблемах, які мають бути вирішені перед вибухом інтелекту. Варто також бути обережними й намагатися не зашкодити (скажімо, утриматися від створення потенційно ризикованих технологій). Так, наприклад, не варто поки що поспішати з розв'язанням деяких технічних проблем

штучного інтелекту, які пришвидшать його створення, але не вплинуть на прогрес у контролі роботи ШІ та його безпеки.

Окреслити коло таких корисних та нагальних проблем непросто. Через стратегічну невизначеність способів зниження екзистенційних ризиків навіть цілком позитивний на перший погляд вплив може виявитися неефективним або навіть нашкодити. Щоб убезпечити себе від відверто шкідливих чи морально хибних дій, треба вибирати *надійно корисні* напрями роботи (які позитивно вплинули б на цілу низку можливих сценаріїв) та використовувати в ній загально виправдані засоби (прийнятні для низки моральних позицій).

Існують інші фактори, які варто врахувати під час визначення пріоритетності завдань. Ми любимо, коли проблеми, над якими нам доводиться працювати, *еластичні* до спрямованих на них зусиль. Високоеластичні проблеми можуть бути вирішені значно швидше або значно більше за тої самої кількості зусиль. Дуже важливо, вкрай потрібно і безперечно корисно для всіх, щоб світ був лагіднішим і добрішим, але це проблема з низькою еластичністю: досі ніхто не знає, як це зробити. Так само добре було б досягти миру у всьому світі, проте, зважаючи на численні спроби й нездоланні перепони на шляху будь-якого швидкого вирішення цієї проблеми, зусилля ще кількох людей навряд чи змінять ситуацію.

Щоб запобігти ризикам революції штучного інтелекту і водночас врахувати згадані фактори, ми пропонуємо зосередитися на аналізі стратегій та нарощуванні потенціалу. Адже можна бути впевненим — ніколи не зайве усвідомлювати можливі стратегії розвитку подій та мати технологічний потенціал. Ці завдання достатньо еластичні: порівняно незначне збільшення інвестицій у цю діяльність може значно покращити результати. Зрештою, такі дії завжди на часі, бо прогрес у цих напрямах акумулюватиметься і стане каталізатором для будь-яких наступних досліджень. Крім двох названих напрямів, варто розглянути ще кілька перспективних цілей.

Пошук стратегічного просвітлення

Аналітичне осмислення ситуації здається особливо доречним та потрібним кроком на фоні плутанини й непевності[552]. Розуміючи стратегічне становище, ми зможемо точніше спрямовувати свої дії. В

умовах радикальної невизначеності не лише в якій-небудь другорядній деталі, а навіть у ключових властивостях основного об'єкта, стратегічний аналіз потрібен як ніколи. Часто ми нічого не знаємо навіть про характер ключових параметрів — не можемо сказати, у який бік їх варто змінювати. Однак, імовірно, наше невігластво можна вилікувати. Цей досі не пещений увагою науки напрям може приховувати у своїй товщі діаманти нового знання буквально відразу під поверхнею.

Під стратегічним аналізом ми розуміємо пошук *вирішальних знань*: ідей чи аргументів, що здатні змінити наш погляд не тільки на деталі реалізації, але й на загальну топологію бажаного[553]. Утрата найменшої зернини такого вирішального знання може зруйнувати наші найблагородніші зусилля або й зовсім обернути їх проти нас, як солдата, що зайняв не той бік. У пошуку вирішальних знань (як описових, так і нормативних) нам часто доведеться перетинати межі дисциплін та наук. Через відсутність усталеної методології потрібно буде віднайти оригінальні підходи та методи такого пошуку.

Нарощення потенціалу

Ще одна важлива діяльність поряд зі стратегічним аналізом, потрібна за будь-якого сценарію розвитку подій, — формування відповідної комплексної ресурсно-світоглядної основи для конструктивного сприйняття імовірного майбутнього. Завдання цієї основи — акумулювати ресурси, які можуть знадобитися для досліджень та аналізу. У разі потреби ресурси можна буде швидко переспрямувати відповідно до нових пріоритетів. Така основа є активом загального призначення, який можна застосовувати відповідно до поточних наукових потреб.

Одним із компонентів такого цінного активу може стати мережа донорів — спільнота раціональних філантропів, які усвідомлюють небезпеку екзистенційних загроз та бажають знайти найкращий метод захисту від них. Від мудрості й альтруїзму першопрохідців залежить становлення культури досліджень і створення суперінтелекту, перш ніж сформуються інтереси учасників і накопичиться досвід їхньої взаємодії, а також, імовірно, ринкові відносини між ними. Початкові умови формуватимуть насамперед люди, яких залучатимуть до роботи

засновники напряму. Спершу, щоб упевнитися, що дослідники приділяють достатньо уваги безпеці та теоретичним основам (і залучатимуть у колектив людей зі схожим ставленням), їм не варто ставити високі цілі.

Важливою змінною процесу створення суперінтелекту та проектів, які беруть у ньому участь, є якість «соціальної епістемології». Займатися пошуком вирішальних знань варто, лише якщо вони будуть враховані в ухваленні рішень. Адже це не завжди можливо. Уявіть проект, учасники якого багато років працювали над створенням прототипу ШІ, витратили багато мільйонів доларів і подолали багато технічних перешкод. Проект саме почав активно розвиватися, і для успішного завершення залишилося докласти лише мінімум зусиль. І раптом виявляється, що для того, щоб ШІ був безпечнішим, треба було застосувати зовсім інший підхід. Невже такий проект, як зганьблений самурай, накладе на себе руки, а учасники знищать недостатньо безпечні результати своєї праці? Чи, може, як сполоханий восьминіг, випустить хмару мотивованого скептицизму в надії уникнути нападу? Звісно, з погляду безпеки, краще, щоб такий проект обрав долю самурая[554]. Проте дуже важко уявити процеси та інституції, які готові вчинити сеппуку через сумнівні заяви, ґрунтовані на спекулятивних міркуваннях. Ще одне слабке місце соціальної епістемології — це управління чутливою інформацією, зокрема низька здатність запобігти витоку інформації з обмеженим доступом. (Захистити таку інформацію може бути досить важко, зокрема, коли з нею працюють науковці, які звикли постійно торочити лише про свою роботу й публікувати її результати на кожному стовпі).

Конкретні заходи

Крім загальних заходів — стратегічного аналізу та збільшення потенціалу можливостей, — існують деякі конкретніші цілі, які можуть бути цікавими в цьому контексті.

Однією з них є прогрес в безпечності штучного інтелекту. Рухаючись до цієї мети, важливо ефективно протистояти інформаційним загрозам. Методи розв'язання проблеми контролю можуть також допомогти вирішити проблему вибору здібностей. Можливо, варто зовсім відмовитися від тих умінь, які шкодять стабільності ШІ.

Крім того, можна сприяти формуванню та поширенню поміж розробниками ШІ «найкращих практик». Інформація про просування вирішення проблеми контролю має бути доступна всім. Щоб запобігти неконтрольованій появі суперінтелекту, деякі види комп'ютерних процесів, зокрема ті, які використовують метод рекурсивного самовдосконалення, мають відбуватися в умовах жорсткого контролю. З наближенням моменту появи ШІ зростатиме нагальність розроблення конкретних методів гарантування безпеки. Не завадило б, щоб зараз усі учасники процесу створення ШІ брали на себе *зобов'язання безпеки,* зокрема визнавали принцип загального блага та пріоритетність безпекових аспектів у створенні ШІ. Звісно, небезпечну технологію не зробити безпечною лише словами, але слова поступово торують дорогу думці.

Водночас можуть з'явитися інші сприятливі нагоди: наприклад, спосіб знизити який-небудь екзистенційний ризик чи покращити розумові здібності біологічного мозку та наш колективний розум, або навіть гармонізувати відносини між країнами.

Що найкраще в людині, відгукнися

Перед перспективою інтелектуального вибуху, ми — люди — схожі на дітей, які граються з бомбою. Саме таке враження виникає від невідповідності між потужністю нашої іграшки і легковажністю, з якою ми до неї ставимося. Ми не готові зараз, і ще довго не будемо готові до викликів суперінтелекту. Невідомо, коли вибухне ця бомба, але якщо притулити її до вуха, можна почути тихеньке цокання.

Найкраще, що може зробити в такій ситуації дитина, — це обережно покласти предмет на землю, швидко вийти з кімнати і покликати дорослого. Проте ми маємо справу не з однією дитиною, їх багато, і кожна з них має власний детонатор. На жаль, майже неймовірно, що всі вони погодяться покласти небезпечну річ на підлогу. Певно, знайдеться розумник, який натисне на кнопку, щоб просто подивитися, що трапиться.

Тікати немає сенсу — вибух інтелектуальності рознесе всю будівлю. Дорослих поблизу не видно.

Погодьтеся — не надто радісна ситуація. Переляк та паніка — значно відповідніші емоції. Однак мусимо бути рішучими й максимально зібраними, як перед важливим іспитом, від якого залежать усі сподівання та мрії про майбутнє.

Мова не про фанатичність. Найімовірніше, від вибуху інтелектуальності нас відділяє ще багато десятиліть. Ба більше, найголовнішим викликом для нас буде, навпаки, зберегти людяність: утримати зв'язок зі здоровим глуздом, доброзичливістю та гідністю, — навіть коли ми опинимося перед цією найбільш неприродною та нелюдською з проблем. Для її вирішення нам знадобиться вся наша винахідливість.

Але ми не повинні випускати з поля зору глобально важливих речей. Крізь туман буденності ми можемо бачити лише ледь помітні контури головного завдання нашої епохи. І хай яким непевним і малозрозумілим є об'єкт нашої уваги, у цій книжці ми спробували розгледіти деякі його деталі. Водночас усе, що нам залишається, з погляду моралі, — це оминати екзистенційні загрози та дотримуватися цивілізаційної траєкторії взаємно доброзичливого використання наших спільних ресурсів.

ПІСЛЯМОВА

З моменту виходу цієї книжки у твердій обкладинці у ставленні широкого загалу до її проблематики намітилися деякі зрушення. Стало легше розглядати суперінтелект серйозно: що зростання інтелектуальності машин може відбутися в цьому сторіччі; що таке зростання може виявитися однією з найважливіших подій в історії людства; що подія ця може мати як значні переваги, так і величезні ризики; що розумно було б потурбуватися наперед, збільшити ймовірність сприятливого результату. Звісно, жодна журналістська спроба торкнутися теми поки що не обходиться без блазнювання на тему Термінатора. Але за цією науково-популярною мішаниною вже можна — якщо наставити вухо під правильним кутом — почути тиху, але зрілішу дискусію.

Швидкість розвитку галузі машинного навчання перевершила очікування багатьох. Останнім часом перед людством відкрилося широке поле нових ідей, які варто досліджувати: нейронні машини Тюрінга, глибинне навчання з підкріпленням (deep reinforcement learning, DRL), баєсова оптимізація гіперпараметрів, ґратчасті ДКЧП (grid LSTM, grid long short-term memory), мережі пам'яті (memory network), варіаційні автокодувальники (variational autoencoder, VAE), векторне представлення речень (sentence embedding), генеративні змагальні мережі (GAN, generative adversarial network), генеративні моделі, основані на увазі (attention-guided GAN, self-attention GAN), нові підходи ймовірнісного програмування — усе це лише деякі з багатьох напрямів останніх досліджень. Жваво розвивається глибинне навчання. Завдяки зростанню потужності комп'ютерів, доступності великих наборів даних і вдосконаленню математичних алгоритмів з'явилися методи глибинного навчання — переважно ґрунтовані на багатошарових нейронних мережах (multi-layered neural networks, MNN). Вони дали комп'ютерам змогу наздогнати, а іноді перевершити

людину в багатьох видах діяльності, пов'язаних зі сприйняттям: читання рукописного тексту, розпізнавання та класифікація зображень, розпізнавання усного мовлення та облич. Методи глибинного навчання добре себе зарекомендували в перекладі природних мов і науковому аналізі. Усі ці здобутки головно завдячують алгоритмам, які дають змогу, без участі та допомоги людини, без наперед заданих ознак чи знань, виокремлювати абстрактні розподілені представлення з необроблених сенсорних даних. Можливості, які створюють ці алгоритми, згодом можуть стати основою складніших умінь.

Багато технологій, основаних на машинному навчанні, уже досягли такого рівня розвитку, що можуть приносити реальну користь. Це означає, що наступні покращення відразу конвертуються в дохід від комерційного використання оновленої технології. Якщо трохи покращити, скажімо, систему розпізнавання мовлення, яка надто неточна й погано працює, це не дало б багато користі. Натомість, якби система відразу була настільки досконалою, що застосовувалася всюди, то покращення її роботи на один відсоток могло б принести мільярд доларів доходу. Захопливе сходження машинного навчання відразу за кількома напрямами, заряджене комерційним інтересом, приваблює в галузь ще більше грошей і талановитих людей.

Куди сягне ця хвиля успіху — невідомо. Хоч це не увійшло до цієї книжки, але наразі немає сумнівів у тому, що вибух інтелектуальності невідворотний, як і в тому, що швидкість його наближення дуже недооцінена. Цілком імовірно, що незабаром, коли ми вичерпаємо запас масштабування та мікроналаштування параметрів, зростання втратить стрімкість. Щоб досягти суперінтелектуальності, потрібні нові відкриття та нові ідеї, а їх спланувати неможливо. Але я не вірю, що знову настане «зима ШІ», принаймні — не така сувора. Досить багато сил вкладено, щоб сфера ШІ стала достойним напрямом як теоретичних досліджень, так і практичних розробок. Тому, на мою думку, фінансування галузі та увага до неї зростатимуть далі. (Не виключено, що окремі проекти не справдять очікувань інвесторів і перетворяться на інвестиційні бульбашки, які згодом луснуть).

Крім того, є прогрес у розумінні, як підвищити ймовірність сприятливого результату. Теоретики пропонують варті подальших

досліджень ідеї, наприклад, ідея Пола Крістіано про «агентів, що керуються схваленням». На організаційному рівні справи також ідуть краще — власне, якщо врахувати, що початковий рівень був аж надто низьким, покращення просто приголомшливі. Дедалі більше наукових спільнот усвідомлюють важливість проблеми контролю та її вплив на майбутнє штучного інтелекту.

Але не варто перебільшувати прогрес у сфері ШІ. Те, що так багато було зроблено за такий короткий час, безперечно, вражає. Значно зросли інвестиції, але лише на покращення апаратного забезпечення щорічно йдуть удвічі-втричі вищі суми. Так, люди дедалі частіше задумуються про наслідки прогресу у сфері ШІ, але їхній погляд рідко коли сягає далі появи автономної смертельної зброї, впливів на ринки праці, кіберзлочинності, впливу на приватність чи появи самокерованих автомобілів. Про ці речі думати потрібно, але це не те, що має нас турбувати, коли ми говоримо про появу ШІ людського рівня чи суперінтелекту.

Варто визнати, що наукові спільноти, пов'язані з дослідженнями ШІ, починають усвідомлювати довгостроковіші наслідки своєї роботи. Але це усвідомлення переважно часткове та хистке. На думку деяких науковців, публічна дискусія з цього питання виходить з-під контролю. Дається взнаки вся ця безглузда термінаторщина. Зрозуміло, кому сподобається, коли його наукову діяльність, колег, зрештою, справу життя, беруть на глум. У відповідь на занепокоєння уявними арміями агресивних роботів, які ввижаються ошуканій публіці, наукова спільнота може зайняти позицію висміювання будь-якої реальної загрози від штучного інтелекту. Поширенню такої позиції можна протидіяти, здійнявши галас у медіа, тоді підвищена увага мас може навіть допомогти. Але турбує побічна реакція. Якщо через переслідування науковці почнуть уникати публічних обговорень суперінтелекту та його потенційних ризиків, щоб не давати критикам і безумцям приводу для цькування, усі спроби ідентифікувати й відвернути майбутні небезпеки ШІ виявляться марними. Тоді почнеться період, який можна назвати «зимою безпечного ШІ», клімат радикально несприятливий для всього, що я пропоную в цій книжці. Важливо запобігти такому протистоянню. Для всіх буде краще, якщо науковці, які займаються створенням ШІ, та науковці, які

опікуватимуться його безпекою, будуть заодно — власне, якщо вони значною мірою будуть тими самими людьми. Тому закликаю всіх бути терплячими, стриманими, відкритими та готовими до діалогу і співпраці.

Я утримаюся від спокуси відповісти на всі коментарі, які отримав з моменту виходу книжки. Зауважу лиш для читачів цього видання, а особливо для тих, хто давно вже не має часу, власне, читати всі книжки, які купує, — хіба що прогляне зміст та передмову: на кількість сторінок, які відведено на розгляд питання, крім важливості питання, впливає ще ціла низка факторів. Тому не варто, підрахувавши сторінки, робити висновок про мої переконання. Зокрема, я зосереджуюся на ризиках більше, ніж на перевагах. Але це не значить, що я заперечуватиму масштаб переваг, які може принести суперінтелект людству. Лиш вважаю, що як би не було корисно помріяти про астрономічну кількість чудових речей, які міг би дати нам суперінтелект, важливіше наразі точно розуміти, що конкретно може піти не так, — щоб упевнено відвернути від себе цю перспективу. Подібно я відводжу багато сторінок аналізу сценаріїв, у яких один суперінтелектуальний ШІ дістає можливість формувати майбутнє у глобальних масштабах, не означає, що я відкидаю імовірність багатополярних сценаріїв (наприклад, див. розділ 11). Я присвятив багато часу аргументації на користь того, що вирішити проблему контролю буде дуже важко, але не виключаю, що розв'язок може виявитися простим.

Ще раз хочу подякувати всім, хто допоміг мені в написанні цієї книжки, усім, хто підтримав її своїми відгуками, а також тим, хто намагається займати конструктивну позицію в цій незвичній та небезпечній пригоді, у якій опинилося людство.

Нік Бостром
Жовтень 2015 року

ПОДЯКИ

Під час написання цієї книжки ідеї, що народжувалися в моїй голові, часто проникали назовні. Деякі з них стали предметом для обговорень у ширшому колі колег. Саме тоді багато ідей потрапили в текст із зовнішніх джерел. Я намагався дуже ретельно відстежувати цитування і вказувати джерела, але впливів було так багато, що вказати всі я не мав змоги.

Висловлюю вдячність тим багатьом людям, з якими я розмовляв у процесі писання книжки і які допомогли мені прояснити думки. Серед них Сем Альтман, Даріо Амодей, Росс Андерсен, Стюарт Армстронг, Оуен Коттон-Барратт, Нік Бекстед, Йошуа Бенджо, Девід Чалмерс, Пол Крістіано, Мілан Чіркович, Ендрю Крітч, Даніель Деннет, Давид Дюч, Даніель Дьюї, Томас Діттеріх, Ерік Дрекслер, Девід Дувенауд, Пітер Екерслі, Амнон Еден, Орен Етціоні, Оуайн Еванс, Бенжа Фалленштейн, Алекс Флінт, Карл Фрей, Зуубін Гахрамані, Ян Голдін, Катя Ґрейс, Роджер Гроссе, Том Гюнтер, Джей Сторрс Голл, Робін Хансон, Деміс Хасабіс, Джеффрі Хінтон, Джеймс Г'юз, Маркус Гаттер, Гаррі Каспаров, Марцін Кульчицький, Патрік Лавіч, Шейн Легг, Моше Візир, Віллам Макаскілл, Ерік Мандельбаум, Гаррі Маркус, Джеймс Мартін, Ліліан Мартін, Роко Міітік, Вінсент Мюллер, Ілон Маск, Шон О'Геґертіг, Крістофер Олах, Тобі Орд, Лоран Орсо, Майкл Осборн, Ларрі Пейдж, Денніс Памлін, Дерек Парфіт, Девід Пірс, Г'ю Прайс, Гай Равін, Мартін Різ, Білл Роско, Франческа Россі, Стюарт Расселл, Анна Саламон, Лу Салкінд, Андерс Сандберг, Джуліан Савелеску, Юрген Шмідгубер, Барт Сельман, Ніколас Шекель, Мюррей Шанахан, Ноель Шаркі, Карл Шульман, Пітер Сінгер, Нейт Соарес, Ден Стойческу, Мустафа Сулейман, Яан Таллінн, Олександр Тамас, Джессіка Тейлор, Макс Теґмарк, Роман Ямпольський та Еліезер Юдковський.

За особливо детальні коментарі до цього тексту дякую Міланові Чірковічу, Даніелю Дьюї, Оуайну Евансу, Ніку Гаю, Кейт Менсфілд,

Люку Муелю Гаузеру, Тобі Орду, Джесс Рідель, Андерсу Сандбергу, Мюрреєві Шанахану і Карлу Шульману. За поради й допомогу під час дослідження вдячний Стюарту Армстронгу, Деніелу Д'юі, Еріку Дрекслеру, Александрові Ерлеру, Ребецці Рош й Андерсу Сендбергу.

За допомогу в підготуванні рукопису дякую Калебу Белу, Мало Бургону, Робіну Брандту, Ленсові Бушу, Кеті Дуглас, Олександру Ерлеру, Джону Кінгу, Крістіану Ренну, Сьюзан Роджерс, Кайлові Скотту, Ендрю Снайдеру Бітті, Сесілії Тіллі та Алексу Вермеєру. Особлива моя подяка Кейт Менсфілд, яка дуже підбадьорювала мене під час роботи над проектом.

Перепрошую тих, кого мав згадати, але не згадав у цій частині книжки.

Найщиріше дякую спонсорам, друзям і своїй родині: без вашої підтримки ця праця ніколи не з'явилася б.

СЛОВНИК ТЕРМІНІВ І ПОНЯТЬ

Антропічне захоплення — гіпотетична ситуація, у якій ШІ думає, що перебуває в симульованому середовищі й намагається поводитися так, щоб його поведінка задовольняла гіпотетичних симуляторів.

Апаратне забезпечення рівня людини — апаратне забезпечення з можливостями обробки інформації, подібними до можливостей людського мозку.

Асоціативне накопичення цінностей — спосіб визначення цінностей ШІ, який полягає в поступовому засвоєнні штучним інтелектом цінностей на основі особистого життєвого досвіду. (Цей механізм подібний до того, як засвоюють життєві цінності люди).

Багатополярний сценарій — сценарій настання ери ШІ, унаслідок якого з'являються кілька конкурентних суперінтелектів.

Безосібна перспектива — перспектива оцінки відповідності деякої дії загальним інтересам усіх зацікавлених сторін — включно з тими, чиє існування буде зумовлене вибором (пор. «Особистісна перспектива»).

Взаємозалежність технологій — ступінь визначеності причинно-наслідкового зв'язку між гіпотетичними технологіями, за якого можливо з певною точністю передбачити часові межі їхньої взаємної появи: коли одна є необхідною умовою іншої або безальтернативним способом її застосування. Наприклад, нейроморфний ШІ взаємозалежний з емуляцією цілого мозку, бо технологія, потрібна для створення емуляції, ще на ранніх етапах свого розвитку може уможливити створення ШІ, подібного до мозку (і для цього існуватимуть усі передумови).

Вибухове зростання інтелектуальності — гіпотетичний процес, протягом якого розумові здібності ШІ швидко зростають від «порівняно скромних» до таких, що значно перевищують можливості інтелекту людини (зазвичай вважають, що це відбудеться внаслідок рекурсивного самовдосконалення).

Визначення цінностей ШІ за допомогою еволюційної селекції — підхід до вирішення проблеми визначення цінностей для ШІ за допомогою реалізації процесу багатократної селекції, аналогічної до природної еволюційної селекції, результатом якої стала поява людини.

Вирішальна стратегічна перевага — стратегічна перевага (технологічна чи інша), яка дає агенту можливість досягти повної влади над світом.

Вирішальні знання — ідея чи аргумент, здатний вирішально вплинути або, можливо, докорінно змінити напрям чи пріоритети діяльності, наприклад, інвертувати знак бажаності

деякого важливого впливу.

Джин — ШІ, призначений для почергового виконання загальних комплексних команд.

Доповнення — спосіб створення суперінтелекту з необхідною мотивацією, узявши за основу систему з потрібними характеристиками мотивації (наприклад, людину) та збільшивши її інтелектуальну потужність замість формування мотивації з нуля.

Друга проблема принципала-агента — проблема принципала-агента, з якою стикається людина («принципал»), яка бажає створити суперінтелектуальний ШІ («агент») для використання його можливостей у досягненні власних цілей. Також відома як «проблема контролю».

Думкозлочин — неправильне поводження з морально значущими обчислювальними процесами (які існують у симуляціях, є частинами ШІ, створені з інструментальною метою).

Емуляція цілого мозку — штучний інтелект, структура якого детально відтворює людський мозок.

Зерно ШІ — функціональний ШІ з базовими можливостями, який може самостійно вдосконалювати та доповнювати свою структуру.

Колективний суперінтелект — інтелектуальна система, яка складається з менших інтелектів, так що її комплексні розумові можливості в багатьох узагальнених сферах діяльності значно перевищують будь-яку сучасну інтелектуальну систему.

Консервативність — внутрішній опір системи вдосконаленням.

Контроль здібностей — обмеження можливостей ШІ, щоб уникнути небажаних наслідків його діяльності.

Космічні ресурси людства — обсяг фізичних ресурсів у Всесвіті, досяжних для технологічно зрілої земної цивілізації (починаючи від сучасної Землі).

Макроструктурний акселератор розвитку — уявний важіль (на кшталт тих, що використовують під час розумових експериментів), який би змінював швидкість перебігу макроструктурних процесів (як-от технологічний розвиток чи геополітика), водночас не впливаючи на темп життя та діяльність пересічних людей.

Методи заохочення — методи контролю ШІ, які передбачають формування середовища діяльності ШІ так, що навіть за відсутності кінцевих цілей ШІ завжди має інструментальні причини дій.

Модуляція емуляції — формування мотивації емуляції мозку за допомогою цифрових аналогів хімічних препаратів або інших засобів.

Навчання цінностей — спосіб завантаження цінностей у ШІ через процес навчання.

Непряма нормативність — підхід до визначення цінностей через встановлення певних критеріїв чи методик, за допомогою яких ШІ, використовуючи власні інтелектуальні можливості, може визначити конкретні цінності.

Обмеження — створення для ШІ спеціального середовища, в якому реалізовані засоби обмеження взаємодії зі зовнішнім світом. Наприклад, запуск ШІ в ізольованій симуляції

реальності з можливістю спілкуватися лише з людьми, які наглядають за його роботою.

Одомашнення — вирішення проблеми контролю ШІ за допомогою обмеження амбіцій, сфер зацікавлення чи втручання.

Оптимізаційна сила — міра якісно зваженого творчого зусилля, спрямованого на збільшення інтелектуальності системи.

Оракул — ШІ, функція якого лише відповідати на питання.

Особистісна перспектива — перспектива оцінки дій з погляду інтересів усіх, хто вже існує або існуватиме, незалежно від результату вибору (пор. «Безосібна перспектива»).

Перша проблема принципала-агента — широко відома управлінська проблема, коли одна людина («принципал») делегує іншій («агенту») повноваження діяти у своїх («принципала») інтересах. Наприклад, робітник та роботодавець.

Підхід до вирішення проблеми контролю через навчання з підкріпленням — коли ШІ самостійно навчається накопичувати нагороду певного виду (представлену як сигнал, який визначає або безпосередньо подає людина, та призначену для того, щоб стимулювати ШІ до потрібних дій).

Поріг стійкості поміркованого синґлтону — мінімальний набір здібностей, з яким цілеспрямована та завбачлива інтелектуальна система за умови відсутності інтелектуальної протидії може успішно колонізувати й контролювати велику частину доступного Всесвіту.

Приваблива обгортка — вказівки, приєднані до цілей ШІ, які визначають окрему винагороду чи можливість впливати на діяльність ШІ для тих, хто долучився до його створення.

Пригнічення — метод контролю ШІ, що полягає в обмеженні доступу до інформації чи встановленні фізичних обмежень когнітивного функціоналу.

Принцип диференційного технологічного розвитку — уповільнювати розвиток небезпечних і шкідливих технологій, які мають більший рівень екзистенційного ризику; пришвидшувати розвиток корисних технологій, особливо тих, які знижують екзистенційні ризики — природні чи техногенні.

Принцип епістемічної переваги — коли з’явиться суперінтелект, він буде епістемічно досконаліший: його судження (ймовірно, у більшості випадків) буде більш правильним. Тому ми повинні намагатися віддавати перевагу судженням суперінтелекту.

Принцип загального блага — суперінтелект має створюватися на користь усьому людству та керуватися у своїй діяльності загальноприйнятими етичними нормами.

Припущення про технологічну завершеність — якщо зусилля, які зумовлюють науковий та технологічний розвиток, не припиняться, усі ключові можливості деякої майбутньої технології будуть досягнуті.

Проблема прищеплення цінностей — завдання встановлення ШІ кінцевої цілі, яка полягає в слідуванні важливим для нас цінностям.

Програмне забезпечення рівня людини — програмне забезпечення, яке може з подібною до людського мозку алгоритмічною ефективністю виконувати характерні для людського мозку

завдання.

Прямі вимоги — підхід до вирішення проблеми контролю, який полягає в тому, що програмісти визначають, що людям потрібно від ШІ, а потім пишуть код, у якому прямо визнають відповідні цінності чи правила функціонування.

Рекурсивне самовдосконалення — процес ітеративного вдосконалення ШІ власного інтелекту, за якого більша потужність інтелекту означає прикладання більшої оптимізаційної сили на кожному наступному етапі самовдосконалення.

Ризики дії — рівень ризику, притаманний певній діяльності. Водночас цей рівень не є простою функцією тривалості дії. Наприклад, небезпека руху мінним полем не менша, якщо рухатися швидше.

Ризики стану — ризики, зумовлені певним станом, водночас рівень ризику пропорційний часу перебування в цьому стані. Наприклад, вразливість Землі до падіння астероїдів створює для нас ризик, рівень якого пропорційний до тривалості нашого перебування в стані вразливості.

Розумове покращення — покращення інтелектуальних здібностей системи.

Синґлтон — світовий порядок, у якому найвищий щабель ухвалення рішень посідає одна єдина сила (яка, однак, може мати складну внутрішню організацію), за умови, що всі основні проблеми глобальної координації вирішені. Прикладами може бути світова демократія, всесвітня диктатура одного диктатора або суперінтелектуальний ШІ, достатньо потужний, щоб подолати всіх потенційних супротивників.

Стрибок — перехід від існування штучного інтелекту рівня людини до радикально суперінтелектуального ШІ. Зазвичай розглядають імовірну швидкість стрибка: «повільний стрибок» триватиме від десятиліть до століття, «помірний стрибок» — від місяців до років, «швидкий стрибок» — може тривати дні або й менше.

Суверен — ШІ, який діє автономно для досягнення широкого спектра довгострокових цілей.

Суперінтелект — інтелектуальна система будь-якої природи, розумові здібності якої значно перевищують здібності людей у всіх важливих сферах діяльності.

Теза інструментальної конвергентності — існують «конвергентні інструментальні цілі» — проміжні цілі, які можуть бути однаково корисними в низці сценаріїв для досягнення будь-якої з імовірних кінцевих цілей у різних імовірних ситуаціях та середовищах для інтелектуальних агентів різного типу.

Теза ортогональності — рівень інтелекту та кінцеві цілі ортогональні: інтелект будь-якого рівня може мати будь-які кінцеві цілі.

Хибна реалізація — ефективний спосіб досягнення кінцевої мети ШІ у спосіб, не передбачений та небажаний для програмістів, які визначали критерій мети (наприклад, «примусити людей сміятися», паралізувавши їхні мімічні м’язи у гримасі, що дуже схожа на усмішку).

Шаблон мотивації — спосіб установлення цінностей ШІ, коли спочатку визначаються прості цілі, а згодом, коли ШІ розвиває в собі здатність сприймати складні поняття і представлення, вони заміняються на складніші — ціннісні — цілі.

Швидкий суперінтелект — інтелектуальна система, яка може робити все те, що може робити пересічна людина, але значно швидше.

ШІ-інструмент — ШІ, який більше схожий не на повноцінного агента, а на гнучкішу та потужнішу комп'ютерну програму. Зокрема, вона не має якоїсь визначеної мети.

ШІ моральної правоти (ШІ-МП) — ШІ, який прагне робити морально правильні вчинки.

ШІ рівня людини — ШІ, здібності якого у всіх важливих сферах діяльності приблизно відповідають здібностям пересічної людини (трактування цього терміна неоднозначне).

Якісний суперінтелект — інтелектуальна система, така сама швидка, як і розум пересічної людини, але якісно значно розумніша.

Цей словник містить лише найважливіші терміни, використані в книжці. Не всі визначення, наведені тут, вичерпно характеризують поняття. Для повного розуміння завжди краще віддавати перевагу поясненню, наведеному в основному тексті. (За першу версію цього словника та за право вносити свої корективи я вдячний Стефані Золайвар та Каті Ґрейс з AI Impacts та Machine Intelligence Research Institute. Оригінал можна знайти за адресою: aiimpacts.org/ai-risk-terminology/).

БІБЛІОГРАФІЯ

Acemoglu, Daron. 2003. "Labor- and Capital-Augmenting Technical Change." *Journal of the European Economic Association* 1 (1): 1—37.

Albertson, D. G., and Thomson, J. N. 1976. "The Pharynx of *Caenorhabditis Elegans*." *Philosophical Transactions of the Royal Society B: Biological Sciences* 275 (938): 299—325.

Allen, Robert C. 2008. "A Review of Gregory Clark's A Farewell to Alms: A Brief Economic History of the World." *Journal of Economic Literature* 46 (4): 946—73.

American Horse Council. 2005. "National Economic Impact of the US Horse Industry." Retrieved July 30, 2013. Available at horsecouncil.org/national-economic-impact-us-horse-industry.

Anand, Paul, Pattanaik, Prasanta, and Puppe, Clemens, eds. 2009. *The Oxford Handbook of Rational and Social Choice*. New York: Oxford University Press.

Andres, B., Koethe, U., Kroeger, T., Helmstaedter, M., Briggman, K. L., Denk, W., and Hamprecht, F. A. 2012. "3D Segmentation of SBFSEM Images of Neuropil by a Graphical Model over Supervoxel Boundaries." *Medical Image Analysis* 16 (4): 796—805.

Armstrong, Alex. 2012. "Computer Competes in Crossword Tournament." *I Programmer*, March 19.

Armstrong, Stuart. 2007. "Chaining God: A Qualitative Approach to AI, Trust and Moral Systems." Unpublished manuscript, October 20. Retrieved December 31, 2012. Available at neweuropeancentury.org/GodAI.pdf.

Armstrong, Stuart. 2010. *Utility Indifierence*, Technical Report 2010-1. Oxford: Future of Humanity Institute, University of Oxford.

Armstrong, Stuart. 2013. "General Purpose Intelligence: Arguing the Orthogonality Thesis." *Analysis and Metaphysics* 12: 68—84.

Armstrong, Stuart, and Sandberg, Anders. 2013. "Eternity in Six Hours: Intergalactic Spreading of Intelligent Life and Sharpening the Fermi Paradox." *Acta Astronautica* 89: 1—13.

Armstrong, Stuart, and Sotala, Kaj. 2012. "How We're Predicting AI—or Failing To." In *Beyond AI: Artificial Dreams*, edited by Ian Romportl, Pavel Ircing, Eva Zackova, Michal Polak, and Radek Schuster, 52—75. Pilsen: University of West Bohemia. Retrieved February 2, 2013.

Asimov, Isaac. 1942. "Runaround." *Astounding Science Fiction*, March, 94—103.

Asimov, Isaac. 1985. *Robots and Empire*. New York: Doubleday.

Aumann, Robert I. 1976. "Agreeing to Disagree." *Annals of Statistics* 4 (6): 1236—9.

Averch, Harvey Allen. 1985. *A Strategic Analysis of Science and Technology Policy*. Baltimore: Johns Hopkins University Press.

Azevedo, F. A. C., Carvalho, L. R. B., Grinberg, L. T., Farfel, I. M., Ferretti, R. E. L., Leite, R. E. P., Iacob, W, Lent, R., and Herculano-Houzel, S. 2009. "Equal Numbers of Neuronal and Nonneuronal Cells Make the Human Brain an Isometrically Scaled-up Primate Brain." *Journal of Comparative Neurology* 513 (5): 532—41.

Baars, Bernard I. 1997. *In the Theater of Consciousness: The Workspace of the Mind*. New York: Oxford University Press.

Baratta, Ioseph Preston. 2004. *The Politics of World Federation: United Nations, UN Reform, Atomic Control*. Westport, CT: Praeger.

Barber, E. J. W. 1991. *Prehistoric Textiles: The Development of Cloth in the Neolithic and Bronze Ages with Special Reference to the Aegean*. Princeton, NJ: Princeton University Press.

Bartels, J., Andreasen, D., Ehirim, P., Mao, H., Seibert, S., Wright, E. J., and Kennedy, P. 2008. "Neurotrophic Electrode: Method of Assembly and Implantation into Human Motor Speech Cortex." *Journal of Neuroscience Methods* 174 (2): 168—76.

Bartz, Jennifer A., Zaki, Iamil, Bolger, Niall, and Ochsner, Kevin N. 2011. "Social Effects of Oxytocin in Humans: Context and Person Matter." *Trends in Cognitive Science* 15 (7): 301—9.

Basten, Stuart, Lutz, Wolfgang, and Scherbov, Sergei. 2013. "Very Long Range Global Population Scenarios to 2300 and the Implications of Sustained Low Fertility." *Demographic Research* 28: 1 145—66.

Baum, Eric B. 2004. *What Is Thought?* Bradford Books. Cambridge, MA: MIT Press.

Baum, Seth D., Goertzel, Ben, and Goertzel, Ted G. 2011. "How Long Until Human—Level AI? Results from an Expert Assessment." *Technological Forecasting and Social Change* 78 (1): 185—95.

Beal, I., and Winston, P. 2009. "Guest Editors' Introduction: The New Frontier of Human-Level Artificial Intelligence." *IEEE Intelligent Systems* 24 (4): 21—3.

Bell, C. Gordon, and Gemmell, Iim. 2009. *Total Recall: How the E-Memory Revolution Will Change Everything*. New York: Dutton.

Benyamin, B., St. Pourcain, B., Davis, 0. S., Davies, G., Hansell, M. K., Brion, M.-I. A., Kirkpatrick, R. M., et al. 2013. "Childhood Intelligence is Heritable, Highly Polygenic and Associated With FNBP1L." *Molecular Psychiatry* (January 23).

Berg, once E., and Rietz, Thomas A. 2003. "Prediction Markets as Decision Support Systems." *Information Systems Frontiers* 5 (1): 79—93.

Berger, Theodore W, Chapin, I. K., Gerhardt, G. A., Soussou, W. V., Taylor, D. M., and Tresco, P. A., eds. 2008. *Brain—Computer Interfaces: An International Assessment of Research and Development Trends*. Springer.

Berger, T. W., Song, D., Chan, R. H., Marmarelis, V. Z., LaCoss, I., Wills, I., Hampson, R. E., Deadwyler, S. A., and Granacki, I. I. 2012. "A Hippocampal Cognitive Prosthesis: Multi-Input, Multi-Output Nonlinear Modeling and VLSI Implementation." *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 20(2): 198—211.

Berliner, Hans I. 1980a. "Backgammon Computer-Program Beats World Champion." *Artificial Intelligence* 14 (2): 205—220.

Berliner, Hans I. 1980b. "Backgammon Program Beats World Champ." *SIGART Newsletter* 69: 6—9.

Bernardo, Iosé M., and Smith, Adrian F. M. 1994. *Bayesian Theory*, 1st ed. Wiley Series in Probability 81 Statistics. New York: Wiley.

Birbaumer, N., Murguialday, A. R., and Cohen, L. 2008. "Brain—Computer Interface in Paralysis." *Current Opinion in Neurology* 21 (6): 634—8.

Bird, Ion, and Layzell, Paul. 2002. "The Evolved Radio and Its Implications for Modelling the Evolution of Novel Sensors." In *Proceedings of the 2002 Congress on Evolutionary Computation*, 2: 1836—41.

Blair, Clay, Ir. 1957. "Passing of a Great Mind: John von Neumann, a Brilliant, Jovial Mathematician, was a Prodigious Servant of Science and His Country" *Life*, February 25, 89— 104.

Bobrow, Daniel G. 1968. "Natural Language Input for a Computer Problem Solving System." In *Semantic Information Processing*, edited by Marvin Minsky, 146—227. Cambridge, MA: MIT Press.

Bostrom, Nick. 1997. "Predictions from Philosophy? How Philosophers Could Make Themselves Useful." Unpublished manuscript. Last revised September 19, 1998.

Bostrom, Nick. 2002a. *Anthropic Bias: Observation Selection Eflects in Science and Philosophy*. New York: Routledge.

Bostrom, Nick. 2002b. "Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards." *Journal of Evolution and Technology* 9.

Bostrom, Nick. 2003a. "Are We Living in a Computer Simulation?" *Philosophical Quarterly* 53 (211): 243—55.

Bostrom, Nick. 2003b. "Astronomical Waste: The Opportunity Cost of Delayed Technological Development." *Utilitas* 15 (3): 308—314.

Bostrom, Nick. 2003c. "Ethical Issues in Advanced Artificial Intelligence." In *Cognitive, Emotive and Ethical Aspects of Decision Making in Humans and in Artificial Intelligence*, edited by Iva Smit and George E. Lasker, 2: 12—17. Windsor, ON: International Institute for Advanced Studies in Systems Research / Cybernetics.

Bostrom, Nick. 2004. "The Future of Human Evolution." In *Two Hundred Years After Kant, Fifty Years After Turing*, edited by Charles Tandy, 2: 339—371. Death and Anti—Death. Palo Alto, CA: Ria University Press.

Bostrom, Nick. 2006a. "How Long Before Superintelligence?" *Linguistic and Philosophical Investigations* 5(1): 11—30.

Bostrom, Nick. 2006b. "Quantity of Experience: Brain-Duplication and Degrees of Consciousness." *Minds and Machines* 16(2): 185—200.

Bostrom, Nick. 2006c. "What is a Singleton?" *Linguistic and Philosophical Investigations* 5 (2): 43—54.

Bostrom, Nick. 2007. "Technological Revolutions: Ethics and Policy in the Dark." In *Nanoscale: Issues and Perspectives for the Nano Century*, edited by Nigel M. de S. Cameron and M. Ellen Mitchell, 129

—52. Hoboken, NI: Wiley.

Bostrom, Nick. 2008a. "Where Are They? Why I Hope the Search for Extraterrestrial Life Finds Nothing." *MIT Technology Review*, May/June issue, 72—7.

Bostrom, Nick. 2008b. "Why I Want to Be a Posthuman When I Grow Up." In *Medical Enhancement and Posthumanity*, edited by Bert Gordijn and Ruth Chadwick, 107—37. New York: Springer.

Bostrom, Nick. 2008c. "Letter from Utopia." *Studies in Ethics, Law, and Technology* 2 (1): 1—7.

Bostrom, Nick. 2009a. "Moral Uncertainty — Towards a Solution?" *Overcoming Bias* (blog), January 1.

Bostrom, Nick. 2009b. "Pascal's Mugging." *Analysis* 69 (3): 443—5.

Bostrom, Nick. 2009c. "The Future of Humanity." In *New Waves in Philosophy of Technology*, edited by Jan Kyrre Berg Olsen, Evan Selinger, and Soren Riis, 186—215. New York: Palgrave Macmillan.

Bostrom, Nick. 2011a. "Information Hazards: A Typology of Potential Harms from Knowledge." *Review of Contemporary Philosophy* 10: 44—79.

Bostrom, Nick. 2011b. "Infinite Ethics." *Analysis and Metaphysics* 10: 9—59.

Bostrom, Nick. 2012. "The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents." In "Theory and Philosophy of AI," edited by Vincent C. Miiller, special issue, *Minds and Machines* 22 (2): 71—85.

Bostrom, Nick, and Ćirković, Milan M. 2003. "The Doomsday Argument and the Self- Indication Assumption: Reply to Olum." *Philosophical Quarterly* 53 (210): 83—91.

Bostrom, Nick, and Ord, Toby. 2006. "The Reversal Test: Eliminating the Status Quo Bias in Applied Ethics." *Ethics* 1 16 (4): 656—79.

Bostrom, Nick, and Roache, Rebecca. 2011. "Smart Policy: Cognitive Enhancement and the Public Interest." In *Enhancing Human Capacities*, edited by Iulian Savulescu, Ruud ter Meulen, and Guy Kahane, 138—49. Malden, MA: Wiley-Blackwell.

Bostrom, Nick and Sandberg, Anders. 2009a. "Cognitive Enhancement: Methods, Ethics, Regulatory Challenges." *Science and Engineering Ethics* 15 (3): 31 1—41.

Bostrom, Nick and Sandberg, Anders. 2009b. "The Wisdom of Nature: An Evolutionary Heuristic for Human Enhancement." In *Human Enhancement*, 1st ed., edited by Julian Savulescu and Nick Bostrom, 375—416. New York: Oxford University Press.

Bostrom, Nick, Sandberg, Anders, and Douglas, Tom. Forthcoming. "The Unilateralist's Curse: The Case for a Principle of Conformity." *Social Epistemology*.

Bostrom, Nick, and Yudkowsky, Eliezer. 2015. "The Ethics of Artificial Intelligence." In *Cambridge Handbook of Artificial Intelligence*, edited by Keith Frankish and William M. Ramsey, 315—334. New York: Cambridge University Press.

Boswell, Iames. 1917. *Boswell's Life of Johnson*. New York: Oxford University Press.

Bouchard, T. I. 2004. "Genetic Influence on Human Psychological Traits: A Survey." *Current Directions in Psychological Science* 13 (4): 148—51.

Bourget, David, and Chalmers, David. 2009. "The PhilPapers Surveys." November. Available at philpapers.org/surveys/ .

Bradbury, Robert I. 1999. "Matrioshka Brains." Archived version. As revised August 16, 2004. Available at web.archive.org/web/20090615040912/http://www.aeiveos.com/ ~bradbury/ Matrio shkaBrains/MatrioshkaBrainsPaper.html.

Brinton, Crane. 1965. *The Anatomy of Revolution*. Revised ed. New York: Vintage Books.

Bryson, Arthur E., Jr., and Ho, Yu-Chi. 1969. *Applied Optimal Control: Optimization, Estimation, and Control*. Waltham, MA: Blaisdell.

Buehler, Martin, Iagnemma, Karl, and Singh, Sanjiv, eds. 2009. *The DARPA Urban Challenge: Autonomous Vehicles in City Traflic*. Springer Tracts in Advanced Robotics 56. Berlin: Springer.

Burch-Brown, I. 2014. "Clues for Consequentialists." *Utilitas* 26 (1): 105—19.

Burke, Colin. 2001. "Agnes Meyer Driscoll vs. the Enigma and the Bombe." Unpublished manuscript. Retrieved February 22, 2013. Available at userpages.umbc. edu/ ~burke/ driscoll1-201 1.pdf.

Canback, S., Samouel, P., and Price, D. 2006. "Do Diseconomies of Scale Impact Firm Size and Performance? A Theoretical and Empirical Overview." *Journal of Managerial Economics* 4 (1): 27—70.

Carmena, J. M., Lebedev, M. A., Crist, R. E., O'Doherty, J. E., Santucci, D. M., Dimitrov, D. F., Patil, P. G., Henriquez, C. S., and Nicolelis, M. A. 2003. "Learning to Control a Brain—Machine Interface for Reaching and Grasping by Primates." *Public Library of Science Biology* 1 (2): 193—208.

Carroll, Bradley W, and Ostlie, Dale A. 2007. *An Introduction to Modern Astrophysics*. 2nd ed. San Francisco, CA: Pearson Addison Wesley.

Carroll, John B. 1993. *Human Cognitive Abilities: A Survey of Factor-Analytic Studies*. New York: Cambridge University Press.

Carter, Brandon. 1983."The Anthropic Principle and its Implications for Biological Evolution." *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 310 (1512): 347—63.

Carter, Brandon. 1993. "The Anthropic Selection Principle and the Ultra-Darwinian Synthesis." In *The Anthropic Principle: Proceedings of the Second Venice Conference on Cosmology and Philosophy*, edited by F. Bertola and U. Curi, 33—66. Cambridge: Cambridge University Press.

CFTC 81 SEC (Commodity Futures Trading Commission and Securities 81 Exchange Commission). 2010. *Findings Regarding the Market Events of May 6, 2010: Report of the Stafls of the CFTC and SEC to the Joint Advisory Committee on Emerging Regulatory Issues*. Washington, DC.

Chalmers, David John. 2010. "The Singularity: A Philosophical Analysis." *Journal of Consciousness Studies* 17 (9—10): 7—65.

Chason, R. J., Csokmay, J., Segars, J. H., DeCherney, A. H., and Armant, D. R. 2011. "Environmental and Epigenetic Effects Upon Preimplantation Embryo Metabolism and Development." *Trends in Endocrinology and Metabolism* 22 (10): 412—20.

Chen, S., and Ravallion, M. 2010. "The Developing World Is Poorer Than We Thought, But No Less Successful in the Fight Against Poverty." *Quarterly Journal of Economics* 125 (4): 1577—1625.

Chislenko, Alexander. 1996. “Networking in the Mind Age: Some Thoughts on Evolution of Robotics and Distributed Systems.” Unpublished manuscript.

Chislenko, Alexander. 1997. “Technology as Extension of Human Functional Architecture.” *Extropy Online*.

Chorost, Michael. 2005. *Rebuilt: How Becoming Part Computer Made Me More Human*. Boston: Houghton Mifflin.

Christiano, Paul F. 2012. “‘Indirect Normativity’ Write—up.” *Ordinary Ideas* (blog), April 21.

CIA. 2013. *The World Factbook*. Central Intelligence Agency. Retrieved August 3. Available at cia.gov/library/publications/the-world—factbook/rankorder/ 2127rank.html?countryname=United%20States&countrycode=us&regionCode=noa&rank=121#us.

Cicero. 1923. “On Divination.” In *On Old Age, on Friendship, on Divination*, translated by W. A. Falconer. Loeb Classical Library. Cambridge, MA: Harvard University Press.

Cirasella, Jill, and Kopec, Danny. 2006. “The History of Computer Games.” Exhibit at Dartmouth Artificial Intelligence Conference: The Next Fifty Years (AI@50), Dartmouth College, July 13—15.

Ćirković, Milan M. 2004. “Forecast for the Next Eon: Applied Cosmology and the Long- Term Fate of Intelligent Beings.” *Foundations of Physics* 34 (2): 239—61.

Ćirković, Milan M., Sandberg, Anders, and Bostrom, Nick. 2010. “Anthropic Shadow: Observation Selection Effects and Human Extinction Risks.” *Risk Analysis* 30 (10): 1495—1506.

Clark, Andy, and Chalmers, David J. 1998. “The Extended Mind.” *Analysis* 58 (1): 7—19.

Clark, Gregory. 2007. *A Farewell to Alms: A Brief Economic History of the World*. 1st ed. Princeton, NJ: Princeton University Press.

Clavin, Whitney. 2012. “Study Shows Our Galaxy Has at Least 100 Billion Planets.” *Jet Propulsion Laboratory*, January 1 1.

CME Group. 2010. *What Happened on May 6th?* Chicago, May 10.

Coase, R. H. 1937. “The Nature of the Firm.” *Economica* 4 (16): 386—405.

Cochran, Gregory, and Harpending, Henry. 2009. *The 10,000 Year Explosion: How Civilization Accelerated Human Evolution*. New York: Basic Books.

Cochran, G., Hardy, J., and Harpending, H. 2006. “Natural History of Ashkenazi Intelligence.” *Journal of Biosocial Science* 38 (5): 659—93.

Cook, James Gordon. 1984. *Handbook of Textile Fibres: Natural Fibres*. Cambridge: Woodhead.

Cope, David. 1996. *Experiments in Musical Intelligence. Computer Music and Digital Audio Series*. Madison, WI: A-R Editions.

Cotman, Carl W, and Berchtold, Nicole C. 2002. “Exercise: A Behavioral Intervention to Enhance Brain Health and Plasticity.” *Trends in Neurosciences* 25 (6): 295—301.

Cowan, Nelson. 2001. “The Magical Number 4 in Short—Term Memory: A Reconsideration of Mental Storage Capacity.” *Behavioral and Brain Sciences* 24 (1): 87—114.

Crabtree, Steve. 1999. “New Poll Gauges Americans’ General Knowledge Levels.” *Gallup News*, July 6.

Cross, Stephen E., and Walker, Edward. 1994. "Dart: Applying Knowledge Based Planning and Scheduling to Crisis Action Planning." In *Intelligent Scheduling*, edited by Monte Zweben and Mark Fox, 711—29. San Francisco, CA: Morgan Kaufmann.

Crow, James F. 2000. "The Origins, Patterns and Implications of Human Spontaneous Mutation." *Nature Reviews Genetics* 1 (1): 40—7.

Cyranoski, David. 2013. "Stem Cells: Egg Engineers." *Nature* 500 (7463): 392—4.

Dagnelie, Gislin. 2012. "Retinal Implants: Emergence of a Multidisciplinary Field." *Current Opinion in Neurology* 25 (1): 67—75.

Dai, Wei. 2009. "Towards a New Decision Theory." *Less Wrong* (blog), August 13.

Dalrymple, David. 2011. "Comment on Kaufman, I. 'Whole Brain Emulation: Looking at Progress on C. Elegans.' " *Less Wrong* (blog), October 29.

Davies, G., Tenesa, A., Payton, A., Yang, J., Harris, S. E., Liewald, D., Ke, X., et al. 2011. "Genome—Wide Association Studies Establish That Human Intelligence Is Highly Heritable and Polygenic." *Molecular Psychiatry* 16 (10): 996—1005.

Davis, Oliver S. P., Butcher, Lee M., Docherty, Sophia J., Meaburn, Emma L., Curtis, Charles J. C., Simpson, Michael A., Schalkwyk, Leonard C., and Plomin, Robert. 2010. "A Three-Stage Genome-Wide Association Study of General Cognitive Ability: Hunting the Small Effects." *Behavior Genetics* 40 (6): 759—767.

Dawkins, Richard. 1995. *River Out of Eden: A Darwinian View of Life*. Science Masters Series. New York: Basic Books.

De Blanc, Peter. 2011. *Ontological Crises in Artificial Agents' Value Systems*. Berkeley, CA: Machine Intelligence Research Institute, May 19.

De Long, J. Bradford. 1998. "Estimates of World GDP, One Million B.C.—Present." Unpublished manuscript.

De Raedt, Luc, and Flach, Peter, eds. 2001. *Machine Learning: ECML 2001: 12th European Conference on Machine Learning, Freiburg, Germany, September 5—7, 2001. Proceedings*. Lecture Notes in Computer Science 2167. New York: Springer.

Dean, Cornelia. 2005. "Scientific Savvy? In US, Not Much." *New York Times*, August 30.

Deary, Ian J. 2001. "Human Intelligence Differences: A Recent History." *Trends in Cognitive Sciences* 5 (3): 127—30.

Deary, Ian J. 2012. "Intelligence." *Annual Review of Psychology* 63: 453—82.

Deary, Ian J., Penke, L., and Johnson, W. 2010. "The Neuroscience of Human Intelligence Differences." *Nature Reviews Neuroscience* 11 (3): 201— 11.

Degnan, G. G., Wind, T. C., Jones, E. V., and Edlich, R. F. 2002. "Functional Electrical Stimulation in Tetraplegic Patients to Restore Hand Function." *Journal of Long—Term Eflects of Medical Implants* 12 (3): 175—88.

Devlin, B., Daniels, M., and Roeder, K. 1997. "The Heritability of IQ." *Nature* 388 (6641): 468—71.

Dewey, Daniel. 2011. "Learning What to Value." In *Artificial General Intelligence: 4th International Conference, AGI 2011, Mountain View, CA, USA, August 3—6, 2011. Proceedings*, edited by Jürgen Schmidhuber, Kristinn R. Thorisson, and Moshe Looks, 309—14. Lecture Notes in Computer Science 6830. Berlin: Springer.

Dowe, D. L., and Hernandez-Orallo, J. 2012. "IQ Tests Are Not for Machines, Yet." *Intelligence* 40 (2): 77—81.

Drescher, Gary L. 2006. *Good and Real: Demystifying Paradoxes from Physics to Ethics*. Bradford Books. Cambridge, MA: MIT Press.

Drexler, K. Eric. 1986. *Engines of Creation*. Garden City, NY: Anchor.

Drexler, K. Eric. 1992. *Nanosystems: Molecular Machinery, Manufacturing, and Computation*. New York: Wiley.

Drexler, K. Eric. 2013. *Radical Abundance: How a Revolution in Nanotechnology Will Change Civilization*. New York: PublicAffairs.

Driscoll, Kevin. 2012. "Code Critique: 'Altair Music of a Sort.' " Paper presented at Critical Code Studies Working Group Online Conference, February 6.

Dyson, Freeman J. 1960. "Search for Artificial Stellar Sources of Infrared Radiation." *Science* 131 (3414): 1667—1668.

Dyson, Freeman J. 1979. *Disturbing the Universe*. 1st ed. Sloan Foundation Science Series. New York: Harper 81 Row.

Elga, Adam. 2004. "Defeating Dr. Evil with Self-Locating Belief." *Philosophy and Phenomenological Research* 69 (2): 383—96.

Elga, Adam. 2007. "Reflection and Disagreement." *Noûs* 41 (3): 478—502.

Eliasmith, Chris, Stewart, Terrence C., Choo, Xuan, Bekolay, Trevor, DeWolf, Travis, Tang, Yichuan, and Rasmussen, Daniel. 2012. "A Large—Scale Model of the Functioning Brain." *Science* 338(61 11): 1202—5.

Ellis, J. H. 1999. "The History of Non-Secret Encryption." *Cryptologia* 23 (3): 267—73.

Elyasaf, Achiya, Hauptmann, Ami, and Sipper, Moche. 2011. "Ga—Freecell: Evolving Solvers for the Game of Freecell." In *Proceedings of the 13th Annual Genetic and Evolutionary Computation Conference*, 1931— 1938. GECCO '11. New York: ACM.

Eppig, C., Fincher, C. L., and Thornhill, R. 2010. "Parasite Prevalence and the Worldwide Distribution of Cognitive Ability." *Proceedings of the Royal Society B: Biological Sciences* 277 (1701): 3801—8.

Espenshade, T. J., Guzman, J. C., and Westoff, C. F. 2003. "The Surprising Global Variation in Replacement Fertility." *Population Research and Policy Review* 22 (5—6): 575—83.

Evans, Thomas G. 1964. "A Heuristic Program to Solve Geometric—Analogy Problems." In *Proceedings of the April 21—23, 1964, Spring Joint Computer Conference*, 327—338. AFIPS '64. New York: ACM.

Evans, Thomas G. 1968. "A Program for the Solution of a Class of Geometric—Analogy Intelligence-Test Questions." In *Semantic Information Processing*, edited by Marvin Minsky, 271—353.

Cambridge, MA: MIT Press.

Faisal, A. A., Selen, L. P., and Wolpert, D. M. 2008. "Noise in the Nervous System." *Nature Reviews Neuroscience* 9 (4): 292—303.

Faisal, A. A., White, J. A., and Laughlin, S. B. 2005. "Ion—Channel Noise Places Limits on the Miniaturization of the Brain's Wiring." *Current Biology* 15 (12): 1143—9.

Feldman, Jacob. 2000. "Minimization of Boolean Complexity in Human Concept Learning." *Nature* 407 (6804): 630—3.

Feldman, J. A., and Ballard, Dana H. 1982. "Connectionist Models and Their Properties." *Cognitive Science* 6 (3): 205—254.

Foley, J. A., Monfreda, C., Ramankutty, N., and Zaks, D. 2007. "Our Share of the Planetary Pie." *Proceedings of the National Academy of Sciences of the United States of America* 104 (31): 12585—6.

Forgas, Joseph P., Cooper, Joel, and Crano, William D., eds. 2010. *The Psychology of Attitudes and Attitude Change*. Sydney Symposium of Social Psychology. New York: Psychology Press.

Frank, Robert H. 1999. *Luxury Fever: Why Money Fails to Satisfy in an Era of Excess*. New York: Free Press.

Fredriksen, Kaja Bonesmo. 2012. *Less Income Inequality and More Growth—Are They Compatible?: Part 6. The Distribution of Wealth*. Technical report, OECD Economics Department Working Papers 929. OECD Publishing.

Freitas, Robert A., Jr. 1980. "A Self-Replicating Interstellar Probe." *Journal of the British Interplanetary Society* 33: 251—64.

Freitas, Robert A., Jr. 2000. "Some Limits to Global Ecophagy by Biovorous Nanoreplicators, with Public Policy Recommendations." Foresight Institute. April. Retrieved July 28, 2013. Available at foresight.org/nano/Ecophagy.html.

Freitas, Robert A., Jr., and Merkle, Ralph C. 2004. *Kinematic Self—Replicating Machines*. Georgetown, TX: Landes Bioscience.

Gaddis, John Lewis. 1982. *Strategies of Containment: A Critical Appraisal of Postwar American National Security Policy*. New York: Oxford University Press.

Gammoned.net. 2012. "Snowie." Archived version. Retrieved June 30. Available at web.archive.org/web/20070920191840/http://www.gammoned.com/snowie.html.

Gates, Bill. 1975. "Software Contest Winners Announced." *Computer Notes* 1 (2): 1.

Georgieff, Michael K. 2007. "Nutrition and the Developing Brain: Nutrient Priorities and Measurement." *American Journal of Clinical Nutrition* 85 (2): 614S—620S.

Gianaroli, Luca. 2000. "Preimplantation Genetic Diagnosis: Polar Body and Embryo Biopsy." Supplement, *Human Reproduction* 15 (4): 69—75.

Gilovich, Thomas, Griffin, Dale, and Kahneman, Daniel, eds. 2002. *Heuristics and Biases: The Psychology of Intuitive Judgment*. New York: Cambridge University Press.

Gilster, Paul. 2012. "ESO: Habitable Red Dwarf Planets Abundant." *Centauri Dreams* (blog), March 29.

Goldstone, Jack A. 1980. "Theories of Revolution: The Third Generation." *World Politics* 32 (3): 425—53.

Goldstone, Jack A. 2001. "Towards a Fourth Generation of Revolutionary Theory." *Annual Review of Political Science* 4: 139—87.

Good, Irving John. 1965. "Speculations Concerning the First Ultraintelligent Machine." In *Advances in Computers*, edited by Franz L. Alt and Morris Rubinoff, 6: 31—88. New York: Academic Press.

Good, Irving John. 1970. "Some Future Social Repercussions of Computers." *International Journal of Environmental Studies* 1 (1—4): 67—79.

Good, Irving John. 1976. "Book review of 'The Thinking Computer: Mind Inside Matter" In *International Journal of Man-Machine Studies* 8: 617—20.

Good, Irving John. 1982. "Ethical Machines." In *Intelligent Systems: Practice and Perspective*, edited by I. E. Hayes, Donald Michie, and Y.-H. Pao, 555—60. Machine Intelligence 10. Chichester: Ellis Horwood.

Goodman, Nelson. 1954. *Fact, Fiction, and Forecast*. 1st ed. London: Athlone Press.

Gott, I. R., Iuric, M., Schlegel, D., Hoyle, F., Vogeley, M., Tegmark, M., Bahcall, N., and Brinkmann, I. 2005. "A Map of the Universe." *Astrophysical Journal* 624 (2): 463—83.

Gottfredson, Linda S. 2002. "G: Highly General and Highly Practical." In *The General Factor of Intelligence: How General Is It?*, edited by Robert J. Sternberg and Elena L. Grigorenko, 331—80. Mahwah, NI: Lawrence Erlbaum.

Gould, S. J. 1990. *Wonderful Life: The Burgess Shale and the Nature of History*. New York: Norton.

Graham, Gordon. 1997. *The Shape of the Past: A Philosophical Approach to History*. New York: Oxford University Press.

Gray, C. M., and McCormick, D. A. 1996. "Chattering Cells: Superficial Pyramidal Neurons Contributing to the Generation of Synchronous Oscillations in the Visual Cortex." *Science* 274 (5284): 109—13.

Greene, Kate. 2012. "Intel's Tiny Wi-Fi Chip Could Have a Big Impact." *MIT Technology Review*, September 21.

Guizzo, Erico. 2010. "World Robot Population Reaches 8.6 Million." *IEEE Spectrum*, April 14.

Gunn, James E. 1982. *Isaac Asimov: The Foundations of Science Fiction*. Science-Fiction Writers. New York: Oxford University Press.

Haberl, Helmut, Erb, Karl-Heinz, and Krausmann, Fridolin. 2013. "Global Human Appropriation of Net Primary Production (HANPP)." *Encyclopedia of Earth*, September 3.

Haberl, H., Erb, K. H., Krausmann, F., Gaube, V., Bondeau, A., Plutzar, C., Gingrich, S., Lucht, W, and Fischer—Kowalski, M. 2007. "Quantifying and Mapping the Human Appropriation of Net Primary Production in Earth's Terrestrial Ecosystems." *Proceedings of the National Academy of Sciences of the United States of America* 104 (31): 12942—7.

Hájek, Alan. 2009. "Dutch Book Arguments." In *Anand, Pattanaik, and Puppe* 2009, 173—95.

Hall, John Storrs. 2007. *Beyond AI: Creating the Conscience of the Machine*. Amherst, NY: Prometheus Books.

Hampson, R. B., Song, D., Chan, R. H., Sweatt, A. J., Riley, M. R., Gerhardt, G. A., Shin, D. C., Marmarelis, V. Z., Berger, T. W, and Deadwyler, S. A. 2012. "A Nonlinear Model for Hippocampal Cognitive Prosthesis: Memory Facilitation by Hippocampal Ensemble Stimulation." *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 20 (2): 184—97.

Hanson, Robin. 1994. "If Uploads Come First: The Crack of a Future Dawn." *Extropy* 6 (2).

Hanson, Robin. 1995. "Could Gambling Save Science? Encouraging an Honest Consensus." *Social Epistemology* 9 (1): 3—33.

Hanson, Robin. 1998a. "Burning the Cosmic Commons: Evolutionary Strategies for Inter- stellar Colonization." Unpublished manuscript, July 1. Retrieved April 26, 2012. http://hanson.gmu.edu/filluniv.pdf.

Hanson, Robin. 1998b. "Economic Growth Given Machine Intelligence." Unpublished manuscript. Retrieved May 15, 2013. Available at hanson.gmu.edu/aigrow.pdf.

Hanson, Robin. 1998c. "Long-Term Growth as a Sequence of Exponential Modes." Unpublished manuscript. Last revised December 2000. Available at hanson.gmu.edu/longgrow.pdf.

Hanson, Robin. 1998d. "Must Early Life Be Easy? The Rhythm of Major Evolutionary Transitions." Unpublished manuscript, September 23. Retrieved August 12, 2012. Available at hanson.gmu.edu/hardstep.pdf.

Hanson, Robin. 2000. "Shall We Vote on Values, But Bet on Beliefs?" Unpublished manuscript, September. Last revised October 2007. Available at hanson.gmu.edu/futarchy.pdf.

Hanson, Robin. 2006. "Uncommon Priors Require Origin Disputes." *Theory and Decision* 61 (4): 319—328.

Hanson, Robin. 2008. "Economics of the Singularity." *IEEE Spectrum* 45 (6): 45—50.

Hanson, Robin. 2009. "Tiptoe or Dash to Future?" *Overcoming Bias* (blog), December 23.

Hanson, Robin. 2012. "Envisioning the Economy, and Society, of Whole Brain Emula- tions." Paper presented at the AGI Impacts conference, Oxford, December 8—1 1.

Hart, Oliver. 2008. "Economica Coase Lecture Reference Points and the Theory of the Firm." *Economica* 75 (299): 404—11.

Hay, Nicholas James. 2005. "Optimal Agents." B.Sc. thesis, University of Auckland.

Hedberg, Sara Reese. 2002. "Dart: Revolutionizing Logistics Planning." *IEEE Intelligent Systems* 17 (3): 81—3.

Helliwell, John, Layard, Richard, and Sachs, Jeffrey. 2012. *World Happiness Report*. The Earth Institute.

Helmstaedter, M., Briggman, K. L., and Denk, W. 2011. "High-Accuracy Neurite Reconstruction for High-Throughput Neuroanatomy." *Nature Neuroscience* 14 (8): 1081—8.

Heyl, Jeremy S. 2005. "The Long—Term Future of Space Travel." *Physical Review* D 72 (10): 1—4.

Hibbard, Bill. 2011. "Measuring Agent Intelligence via Hierarchies of Environments." In *Artificial General Intelligence: 4th International Conference, AGI 2011, Mountain View, CA, USA, August 3—6,*

2011. Proceedings, edited by Jürgen Schmidhuber, Kristinn R. Thorisson, and Moshe Looks, 303—8. Lecture Notes in Computer Science 6830. Berlin: Springer.

Hinke, R. M., Hu, X., Stillman, A. E., Herkle, H., Salmi, R., and Ugurbil, K. 1993. "Functional Magnetic Resonance Imaging of Broca's Area During Internal Speech." *Neuroreport* 4 (6): 675—8.

Hinxton Group. 2008. *Consensus Statement: Science, Ethics and Policy Challenges of Pluripotent Stem Cell-Derived Gametes*. Hinxton, Cambridgeshire, UK, April 11. Available at hinxtongroup.org/Consensus_HG08_FINAL.pdf.

Hoffman, David E. 2009. *The Dead Hand: The Untold Story of the Cold War Arms Race and Its Dangerous Legacy*. New York: Doubleday.

Hofstadter, Douglas R. (1979) 1999. *Godel, Escher, Bach: An Eternal Golden Braid*. New York: Basic Books.

Holley, Rose. 2009. "How Good Can It Get? Analysing and Improving OCR Accuracy in Large Scale Historic Newspaper Digitisation Programs." *D-Lib Magazine* 15 (3—4).

Horton, Sue, Alderman, Harold, and Rivera, Juan A. 2008. *Copenhagen Consensus 2008 Challenge Paper: Hunger and Malnutrition*. Technical report. Copenhagen Consensus Center, May 11.

Howson, Colin, and Urbach, Peter. 1993. *Scientific Reasoning: The Bayesian Approach*. 2nd ed. Chicago: Open Court.

Hsu, Stephen. 2012. "Investigating the Genetic Basis for Intelligence and Other Quantitative Traits." Lecture given at UC Davis Department of Physics Colloquium, Davis, CA, February 13.

Huebner, Bryce. 2008. "Do You See What We See? An Investigation of an Argument Against Collective Representation." *Philosophical Psychology* 21 (1): 91—112.

Huff, C. D., Xing, J., Rogers, A. R., Witherspoon, D., and Jorde, L. B. 2010. "Mobile Elements Reveal Small Population Size in the Ancient Ancestors of *Homo Sapiens*." *Proceedings of the National Academy of Sciences of the United States of America* 107 (5): 2147—52.

Huffman, W. Cary, and Pless, Vera. 2003. *Fundamentals of Error—Correcting Codes*. New York: Cambridge University Press.

Hunt, Patrick. 201 1. "Late Roman Silk: Smuggling and Espionage in the 6th Century CE." *Philolog, Stanford University* (blog), August 2.

Hutter, Marcus. 2001. "Towards a Universal Theory of Artificial Intelligence Based on Algorithmic Probability and Sequential Decisions." In De Raedt and Flach 2001, 226—38.

Hutter, Marcus. 2005. *Universal Artificial Intelligencet: Sequential Decisions Based On Algorithmic Probability*. Texts in Theoretical Computer Science. Berlin: Springer.

Iliadou, A. N., Ianson, P. C., and Cnattingius, S. 201 1. "Epigenetics and Assisted Reproduc- tive Technology." *Journal of Internal Medicine* 270 (5): 414—20.

Isaksson, Anders. 2007. *Productivity and Aggregate Growth: A Global Picture*. Technical report 05/2007. Vienna, Austria: UNIDO (United Nations Industrial Development Organization) Research and Statistics Branch.

Jones, Garret. 2009. "Artificial Intelligence and Economic Growth: A Few Finger- Exercises." Unpublished manuscript, January. Retrieved November 5, 2012. Available at mason.gmu.edu/~gjonesb/AIandGrowth.

Jones, Vincent C. 1985. *Manhattan: The Army and the Atomic Bomb*. United States Army in World War II. Washington, DC: Center of Military History.

Joyce, James M. 1999. *The Foundations of Causal Decision Theory*. Cambridge Studies in Probability, Induction and Decision Theory. New York: Cambridge University Press.

Judd, K. L., Schmedders, K., and Yeltekin, S. 2012. "Optimal Rules for Patent Races." *International Economic Review* 53 (1): 23—52.

Kalfoglou, A., Suthers, K., Scott, J., and Hudson, K. 2004. *Reproductive Genetic Testing: What America Thinks*. Genetics and Public Policy Center.

Kamm, Frances M. 2007. *Intricate Ethics: Rights, Responsibilities, and Permissible Harm*. Oxford Ethics Series. New York: Oxford University Press.

Kandel, Eric R., Schwartz, James H., and Jessell, Thomas M., eds. 2000. *Principles of Neural Science*. 4th ed. New York: McGraw-Hill.

Kansa, Eric. 2003. "Social Complexity and Flamboyant Display in Competition: More Thoughts on the Fermi Paradox." Unpublished manuscript, archived version.

Karnofsky, Holden. 2012. "Comment on 'Reply to Holden on Tool AI.' " *Less Wrong* (blog), August 1.

Kasparov, Garry. 1996. "The Day That I Sensed a New Kind of Intelligence." *Time*, March 25, no. 13.

Kaufman, Jeff. 201 1. "Whole Brain Emulation and Nematodes." *Jeff Kaufman's Blog* (blog), November 2.

Keim, G. A., Shazeer, N. M., Littman, M. L., Agarwal, S., Cheves, C. M., Fitzgerald, J., Grosland, J., Jiang, F., Pollard, S., and Weinmeister, K. 1999. "Proverb: The Probabilistic Cruciverbalist." In *Proceedings of the Sixteenth National Conference on Artificial Intelligence*, 710—17. Menlo Park, CA: AAAI Press.

Kell, Harrison J., Lubinski, David, and Benbow, Camilla P. 2013. "Who Rises to the Top? Early Indicators." *Psychological Science* 24 (5): 648—59.

Keller, Wolfgang. 2004. "International Technology Diffusion." *Journal of Economic Literature* 42 (3): 752—82.

KGS Go Server. 2012. "KGS Game Archives: Games of KGS player zen19." Retrieved July 22, 2013. Available at gokgs.com/gameArchives.jsp?user= zen19d81oldAcco unts=t81year=201281month=3.

Knill, Emanuel, Laflamme, Raymond, and Viola, Lorenza. 2000. "Theory of Quantum Error Correction for General Noise." *Physical Review Letters* 84 (11): 2525—8.

Koch, K., McLean, J., Segev, R., Freed, M. A., Berry, M. J., Balasubramanian, V., and Sterling, P. 2006. "How Much the Eye Tells the Brain." *Current Biology* 16 (14): 1428—34.

Kong, A., Frigge, M. L., Masson, G., Besenbacher, S., Sulem, P., Magnusson, G., Gudjonsson, S. A., Sigurdsson, A., et al. 2012. "Rate of De Novo Mutations and the Importance of Father's Age to Disease Risk." *Nature* 488: 471—5.

Koomey, Jonathan G. 2011. *Growth in Data Center Electricity Use 2005 to 2010*. Technical report, 08/01/2011. Oakland, CA: Analytics Press.

Koubi, Vally. 1999. "Military Technology Races." *International Organization* 53 (3): 537—65.

Koubi, Vally, and Lalman, David. 2007. "Distribution of Power and Military R81D." *Journal of Theoretical Politics* 19 (2): 133—52.

Koza, J. R., Keane, M. A., Streeter, M. J., Mydlowec, W., Yu, J., and Lanza, G. 2003. *Genetic Programming IV: Routine Human—Competitive Machine Intelligence*. 2nd ed. Genetic Programming. Norwell, MA: Kluwer Academic.

Kremer, Michael. 1993. "Population Growth and Technological Change: One Million BC. to 1990." *Quarterly Journal of Economics* 108 (3): 681—716.

Kruel, Alexander. 2011. "Interview Series on Risks from AI." *Less Wrong Wiki* (blog). Retrieved Oct 26, 2013. Available at wiki.lesswrong.com/wiki/Interview_series_ on_risks_from_AI.

Kruel, Alexander. 2012. "Q81A with Experts on Risks From AI #2." *Less Wrong* (blog), January 9.

Krusienski, D. I., and Shih, I. I. 2011. "Control of a Visual Keyboard Using an Electrocorticographic Brain—Computer Interface." *Neurorehabilitation and Neural Repair* 25 (4): 323—31.

Kuhn, Thomas S. 1962. *The Structure of Scientific Revolutions*. 1st ed. Chicago: University of Chicago Press.

Kuipers, Benjamin. 2012. "An Existing, Ecologically—Successful Genus of Collectively Intelligent Artificial Creatures." Paper presented at the 4th International Conference, ICCCI 2012, Ho Chi Minh City, Vietnam, November 28—30.

Kurzweil, Ray. 2001. "Response to Stephen Hawking." Kurzweil Accelerating Intelligence. September 5. Retrieved December 31, 2012. Available at kurzweilai.net/ response-to-stephen-hawking.

Kurzweil, Ray. 2005. *The Singularity Is Near: When Humans Transcend Biology*. New York: Viking.

Laffont, Jean-Jacques, and Martimort, David. 2002. *The Theory of Incentives: The Principal- Agent Model*. Princeton, NJ: Princeton University Press.

Lancet, The. 2008. "Iodine Deficiency—Way to Go Yet." *The Lancet* 372 (9633): 88.

Landauer, Thomas K. 1986. "How Much Do People Remember? Some Estimates of the Quantity of Learned Information in Long-Term Memory." *Cognitive Science* 10 (4): 477—93.

Lebedev, Anastasiya. 2004. "The Man Who Saved the World Finally Recognized." *Mos-News*, May 21.

Lebedev, M. A., and Nicolelis, M. A. 2006. "Brain—Machine Interfaces: Past, Present and Future." *Trends in Neuroscience* 29 (9): 536—46.

Legg, Shane. 2008. "Machine Super Intelligence." PhD dissertation, University of Lugano.

Leigh, E. G., Ir. 2010. "The Group Selection Controversy." *Journal of Evolutionary Biology* 23 (1): 6—19.

Lenat, Douglas B. 1982. "Learning Program Helps Win National Fleet Wargame Tournament." *SIGART Newsletter* 79: 16—17.

Lenat, Douglas B. 1983. "EURISKO: A Program that Learns New Heuristics and Domain Concepts." *Artificial Intelligence* 21 (1—2): 61—98.

Lenman, James. 2000. "Consequentialism and Cluelessness." *Philosophy & Public Aflairs* 29 (4): 342—70.

Lerner, Josh. 1997. "An Empirical Exploration of a Technology Race." *RAND Journal of Economics* 28 (2): 228—47.

Leslie, John. 1996. *The End of the World: The Science and Ethics of Human Extinction*. London: Routledge.

Lewis, David. 1988. "Desire as Belief." *Mind: A Quarterly Review of Philosophy* 97 (387): 323—32.

Li, Ming, and Vitanyi, Paul M. B. 2008. *An Introduction to Kolmogorov Complexity and Its Applications*. Texts in Computer Science. New York: Springer.

Lin, Thomas, Mausam, and Etzioni, Oren. 2012. "Entity Linking at Web Scale." In *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web—scale Knowledge Extraction (AKBC-WEKEX '12),* edited by James Fan, Raphael Hoffman, Aditya Kalyanpur, Sebastian Riedel, Fabian Suchanek, and Partha Pratim Talukdar, 84—88. Madison, WI: Omnipress.

Lloyd, Seth. 2000. "Ultimate Physical Limits to Computation." *Nature* 406 (6799): 1047—54.

Louis Harris & Associates. 1969. "Science, Sex, and Morality Survey, study no. 1927." *Life Magazine* (New York) 4.

Lynch, Michael. 2010. "Rate, Molecular Spectrum, and Consequences of Human Mutation." *Proceedings of the National Academy of Sciences of the United States of America* 107 (3): 961—8.

Lyons, Mark K. 201 1. "Deep Brain Stimulation: Current and Future Clinical Applications." *Mayo Clinic Proceedings* 86 (7): 662—72.

MacAskill, William. 2010. "Moral Uncertainty and Intertheoretic Comparisons of Value." BPhil thesis, University of Oxford.

McCarthy, John. 2007. "From Here to Human-Level AI." *Artificial Intelligence* 171 (18): 1 174—82.

McCorduck, Pamela. 1979. *Machines Who Think: A Personal Inquiry into the History and Prospects of Artificial Intelligence*. San Francisco: W. H. Freeman.

Mack, C. A. 2011. "Fifty Years of Moore's Law." *IEEE Transactions on Semiconductor Manufacturing* 24 (2): 202—7.

MacKay, David I. C. 2003. *Information Theory, Inference, and Learning Algorithms*. New York: Cambridge University Press.

McLean, George, and Stewart, Brian. 1979. "Norad False Alarm Causes Uproar." *The National*. Aired November 10. Ottawa, ON: CBC, 2012. News Broadcast.

Maddison, Angus. 1999. "Economic Progress: The Last Half Century in Historical Perspective." In *Facts and Fancies of Human Development: Annual Symposium and Cunningham Lecture*, 1999, edited by Ian Castles. Occasional Paper Series, 1/2000. Acton, ACT: Academy of the Social Sciences in Australia.

Maddison, Angus. 2001. *The World Economy: A Millennial Perspective*. Development Centre Studies. Paris: Development Centre of the Organisation for Economic Cooperation and Development.

Maddison, Angus. 2005. *Growth and Interaction in the World Economy: The Roots of Modernity*. Washington, DC: AEI Press.

Maddison, Angus. 2007. *Contours of the World Economy, 1—2030 AD: Essays in Macro- Economic History*. New York: Oxford University Press.

Maddison, Angus. 2010. "Statistics of World Population, GDP and Per Capita GDP 1- 2008 AD." Retrieved October 26, 2013. Available at ggdc.net/maddison/Historical_Statistics/vertical—file_O2-2010.xls.

Mai, Q., Yu, Y., Li, T., Wang, L., Chen, M. J., Huang, S. Z., Zhou, C., and Zhou, Q. 2007. "Derivation of Human Embryonic Stem Cell Lines from Parthenogenetic Blastocysts." *Cell Research* 17 (12): 1008—19.

Mak, J. N., and Wolpaw, J. R. 2009. "Clinical Applications of Brain—Computer Interfaces: Current State and Future Prospects." *IEEE Reviews in Biomedical Engineering* 2: 187—99.

Mankiw, N. Gregory. 2009. *Macroeconomics*. 7th ed. New York, NY: Worth.

Mardis, Elaine R. 2011. "A Decade's Perspective on DNA Sequencing Technology." *Nature* 470 (7333): 1953—203.

Markoff, John. 2011. "Computer Wins on 'Jeopardy!': Trivial, It's Not." *New York Times*, February 16.

Markram, Henry. 2006. "The Blue Brain Project." *Nature Reviews Neuroscience* 7 (2): 153—160.

Mason, Heather. 2003. "Gallup Brain: The Birth of In Vitro Fertilization." *Gallup*, August 5.

Menzel, Randolf, and Giurfa, Martin. 2001. "Cognitive Architecture of a Mini-Brain: The Honeybee." *Trends in Cognitive Sciences* 5 (2): 62—71.

Metzinger, Thomas. 2003. *Being No One: The Self—Model Theory of Subjectivity*. Cambridge, MA: MIT Press.

Mijic, Roko. 2010. "Bootstrapping Safe AGI Goal Systems." Paper presented at the Road- maps to AGI and the Future of AGI Workshop, Lugano, Switzerland, March 8.

Mike, Mike. 2013. "Face of Tomorrow." Retrieved Iune 30, 2012 . Available at faceoftomorrow.org.

Milgrom, Paul, and Roberts, John. 1990. "Bargaining Costs, Influence Costs, and the Organization of Economic Activity." In *Perspectives on Positive Political Economy*, edited by James E. Alt and Kenneth A. Shepsle, 57—89. New York: Cambridge University Press.

Miller, George A. 1956. "The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information." *Psychological Review* 63 (2): 81—97.

Miller, Geoffrey. 2000. *The Mating Mind: How Sexual Choice Shaped the Evolution of Human Nature*. New York: Doubleday.

Miller, James D. 2012. *Singularity Rising: Surviving and Thriving in a Smarter, Richer, and More Dangerous World*. Dallas, TX: BenBella Books.

Minsky, Marvin. 1967. *Computation: Finite and Infinite Machines*. Englewood Cliffs, NJ: Prentice-Hall.

Minsky, Marvin, ed. 1968. *Semantic Information Processing*. Cambridge, MA: MIT Press.

Minsky, Marvin. 1984. "Afterword to Vernor Vinge's novel, 'True Names.' " Unpublished manuscript, October 1. Retrieved December 31, 2012. Available at web.media.mit.edu/

~minsky/papers/TrueNames.Afterword.html.

Minsky, Marvin. 2006. *The Emotion Machine: Commonsense Thinking, Artificial Intelligence, and the Future of the Human Mind*. New York: Simon 81 Schuster.

Minsky, Marvin, and Papert, Seymour. 1969. *Perceptrons: An Introduction to Computational Geometry*. 1st ed. Cambridge, MA: MIT Press.

Moore, Andrew. 2011. "Hedonism." In *The Stanford Encyclopedia of Philosophy*, Winter 2011, edited by Edward N. Zalta. Stanford, CA: Stanford University.

Moravec, Hans P. 1976. "The Role of Raw Power in Intelligence." Unpublished manuscript, May 12. Retrieved August 12, 2012. Available at frc.ri.cmu.edu/users/hpm/project.archive/general.articles/l975/Raw.Power.html.

Moravec, Hans P. 1980. "Obstacle Avoidance and Navigation in the Real World by a Seeing Robot Rover." PhD dissertation, Stanford University.

Moravec, Hans P. 1988. *Mind Children: The Future of Robot and Human Intelligence*. Cambridge, MA: Harvard University Press.

Moravec, Hans P. 1998. "When Will Computer Hardware Match the Human Brain?" *Journal of Evolution and Technology* 1.

Moravec, Hans P. 1999. "Rise of the Robots." *Scientific American*, December, 124—35.

Muehlhauser, Luke, and Helm, Louie. 2012. "The Singularity and Machine Ethics." In *Singularity Hypotheses: A Scientific and Philosophical Assessment*, edited by Amnon Eden, Johnny Søraker, James H. Moor, and Eric Steinhart. The Frontiers Collection. Berlin: Springer.

Muehlhauser, Luke, and Salamon, Anna. 2012. "Intelligence Explosion: Evidence and Import." In *Singularity Hypotheses: A Scientific and Philosophical Assessment*, edited by Amnon Eden, Johnny Søraker, James H. Moor, and Eric Steinhart. The Frontiers Collection. Berlin: Springer.

Müller, Vincent C., and Bostrom, Nick. 2016. "Future Progress in Artificial Intelligence: A Survey of Expert Opinion." In *Fundamental Issues of Artificial Intelligence*, edited by Vincent C. Müller. Synthese Library; Berlin: Springer.

Murphy, Kevin P. 2012. *Machine Learning: A Probabilistic Perspective*. Adaptive Computation and Machine Learning. Cambridge, MA: MIT Press.

Nachman, Michael W., and Crowell, Susan L. 2000. "Estimate of the Mutation Rate per Nucleotide in Humans." *Genetics* 156 (1): 297—304.

Nagy, Z. P., and Chang, C. C. 2007. "Artificial Gametes." *Theriogenology* 67 (1): 99—104.

Nagy, Z. P., Kerkis, I., and Chang, C. C. 2008. "Development of Artificial Gametes." *Reproductive BioMedicine Online* 16 (4): 539—44.

NASA. 2013. "International Space Station: Facts and Figures." Available at nasa.gov/worldbook/intspacestation_worldbook.html.

Newborn, Monty. 2011. *Beyond Deep Blue: Chess in the Stratosphere*. New York: Springer.

Newell, Allen, Shaw, J. C., and Simon, Herbert A. 1958. "Chess-Playing Programs and the Problem of Complexity." *IBM Journal of Research and Development* 2 (4): 320—35.

Newell, Allen, Shaw, J. C., and Simon, Herbert A. 1959. "Report on a General Problem- Solving Program: Proceedings of the International Conference on Information Processing." In *Information Processing*, 256—64. Paris: UNESCO.

Nicolelis, Miguel A. L., and Lebedev, Mikhail A. 2009. "Principles of Neural Ensemble Physiology Underlying the Operation of Brain—Machine Interfaces." *Nature Reviews Neuroscience* 10 (7): 530—40.

Nilsson, Nils J. 1984. *Shakey the Robot*, Technical Note 323. Menlo Park, CA: AI Center, SRI International, April.

Nilsson, Nils J. 2009. *The Quest for Artificial Intelligence: A History of Ideas and Achievements*. New York: Cambridge University Press.

Nisbett, R. E., Aronson, J., Blair, C., Dickens, W, Flynn, J., Halpern, D. F., and Turkheimer, E. 2012. "Intelligence: New Findings and Theoretical Developments." *American Psychologist* 67 (2): 130—59.

Niven, Larry. 1973. "The Defenseless Dead." In *Ten Tomorrows*, edited by Roger Elwood, 91—142. New York: Fawcett.

Nordhaus, William D. 2007. "Two Centuries of Productivity Growth in Computing." *Journal of Economic History* 67(1): 128—59.

Norton, John D. 2011. "Waiting for Landauer." Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics 42 (3): 184—98.

Olds, James, and Milner, Peter. 1954. "Positive Reinforcement Produced by Electrical Stimulation of Septal Area and Other Regions of Rat Brain." *Journal of Comparative and Physiological Psychology* 47 (6): 419—27.

Olum, Ken D. 2002. "The Doomsday Argument and the Number of Possible Observers." *Philosophical Quarterly* 52 (207): 164—84.

Omohundro, Stephen M. 2007. "The Nature of Self-Improving Artificial Intelligence." Paper presented at Singularity Summit 2007, San Francisco, CA, September 8—9.

Omohundro, Stephen M. 2008. "The Basic AI Drives." In *Artificial General Intelligence 2008: Proceedings of the First AGI Conference*, edited by Pei Wang, Ben Goertzel, and Stan Franklin, 483—92. Frontiers in Artificial Intelligence and Applications 171. Amsterdam: IOS.

Omohundro, Stephen M. 2012. "Rational Artificial Intelligence for the Greater Good." In *Singularity Hypotheses: A Scientific and Philosophical Assessment*, edited by Amnon Eden, Johnny Søraker, James H. Moor, and Eric Steinhart. The Frontiers Collection. Berlin: Springer.

O'Neill, Gerard K. 1974. "The Colonization of Space." *Physics Today* 27 (9): 32—40.

Oshima, Hideki, and Katayama, Yoichi. 2010. "Neuroethics of Deep Brain Stimulation for Mental Disorders: Brain Stimulation Reward in Humans." *Neurologia medico-chirurgica* 50 (9): 845—52.

Parfit, Derek. 1986. *Reasons and Persons*. New York: Oxford University Press.

Parfit, Derek. 2011. *On What Matters*. 2 vols. The Berkeley Tanner Lectures. New York: Oxford University Press.

Parrington, Alan J. 1997. "Mutually Assured Destruction Revisited." *Airpower Journal* 11 (4).

Pasqualotto, Emanuele, Federici, Stefano, and Belardinelli, Marta Olivetti. 2012. "Toward Functioning and Usable Brain—Computer Interfaces (BCIs): A Literature Review." *Disability and Rehabilitation: Assistive Technology* 7 (2): 89—103.

Pearl, Judea. 2009. *Causality: Models, Reasoning, and Inference*. 2nd ed. New York: Cambridge University Press.

Perlmutter, J. S., and Mink, J. W. 2006. "Deep Brain Stimulation." *Annual Review of Neuroscience* 29: 229—57.

Pinker, Steven. 2011. *The Better Angels of Our Nature: Why Violence Has Declined*. New York: Viking.

Plomin, R., Haworth, C. M., Meaburn, E. L., Price, T. S., Wellcome Trust Case Control Consortium 2, and Davis, O. S. 2013. "Common DNA Markers Can Account for More Than Half of the Genetic Influence on Cognitive Abilities." *Psychological Science* 24 (2): 562—8.

Popper, Nathaniel. 2012. "Flood of Errant Trades Is a Black Eye for Wall Street." *New York Times*, August 1.

Pourret, Olivier, Naim, Patrick, and Marcot, Bruce, eds. 2008. *Bayesian Networks: A Practical Guide to Applications*. Chichester, West Sussex, UK: Wiley.

Powell, A., Shennan, S., and Thomas, M. G. 2009. "Late Pleistocene Demography and the Appearance of Modern Human Behavior." *Science* 324 (5932): 1298—1301.

Price, Huw. 1991. "Agency and Probabilistic Causality." *British Journal for the Philosophy of Science* 42(2): 157—76.

Qian, M., Wang, D., Watkins, W. E., Gebski, V., Yan, Y. Q., Li, M., and Chen, Z. P. 2005. "The Effects of Iodine on Intelligence in Children: A Meta-Analysis of Studies Conducted in China." *Asia Pacific Journal of Clinical Nutrition* 14 (1): 32—42.

Quine, Willard Van Orman, and Ullian, Ioseph Silbert. 1978. *The Web of Belief*, ed. Richard Malin Ohmann, vol. 2. New York: Random House.

Railton, Peter. 1986. "Facts and Values." *Philosophical Topics* 14 (2): 5—31.

Rajab, Moheeb Abu, Zarfoss, Jay, Monrose, Fabian, and Terzis, Andreas. 2006. "A Multi-faceted Approach to Understanding the Botnet Phenomenon." In *Proceedings of the 6th ACM SIGCOMM Conference on Internet Measurement*, 41—52. New York: ACM.

Rawls, John. 1971. *A Theory of Justice*. Cambridge, MA: Belknap.

Read, J. I., and Trentham, Neil. 2005. "The Baryonic Mass Function of Galaxies." *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 363 (1837): 2693—710.

Repantis, D., Schlattmann, P., Laisney, O., and Heuser, I. 2010. "Modafinil and Methylphenidate for Neuroenhancement in Healthy Individuals: A Systematic Review." *Pharmacological Research* 62 (3): 187—206.

Rhodes, Richard. 1986. *The Making of the Atomic Bomb*. New York: Simon & Schuster.

Rhodes, Richard. 2008. *Arsenals of Folly: The Making of the Nuclear Arms Race*. New York: Vintage.

Rietveld, Cornelius A., Medland, Sarah E., Derringer, Jaime, Yang, Jian, Esko, Tonu, Martin, Nicolas W., Westra, Harm-Jan, Shakhbazov, Konstantin, Abdellaoui, Abdel, et l1. 2013. "GWAS of 126,559 Individuals Identifies Genetic Variants Associated with Educational Attainment." *Science* 340 (6139): 1467—71.

Ring, Mark, and Orseau, Laurent. 2011. "Delusion, Survival, and Intelligent Agents." In *Artificial General Intelligence: 4th International Conference, AGI 2011, Mountain View, CA, USA, August 3—6, 2011. Proceedings*, edited by Jürgen Schmidhuber, Kristinn R. Thorisson, and Moshe Looks, 11—20. Lecture Notes in Computer Science 6830. Berlin: Springer.

Ritchie, Graeme, Manurung, Ruli, and Waller, Annalu. 2007. "A Practical Application of Computational Humour." In *Proceedings of the 4th International Joint Workshop on Computational Creativity*, edited by Amilcar Cardoso and Geraint A. Wiggins, 91—8. London: Goldsmiths, University of London.

Roache, Rebecca. 2008. "Ethics, Speculation, and Values." *NanoEthics* 2 (3): 317—27.

Robles, J. A., Lineweaver, C. H., Grether, D., Flynn, C., Egan, C. A., Pracy, M. B., Holmberg, J., and Gardner, E. 2008. "A Comprehensive Comparison of the Sun to Other Stars: Searching for Self-Selection Effects." *Astrophysical Journal* 684 (1): 691—706.

Roe, Anne. 1953. *The Making of a Scientist*. New York: Dodd, Mead.

Roy, Deb. 2012. "About." Retrieved October 14. Available at web.media.mit. edu/~dkroy/.

Rubin, Jonathan, and Watson, Ian. 2011. "Computer Poker: A Review." *Artificial Intelligence* 175 (5—6): 958—87.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. 1986. "Learning Representations by Back-Propagating Errors." *Nature* 323 (6088): 533—6.

Russell, Bertrand. 1986. "The Philosophy of Logical Atomism." In *The Philosophy of Logical Atomism and Other Essays 1914—1919*, edited by John G. Slater, 8: 157—244. The Collected Papers of Bertrand Russell. Boston: Allen 81 Unwin.

Russell, Bertrand, and Griffin, Nicholas. 2001. *The Selected Letters of Bertrand Russell: The Public Years, 1914—1970*. New York: Routledge.

Russell, Stuart J., and Norvig, Peter. 2010. *Artificial Intelligence: A Modern Approach*. 3rd ed. Upper Saddle River, NJ: Prentice-Hall.

Sabrosky, Curtis W. 1952. "How Many Insects Are There?" In *Insects*, edited by United States Department of Agriculture, 1—7. Yearbook of Agriculture. Washington, DC: United States Government Printing Office.

Salamon, Anna. 2009. "When Software Goes Mental: Why Artificial Minds Mean Fast Endogenous Growth." Working Paper, December 27.

Salem, D. J., and Rowan, A. N. 2001. *The State of the Animals: 2001*. Public Policy Series. Washington, DC: Humane Society Press.

Salverda, W, Nolan, B., and Smeeding, T. M. 2009. *The Oxford Handbook of Economic Inequality*. Oxford: Oxford University Press.

Samuel, A. L. 1959. "Some Studies in Machine Learning Using the Game of Checkers." *IBM Journal of Research and Development* 3 (3): 2 10— 19.

Sandberg, Anders. 1999. "The Physics of Information Processing Superobjects: Daily Life Among the Iupiter Brains." *Journal of Evolution and Technology* 5.

Sandberg, Anders. 2010. "An Overview of Models of Technological Singularity." Paper presented at the Roadmaps to AGI and the Future of AGI Workshop, Lugano, Switzerland, March 8.

Sandberg, Anders. 2013. "Feasibility of Whole Brain Emulation." In *Philosophy and Theory ofArtificial Intelligence*, edited by Vincent C. Muller, 5: 251—64. Studies in Applied Philosophy, Epistemology and Rational Ethics. New York: Springer.

Sandberg, Anders, and Bostrom, Nick. 2006. "Converging Cognitive Enhancements." *Annals of the New York Academy of Sciences* 1093: 201—27.

Sandberg, Anders, and Bostrom, Nick. 2008. *Whole Brain Emulation: A Roadmap*. Technical Report 2008-3. Future of Humanity Institute, University of Oxford.

Sandberg, Anders, and Bostrom, Nick. 2011. *Machine Intelligence Survey. Technical Report* 2011-1. Future of Humanity Institute, University of Oxford.

Sandberg, Anders, and Savulescu, Julian. 2011. "The Social and Economic Impacts of Cognitive Enhancement." In *Enhancing Human Capacities*, edited by Iulian Savulescu, Ruud ter Meulen, and Guy Kahane, 92—112. Malden, MA: Wiley—Blackwell.

Schaeffer, Jonathan. 1997. *One Jump Ahead: Challenging Human Supremacy in Checkers*. New York: Springer.

Schaeffer, J., Burch, N., Bjornsson, Y., Kishimoto, A., Muller, M., Lake, R., Lu, P., and Sutphen, S. 2007. "Checkers Is Solved." *Science* 317 (5844): 1518—22.

Schalk, Gerwin. 2008. "Brain—Computer Symbiosis." *Journal of Neural Engineering* 5 (1): Pl—PIS.

Schelling, Thomas C. 1980. *The Strategy of Conflict*. 2nd ed. Cambridge, MA: Harvard University Press.

Schultz, T. R. 2000. "In Search of Ant Ancestors." *Proceedings of the National Academy of Sciences of the United States of America* 97 (26): 14028—9.

Schultz, W, Dayan, P., and Montague, P. R. 1997. "A Neural Substrate of Prediction and Reward." *Science* 275 (5306): 1593—9.

Schwartz, Jacob T. 1987. "Limits of Artificial Intelligence." In *Encyclopedia of Artificial Intelligence*, edited by Stuart C. Shapiro and David Eckroth, 1: 488—503. New York: Wiley.

Schwitzgebel, Eric. 2013. "If Materialism is True, the United States is Probably Conscious." *Working Paper*, February 8.

Sen, Amartya, and Williams, Bernard, eds. 1982. *Utilitarianism and Beyond*. New York: Cambridge University Press.

Shanahan, Murray. 2010. *Embodiment and the Inner Life: Cognition and Consciousness in the Space of Possible Minds*. New York: Oxford University Press.

Shannon, Robert V. 2012. "Advances in Auditory Prostheses." *Current Opinion in Neurology* 25 (1): 61—6.

Shapiro, Stuart C. 1992. "Artificial Intelligence." In *Encyclopedia of Artificial Intelligence*, 2nd ed., 1: 54—7. New York: Wiley.

Sheppard, Brian. 2002. "World—Championship-Caliber Scrabble." *Artificial Intelligence* 134 (1—2): 241—75.

Shoemaker, Sydney. 1969. "Time Without Change." *Journal of Philosophy* 66 (12): 363—81.

Shulman, Carl. 2010a. *Omohundro's "Basic AI Drives" and Catastrophic Risks*. Berkeley, CA: Machine Intelligence Research Institute.

Shulman, Carl. 2010b. *Whole Brain Emulation and the Evolution of Superorganisms*. Berkeley, CA: Machine Intelligence Research Institute.

Shulman, Carl. 2012. "Could We Use Untrustworthy Human Brain Emulations to Make Trust- worthy Ones?" Paper presented at the AGI Impacts conference, Oxford, December 8—11.

Shulman, Carl, and Bostrom, Nick. 2012. "How Hard is Artificial Intelligence? Evolutionary Arguments and Selection Effects." *Journal of Consciousness Studies* 19 (7—8): 103—30.

Shulman, Carl, and Bostrom, Nick. 2014. "Embryo Selection for Cognitive Enhancement: Curiosity or Game-Changer?" *Global Policy* 5 (1): 85—92.

Shulman, Carl, Ionsson, Henrik, and Tarleton, Nick. 2009. "Which Consequentialism? Machine Ethics and Moral Divergence." In *AP—CAP 2009: The Fifth Asia-Pacific Com- puting and Philosophy Conference, October 1st-2nd, University of Tokyo, Japan.Proceedings*, edited by Carson Reynolds and Alvaro Cassinelli, 23—25. AP—CAP 2009.

Sidgwick, Henry, and Jones, Emily Elizabeth Constance. 2010. *The Methods of Ethics*. Charleston, SC: Nabu Press.

Silver, Albert. 2006. "How Strong Is GNU Backgammon?" Backgammon Galore! September 16. Retrieved October 26, 2013. Available at bkgm.com/gnu/AllAboutGNU.html#how_strong_is_gnu.

Simeral, J. D., Kim, S. P., Black, M. J., Donoghue, J. P., and Hochberg, L. R. 2011. "Neural Control of Cursor Trajectory and Click by a Human with Tetraplegia 1000 Days after Implant of an Intracortical Microelectrode Array." *Journal of Neural Engineering* 8 (2): 025027.

Simester, Duncan, and Knez, Marc. 2002. "Direct and Indirect Bargaining Costs and the Scope of the Firm." *Journal of Business* 75 (2): 283—304.

Simon, Herbert Alexander. 1965. *The Shape of Automation for Men and Management*. New York: Harper 81 Row.

Sinhababu, Neil. 2009. "The Humean Theory of Motivation Reformulated and Defended." *Philosophical Review* 118 (4): 465—500.

Slagle, James R. 1963. "A Heuristic Program That Solves Symbolic Integration Problems in Freshman Calculus." *Journal of the ACM* 10 (4): 507—20.

Smeding, H. M., Speelman, J. D., Koning—Haanstra, M., Schuurman, P. R., Nijssen, P., van Laar, T., and Schmand, B. 2006. "Neuropsychological Effects of Bilateral STN Stimulation in Parkinson

Disease: A Controlled Study." *Neurology* 66 (12): 1830—6.

Smith, Michael. 1987. "The Humean Theory of Motivation." *Mind: A Quarterly Review of Philosophy* 96 (381): 36—61.

Smith, Michael, Lewis, David, and Johnston, Mark. 1989. "Dispositional Theories of Value." *Proceedings of the Aristotelian Society* 63: 89—174.

Sparrow, Robert. 2013. "In Vitro Eugenics." *Journal of Medical Ethics*. doi:10.1136/medethics-2012-101200. Published online April 4, 2013. Available at jme.bmj.com/content/early/2013/02/13/medethics-2012- 101200.full.

Stansberry, Matt, and Kudritzki, Iulian. 2012. *Uptime Institute 2012 Data Center Industry Survey*. Uptime Institute.

Stapledon, Olaf. 1937. *Star Maker*. London: Methuen.

Steriade, M., Timofeev, I., Durmuller, N., and Grenier, F. 1998. "Dynamic Properties of Corticothalamic Neurons and Local Cortical Interneurons Generating Fast Rhythmic (30—40 Hz) Spike Bursts." *Journal of Neurophysiology* 79 (1): 483—90.

Stewart, P. W, Lonky, E., Reihman, J., Pagano, J., Gump, B. B., and Darvill, T. 2008. "The Relationship Between Prenatal PCB Exposure and Intelligence (IQ) in 9-Year-Old Children." *Environmental Health Perspectives* 116 (10): 1416—22.

Sun, W, Yu, H., Shen, Y., Banno, Y., Xiang, Z., and Zhang, Z. 2012. "Phylogeny and Evolutionary History of the Silkworm." Science China Life Sciences 55 (6): 483—96.

Sundet, J., Barlaug, D., and Torjussen, T. 2004. "The End of the Flynn Effect? A Study of Secular Trends in Mean Intelligence Scores of Norwegian Conscripts During Half a Century." *Intelligence* 32 (4): 349—62.

Sutton, Richard S., and Barto, Andrew G. 1998. *Reinforcement Learning: An Introduction*. Adaptive Computation and Machine Learning. Cambridge, MA: MIT Press.

Talukdar, D., Sudhir, K., and Ainslie, A. 2002. "Investigating New Product Diffusion Across Products and Countries." *Marketing Science* 21 (1): 97—1 14.

Teasdale, Thomas W., and Owen, David R. 2008. "Secular Declines in Cognitive Test Scores: A Reversal of the Flynn Effect." *Intelligence* 36 (2): 121—6.

Tegmark, Max, and Bostrom, Nick. 2005. "Is a Doomsday Catastrophe Likely?" *Nature* 438: 754.

Teitelman, Warren. 1966. "Pilot: A Step Towards Man—Computer Symbiosis." PhD dissertation, Massachusetts Institute of Technology.

Temple, Robert K. G. 1986. *The Genius of China: 3000 Years of Science, Discovery, and Invention*. 1st ed. New York: Simon & Schuster.

Tesauro, Gerald. 1995. "Temporal Difference Learning and TD-Gammon." *Communications of the ACM* 38 (3): 58—68.

Tetlock, Philip E. 2005. *Expert Political Judgment: How Good is it? How Can We Know?* Princeton, NI: Princeton University Press.

Tetlock, Philip E., and Belkin, Aaron. 1996. "Counterfactual Thought Experiments in World Politics: Logical, Methodological, and Psychological Perspectives." In *Counterfactual Thought Experiments in World Politics: Logical, Methodological, and Psychological Perspectives*, edited by Philip E. Tetlock and Aaron Belkin, 1—38. Princeton, NJ: Princeton University Press.

Thompson, Adrian. 1997. "Artificial Evolution in the Physical World." In *Evolutionary Robotics: From Intelligent Robots to Artificial Life*, edited by Takashi Gomi, 101—25. ER '97. Carp, ON: Applied AI Systems.

Thrun, S., Montemerlo, M., Dahlkamp, H., Stavens, D., Aron, A., Diebel, J., Fong, P., et al. 2006. "Stanley: The Robot That Won the DARPA Grand Challenge." *Journal of Field Robotics* 23 (9): 661—92.

Trachtenberg, J. T., Chen, B. E., Knott, G. W, Feng, G., Sanes, J. R., Welker, E., and Svoboda, K. 2002. "Long-Term In Vivo Imaging of Experience-Dependent Synaptic Plasticity in Adult Cortex." *Nature* 420 (6917): 788—94.

Traub, Wesley A. 2012. "Terrestrial, Habitable-Zone Exoplanet Frequency from Kepler." *Astrophysical Journal* 745 (1): 1—10.

Truman, James W, Taylor, Barbara J., and Awad, Timothy A. 1993. "Formation of the Adult Nervous System." In *The Development of Drosophila Melanogaster*, edited by Michael Bate and Alfonso Martinez Arias. Plainview, NY: Cold Spring Harbor Laboratory.

Tuomi, Ilkka. 2002. "The Lives and the Death of Moore's Law." *First Monday* 7 (11).

Turing, A. M. 1950. "Computing Machinery and Intelligence." *Mind* 59 (236): 433—60.

Turkheimer, Eric, Haley, Andreana, Waldron, Mary, D'Onofrio, Brian, and Gottesman, Irving I. 2003. "Socioeconomic Status Modifies Heritability of IQ in Young Children." *Psychological Science* 14 (6): 623—8.

Uauy, Ricardo, and Dangour, Alan D. 2006. "Nutrition in Brain Development and Aging: Role of Essential Fatty Acids." Supplement, *Nutrition Reviews* 64 (5): S24—S33.

Ulam, Stanislaw M. 1958. "John von Neumann." *Bulletin of the American Mathematical Society* 64(3): 1—49.

Uncertain Future, The. 2012. "Frequently Asked Questions." The Uncertain Future. Retrieved March 25, 2012. Available at theuncertainfuture.com/faq.html.

US. Congress, Office of Technology Assessment. 1995. *U.S.—Russian Cooperation in Space* ISS-618. Washington, DC: US. Government Printing Office, April.

Van Zanden, Jan Luiten. 2003. *On Global Economic History: A Personal View on an Agenda for Future Research*. Amsterdam: International Institute of Social History, July 23.

Vardi, Moshe Y. 2012. "Artificial Intelligence: Past and Future." *Communications of the ACM* 55 (1): 5.

Vassar, Michael, and Freitas, Robert A., Jr. 2006. "Lifeboat Foundation Nanoshield." Lifeboat Foundation. Retrieved May 12, 2012. Available at lifeboat.com/ex/nanoshield.

Vinge, Vernor. 1993. "The Coming Technological Singularity: How to Survive in the Post-Human Era." In *Vision-21: Interdisciplinary Science and Engineering in the Era of Cyberspace*, 11—22. NASA

Conference Publication 10129. NASA Lewis Research Center.

Visscher, P. M., Hill, W. G., and Wray, N. R. 2008. "Heritability in the Genomics Era: Concepts and Misconceptions." *Nature Reviews Genetics* 9 (4): 255—66.

Vollenweider, Franz, Gamma, Alex, Liechti, Matthias, and Huber, Theo. 1998. "Psycho- logical and Cardiovascular Effects and Short—Term Sequelae of MDMA ('Ecstasy') in MDMA-Naive Healthy Volunteers." *Neuropsychopharmachology* 19 (4): 241—51.

Wade, Michael I. 1976. "Group Selections Among Laboratory Populations of Tribolium." *Proceedings of the National Academy of Sciences of the United States of America* 73 (12): 4604—7.

Wainwright, Martin J., and Jordan, Michael I. 2008. "Graphical Models, Exponential Families, and Variational Inference." *Foundations and Trends in Machine Learning* 1 (1—2): 1—305.

Walker, Mark. 2002. "Prolegomena to Any Future Philosophy." *Journal of Evolution and Technology* 10(1).

Walsh, Nick Paton. 2001. "Alter our DNA or robots will take over, warns Hawking." *The Observer*, September 1. theguardian.com/uk/2001/sep/O2/medicalscience.genetics.

Warwick, Kevin. 2002. *I, Cyborg*. London: Century.

Wehner, M., Oliker, L., and Shalf, J. 2008. "Towards Ultra—High Resolution Models of Climate and Weather." *International Journal of High Performance Computing Applications* 22 (2): 149—65.

Weizenbaum, Joseph. 1966. "Eliza: A Computer Program for the Study of Natural Language Communication Between Man and Machine." *Communications of the ACM* 9 (1): 36—45.

Weizenbaum, Joseph. 1976. *Computer Power and Human Reason: From Judgment to Calculation*. San Francisco, CA: W. H. Freeman.

Werbos, Paul John. 1994. *The Roots of Backpropagation: From Ordered Derivatives to Neural Networks and Political Forecasting*. New York: Wiley.

White, J. G., Southgate, E., Thomson, J. N., and Brenner, S. 1986. "The Structure of the Nervous System of the Nematode *Caenorhabditis Elegans*." *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 314 (1165): 1—340.

Whitehead, Hal. 2003. *Sperm Whales: Social Evolution in the Ocean*. Chicago: University of Chicago Press.

Whitman, William B., Coleman, David C., and Wiebe, William J. 1998. "Prokaryotes: The Unseen Majority." *Proceedings of the National Academy of Sciences of the United States of America* 95 (12): 6578—83.

Wiener, Norbert. 1960. "Some Moral and Technical Consequences of Automation." *Science* 131(3410):1355—8.

Wikipedia. 2012a, s.v. "Computer Bridge." Retrieved June 30, 2013. Available at en.wikipedia.org/wiki/Computer_bridge.

Wikipedia. 2012b, s.v. "Supercomputer." Retrieved June 30, 2013. Available at et.wikipedia.org/wiki/Superarvuti.

Williams, George C. 1966. *Adaptation and Natural Selection: A Critique of Some Current Evolutionary Thought*. Princeton Science Library. Princeton, NI: Princeton University Press.

Winograd, Terry. 1972. *Understanding Natural Language*. New York: Academic Press.

Wood, Nigel. 2007. *Chinese Glazes: Their Origins, Chemistry and Re-creation*. London: A81C Black

World Bank. 2008. *Global Economic Prospects: Technology Diflusion in the Developing World*, 42097. Washington, DC.

World Robotics. 2011. *Executive Summary of 1. World Robotics 2011 Industrial Robots; 2. World Robotics 2011 Service Robots*. Retrieved June 30, 2012. Available at bara.org.uk/pdf/2012/world-robotics/Executive_Summary_WR_2012.pdf.

World Values Survey. 2008. WVS 2005-2008. Retrieved 29 October, 2013. Available at wvsevsdb.com/wvs/WVSAnalizeStudy.jsp.

Wright, Robert. 2001. *Nonzero: The Logic of Human Destiny*. New York: Vintage.

Yaeger, Larry. 1994. "Computational Genetics, Physiology, Metabolism, Neural Systems, Learning, Vision, and Behavior or PolyWorld: Life in a New Context." In *Proceedings of the Artificial Life III Conference*, edited by C. G. Langton, 263—98. Santa Fe Institute Studies in the Sciences of Complexity. Reading, MA: Addison-Wesley.

Yudkowsky, Eliezer. 2001. *Creating Friendly AI 1.0: The Analysis and Design of Benevolent Goal Architectures*. Berkeley, CA: Machine Intelligence Research Institute, June 15.

Yudkowsky, Eliezer. 2002. "The AI-Box Experiment." Retrieved January 15, 2012. Available at yudkowsky.net/singularity/aibox.

Yudkowsky, Eliezer. 2004. *Coherent Extrapolated Volition*. Berkeley, CA: Machine Intelligence Research Institute, May.

Yudkowsky, Eliezer. 2007. "Levels of Organization in General Intelligence." In *Artificial General Intelligence*, edited by Ben Goertzel and Cassio Pennachin, 389—501. Cognitive Technologies. Berlin: Springer.

Yudkowsky, Eliezer. 2008a. "Artificial Intelligence as a Positive and Negative Factor in Global Risk." In *Global Catastrophic Risks*, edited by Nick Bostrom and Milan M. Ćirković, 308—45. New York: Oxford University Press.

Yudkowsky, Eliezer. 2008b. "Sustained Strong Recursion." *Less Wrong* (blog), December 5.

Yudkowsky, Eliezer. 2010. *Timeless Decision Theory*. Berkeley, CA: Machine Intelligence Research Institute.

Yudkowsky, Eliezer. 2011. *Complex Value Systems are Required to Realize Valuable Futures*. Berkeley, CA: Machine Intelligence Research Institute.

Yudkowsky, Eliezer. 2013. *Intelligence Explosion Microeconomics*, Technical Report 2013-1. Berkeley, CA: Machine Intelligence Research Institute.

Zahavi, Amotz, and Zahavi, Avishag. 1997. *The Handicap Principle: A Missing Piece of Darwin's Puzzle*. Translated by N. Zahavi-Ely and M. P. Ely. New York: Oxford University Press.

Zalasiewicz, J., Williams, M., Smith, A., Barry, T. L., Coe, A. L., Bown, P. R., Brenchley, P., et al. 2008. "Are We Now Living in the Anthropocene?" *GSA Today* 18 (2): 4—8.

Zeira, Joseph. 2011. "Innovations, Patent Races and Endogenous Growth." *Journal of Economic Growth* 16 (2): 135—56.

Zuleta, Hernando. 2008. "An Empirical Note on Factor Shares." *Journal of International Trade and Economic Development* 17 (3): 379—90.

ПРИМІТКИ

Передмова

1 Однак не всі примітки містять корисну інформацію.

2 Не знаю, які саме.

Розділ 1

3 Сьогодні прожитковий мінімум становить приблизно 400 доларів (Chen and Ravallion 2010). Отже, прожитковий мінімум мільйона людей становить 400 000 000 доларів. Нинішній світовий валовий продукт — близько 60 000 000 000 000 доларів. Останніми роками він щорічно зростав десь на чотири відсотки (сукупний щорічний приріст від 1950 року, згідно з Maddison [2010]). На цих значеннях ґрунтуються оцінки, наведені в тексті, хоча, звісно, вони дуже приблизні. Якщо ж говорити безпосередньо про кількісні показники зростання населення, то нині населення світу зростає на один мільйон приблизно кожні півтора тижні. Але приріст населення все одно не встигає за зростанням економіки, оскільки дохід на душу населення також збільшується. Від 5000 року до н. е., після аграрної революції, населення світу зростало приблизно на мільйон за кожні двісті років. Це було величезне прискорення у зростанні, адже в доісторичні часи збільшення населення становило мільйон осіб за мільйон років. Проте сучасні цифри теж вражають: зростання економіки, для якого сімсот років тому знадобилося б два століття, тепер займає всього 90 хвилин. А рівень приросту населення світу, який існував двісті років тому, тепер відбувається за півтора тижні. Див. також Maddison (2005).

4 Таке зростання може бути одним зі свідчень імовірного наближення «сингулярності», як передрікав Джон фон Нейман у розмові з математиком Станіславом Улямом: «Ми розмовляли про те, що завдяки постійному пришвидшенню прогресу технологій та темпу життя людства складається враження, що розвиток людства прямує до деякої сингулярності, за якою існування людства у вигляді, у якому ми його знали досі, буде неможливим» (Ulam, 1958).

5 Hanson (2000).

6 Van Zanden (2003); Maddison (1999, 2001); De Long (1998).

7 Vinge (1993); Kurzweil (2005).

8 Sandberg (2010).

9 Дві оптимістичні цитати 1960-х, які часто згадують: «За двадцять років машини зможуть виконувати всю роботу, яку може робити людина» (Simon 1965, 96); «Протягом життя одного покоління … проблему створення штучного інтелекту буде вирішено» (Minsky 1967, 2). Для систематизованої підбірки передбачень щодо ШІ див. Armstrong, Sotala (2012).

10 Наприклад, див. Baum et al. (2011) та Armstrong and Sotala (2012).

11 Це може означати, що дослідники ШІ не надто добре уявляють собі часові межі своєї роботи. Вони можуть як переоцінювати швидкість появи ШІ, так і недооцінювати.

12 Good (1965, 33).

13 Але один з тих, хто допускав, був Норберт Вінер, який мав сумніви щодо можливих наслідків. У 1960 році він писав: «Якщо ми доручаємо свої інтереси механічному посереднику, перервати роботу якого не можемо, бо швидкість його дій надто висока, щоб ми мали змогу їх оцінити до того, як вони завершаться, то нам варто впевнитися, що його цілі справді відповідають нашим, а не лише правдоподібно імітують їх» (Wienner 1960). В інтерв'ю з Памелою Маккордак (у 1979) Ед Фредкін висловлював занепокоєння імовірністю появи суперінтелектуального ШІ. 1970 року Гуд сам пише про ризики і навіть закликає створити асоціацію, яка має протидіяти небезпекам (Good 1970). У його більш пізній статті (Good 1982) він передбачає ідеї «непрямої нормативності», які ми обговорюватимемо в розділі 13). У 1984 році Марвін Мінскі також писав про ключові небезпеки (Minsky 1984).

14 Пор. Юдковський (2008 а). Щодо важливості оцінки етичних аспектів потенційно небезпечних технологій до того, як вони стануть досяжними; див. Roache (2008).

15 Маккордак (1979).

16 Newell et al. (1959).

17 Програма SAINTS, програма ANALOGY і програма STUDENT відповідно. Див. Slagle (1963), Evans (1964, 1968) і Bobrow (1968).

18 Nilsson (1984).

19 Weizenbaum (1966).

20 Winograd (1972).

21 Cope (1996); Weizenbaum (1966); Moravec (1980); Thrun et al. (2006); Buehler et al. (2009); Koza et al. (2003). Департамент автотранспорту штату Невада видав перший дозвіл на автомобіль без водія у травні 2012 року.

22 Система STANDUP (Ritchie et al. 2007).

23 Schwartz (1987). Тут Шварц має на увазі скептицизм, присутній, на його думку, у працях Губерта Дрейфуса.

24 Відомим критиком свого часу був Губерт Дрейфус. Також скептицизм висловлювали Джон Лукас, Роджер Пенроуз та Джон Серль. Проте лише Дрейфус спростовував твердження про те, яких практичних результатів очікувати від наявних парадигм ШІ (втім визнавав, що нові парадигми можуть сягнути далі). Серль здебільшого доскіпувався до функціоналістських теорій філософії розуму, а не інструментальних можливостей ШІ. Лукас і Пенроуз відкидали можливість того, що класичний комп'ютер зможе коли-небудь виконувати все те, що здатна розв'язати людина-математик. Проте не заперечували, що будь-яка окрема функція загалом може бути автоматизована і що ШІ згодом може стати дуже потужним інструментом. Цицерон зауважував «такого абсурду я ще не чув, але так сказав якийсь філософ» (Cicero 1923, 119); проте напрочуд важко назвати хоча б одного відомого науковця, який би заперечував можливість існування штучного суперінтелекту в сенсі, вжитому в цій книжці.

25 Навчання нейронної мережі дещо відрізняється від навчання за допомогою лінійної регресії — статистичного механізму, створеного Адрієном-Марі Лежандром та Карлом Фрідріхом Гауссом на початку 1800-х років.

26 Артур Брайсон та Ю-Ши Хо у 1969 році описали основний алгоритм як метод багатокрокової динамічної оптимізації. Застосувати його в нейронній мережі вперше запропонував у 1974 році Пол Вербос (Paul Werbos 1994), проте лише у 1986 році після праці Девіда Румельхарта, Джеффрі Гінтона та Рональда Вільямсона метод почав помалу завойовувати визнання ширшої наукової спільноти.

27 Раніше демонстрували, що мережі без прихованих шарів мають значні обмеження (Minsky and Papert 1969).

28 Наприклад, MacKay (2003).

29 Murphy (2012).

30 Ми навмисно залишили поза увагою деякі важливі технічні деталі, щоб не переобтяжувати виклад. Ще матимемо нагоду згадати деякі з них у розділі 12.

31 Програма p є описом рядка x, якщо x є результатом виконання p на універсальній машині Тюрінга U. Це записується $U(p) = x$. (Рядок x є представленням імовірного світу). Тоді колмогоровська складність x: $K(x) := \min_p \{l(p) : U(p) = x\}$, де $l(p)$ — довжина p в бітах. А «Соломонова» імовірність x визначається як $M(x) := \sum_{p:U(p)=x} 2^{-l(p)}$, де сума береться по всіх («мінімальних», тобто таких, які необов'язково зупиняються після закінчення виведення x) програмах p, для яких U виводить рядок, що починається з x (Hutter 2005).

32 Баєсова ймовірність за умови спостереження E:

$$P_{\text{апост}}(w) = P_{\text{апр}}(w \mid E) = \frac{P_{\text{апр}}(E \mid w) P_{\text{апр}}(w)}{P_{\text{апр}}(E)}.$$

(Імовірність твердження дорівнює сумі ймовірностей світів, у яких це твердження правдиве).

33 Або випадково вибирає одну з дій, які мають найбільшу очікувану корисність, якщо існує встановлений зв'язок між діями і їхньою корисністю.

34 Якщо коротко, то очікувана корисність дії може бути визначена як $EU(a)=\sum_{w\in\mathbb{W}} U(w)P(w|a)$, де сума взята по всіх можливих світах.

35 Див., наприклад, Howson and Urbach (1993); Bernardo and Smith (1994); Russell and Norvig (2010).

36 Pearl (2009).

37 Wainwright and Jordan (2008). Способів застосування баєсових мереж безліч; див., наприклад, Pourret et al. (2008).

38 Може, ви питаєте: для чого приділяти таку увагу ігровому ШІ — не надто важливій сфері застосування ШІ. Відповідаю: гра — найпростіший спосіб порівняти можливості ШІ з людськими.

39 Newell et al. (1958, 320).

40 Vardi (2012).

41 У 1976 році І. Дж. Гуд писав: «Із появою комп'ютерної програми рівня гросмейстера прийде вік [ультрарозумних машин]» (Good 1976). У 1979 році Дуглас Гофстедтер у своїй книжці «Гедель, Ешер, Бах», за яку він отримав Пулітцера, висловив таку думку: «Питання: Чи з'являться шахові програми, здатні виграти в будь-кого? Відповідь: Ні. Можуть з'явитися програми, які зможуть виграти в будь-кого в шахи, але вони не будуть призначені лише для шахів. Це будуть програми загальної інтелектуальності, і вони будуть такими самими норовливими, як і люди. Хочеш зіграти в шахи? — Ні, шахи це нудно. Поговорімо краще про поезію» (Hofstadter [1979] 1999, 678).

42 Це мінімакс пошук з альфа-бета відсіченням та специфічною для шахів функцією евристичної оцінки ігрових станів. Укомплектований пристойною бібліотекою дебютів, ендшпілів та доброю підбіркою інших хитрощів, такий алгоритм може стати досить здібною шаховою програмою.

43 Однак багато алгоритмів, які він використовує, можуть придатися і для інших ігор, особливо враховуючи нещодавній прогрес у машинному навчанні.

44 Samuel (1959); Schaeffer (1997, ch. 6).

45 Schaeffer et al. (2007).

46 Berliner (1980 a, b).

47 Tesauro (1995).

48 Це програми GNU (див. Silver [2006]) та Snowie (див. Gammoned.net [2012]).

49 Створювати флот Ленату допомагав комп'ютер. Він писав про це так: «Отже, внесок у результат треба поділити 60/40 відсотків між мною та Eurisko, хоч варто підкреслити, що жоден не зміг би виграти без допомоги іншого» (Lenat 1983, 80).

50 Lenat (1982, 1983).

51 Cirasella and Kopec (2006).

52 Kasparov (1996, 55).

53 Newborn (2011).

54 Keim et al. (1999).

55 Див. Armstrong (2012).

56 Sheppard (2002).

57 Wikipedia (2012 a).

58 Markoff (2011).

59 Rubin and Watson (2011).

60 Elyasaf et al. (2011).

61 KGS (2012).

62 Nilsson (2009, 318). Кнут, без сумніву, перебільшував. Існує багато виключно «думальних» завдань, у яких ШІ не демонструє жодних результатів — відкриття нових царин математики, діяльність у сфері філософії, написання детективних романів, планування coup d'ětat (*фр.* — державний переворот), розробка нового цікавого продукту.

63 Shapiro (1992).

64 Імовірно, ШІ не вдалося зрівнятися з людиною у здатностях чуття, моторики, загального розуміння та розуміння мови через те, що людський мозок має спеціалізовані засоби для цієї діяльності — нервові структури, які сформувалися протягом тривалої еволюції нашого виду. Натомість логічне мислення та здатність грати в шахи неприродні для нас. Для виконання цієї діяльності ми задіюємо обмежені засоби пізнавальних механізмів загального призначення. Можливо, виконуючи логічні міркування, наш мозок задіює щось на кшталт «віртуальної машини»: повільну та неоковирну симуляцію комп'ютера загального призначення. У такому разі можна сказати, що класичний програмний ШІ емулює людський розум тією самою мірою, якою людина емулює комп'ютер, коли намагається міркувати логічно.

65 Дещо суперечливий приклад, адже приблизно 20 відсотків дорослих жителів США і десь така сама частина населення багатьох інших розвинених країн вважають, що Сонце обертається навколо Землі (Crabtree 1999; Dean 2005).

66 World Robotics (2011).

67 За оцінками даних наведених у Guizzo (2010).

68 Holley (2009).

69 Також існують гібридні механізми автоматизованого перекладу, проте їх використовують нечасто.

70 Cross and Walker (1994); Hedberg (2002).

71 Згідно зі статистичними даними TABB Group, компанії з філіями в Нью-Йорку та Лондоні, що проводить дослідження ринків капіталу (дані надані особисто).

72 CFTC and SEC (2010). Погляд на події 6 травня 2010 року з іншої перспективи поданий CME Group (2010).

73 Ці слова не варто сприймати як аргументацію проти високочастотної алгоритмічної торгівлі. За нормальних умов така методика здатна позитивно впливати на ліквідність та ефективність ринку.

74 Схожа паніка трапилася на ринку цінних паперів 1 серпня 2012 року, почасти через те, що «запобіжник» не був запрограмований реагувати на екстремальні зміни в *кількості* акцій. Далі ми детальніше говоритимемо про це: складно передбачити всі можливі варіанти, що, по суті, порушуватимуть деяке цілком слушне правило і водночас формально йому не суперечитимуть.

75 Nilsson (2009, 319).

76 Minsky (2006); McCarthy (2007); Beal and Winston (2009).

77 Пітер Норвіг, особисто. Крім того, слідом за окремою хвилею популярності «великих даних» (з подання компаній, як-от Google та Netflix Prize) зросла також популярність курсів з машинного навчання.

78 Armstrong and Sotala (2012).

79 Müller and Bostrom (2016).

80 Див. Baum et al. (2011), де також наведено результати іншого дослідження, а ще Sandberg and Bostrom (2011).

81 Nilsson (2009).

82 Знову ж, за умови відсутності значних катастроф, які можуть знизити цивілізаційний рівень людства. Нільссон наводить таке визначення ШІРЛ: «ШІ, який може виконувати не гірше або й краще за людину 80 відсотків усіх завдань» (Kruel 2012).

83 У таблиці наведено дані чотирьох опитувань та об'єднані результати. Перші два опитування проведені на наукових конференціях: PT-AI, опитування учасників конференції Philosophy and Theory of AI (Філософія і теорія штучного інтелекту) у Салоніках 2011 року (опитування відбулося у 2012), у якому взяли участь 43 із 88 учасників конференції; *AGI,* опитування учасників конференцій Artificial General Intelligence (Штучний загальний інтелект) та Risks of Artificial General Intelligence (Ризики штучного загального інтелекту), проведених в Оксфорді у грудні 2012 року (взяли участь 72/111). EETN — опитування членів Грецької асоціації штучного інтелекту, професійної організації науковців, що мають друковані праці у сфері ШІ, яке відбулося у квітні 2013 року (взяли участь 26/250). Опитування Топ-100 представляє думку сто найбільш цитованих авторів у сфері ШІ станом на травень 2013 року (відповіли 29 зі 100).

84 На момент написання цієї книжки Александер Крюель провів та опублікував на власному інтернет-ресурсі інтерв'ю з 28 науковцями та експертами у сфері ШІ (Kruel 2011).

85 На рисунку наведено унормовані значення медіан. Середні значення суттєво відрізняються. Наприклад, середня оцінка ймовірності «дуже негативного» впливу становила 7,6 відсотка (для Топ-100) і 17,2 відсотка (загальна оцінка об'єднаного пулу експертів).

86 Існує багато літератури про ненадійність експертних прогнозів у багатьох сферах і, на жаль, є всі підстави зараховувати туди галузь штучного інтелекту. Зокрема, науковці схильні перебільшувати свою впевненість у власній версії майбутнього, переоцінювати точність своїх суджень і, відповідно, недооцінювати ймовірність помилки (Tetlock 2005). (Впливають також і інші упередження, див., наприклад, Gilovich et al. [2002]). Проте невпевненість є невід'ємним елементом причин тих чи тих людських рішень і багато з них зумовлені лише очікуваними оцінками довгострокових наслідків, тобто ймовірнісним прогнозуванням. Відмова від відповідальності прогнозування не вирішить епістемічну проблему, а лише тимчасово прибере її з поля зору (Bostrom 2007). Натомість за першої ознаки концентрації впевненості треба розширювати довірчі інтервали (чи «простір впевненості») — наприклад, розмиваючи довірчу функцію, та й взагалі намагатися позбуватися різного роду упереджень, змінювати перспективи та прагнути інтелектуальної чесності. У майбутньому ми могли б розробити методології, навчальні методики та інститути, які б сприяли калібруванню нашого пізнання. Див. також Armstrong and Sotala (2012).

Розділ 2

87 Таке визначення схоже на дане в Bostrom (2003c) та Bostrom (2006a). Його можна порівняти також із визначенням Шейна Легга («Інтелектуальність визначає здатність агента досягати цілей у великій кількості різноманітних середовищ») та його формалізацією (Legg, 2008).

Дуже схоже визначення ультраінтелекту Гуда, наведене в розділі 1 («машини, яка зможе перевершити в інтелектуальній діяльності навіть найрозумнішу людину»).

88 Тому утримаємося від припущень про те, чи може суперінтелектуальна машина мати «правдиві наміри» (з усією повагою до Серля, може, але це виходить за межі теми цієї книжки). Ми також оминемо увагою гарячі філософські дебати про внутрішність/зовнішність розумового вмісту відносно самого розуму та пов'язану з ними тезу «Розширеного розуму» (Clark and Chalmers, 1998).

89 Turing (1950, 456).

90 Turing (1950, 456).

91 Chalmers (2010); Moravec (1976, 1988, 1998, 1999).

92 Див. Moravec (1976). Його думки розвинув Девід Чалмерс (David Chalmers, 2010).

93 Див. також Shulman and Bostrom (2012). Там це питання розглянуто детальніше.

94 Тому, на думку Легга (2008), для відтворення еволюційного прогресу людям знадобиться менше часу й обчислювальних ресурсів (тоді як ресурси, потрібні для повного еволюційного моделювання, на його думку, нам не доступні). Водночас Баум (Baum 2004) вважає, що попередні наукові досягнення можуть допомогти у створенні ШІ, зокрема структура генома є важливим представленням інформації, яке можна використати в еволюційному алгоритмі.

95 Whitman et al. (1998); Sabrosky (1952).

96 Schultz (2000).

97 Menzel and Giurfa (2001, 62); Truman et al. (1993).

98 Sandberg and Bostrom (2008).

99 Див. Legg (2008), де цю думку пояснено детальніше, зокрема досліджено перспективу використання функцій та середовищ, які визначають корисність на основі згладженого ландшафту тестів чистого інтелекту.

100 Щодо систематизації і детального розгляду способів інженерної оптимізації тривалості еволюційної селекції див. Bostrom and Sandberg (2009 b).

101 Аналіз стосується лише нервової системи живих організмів і не враховує потребу симулювати тіла чи віртуальне середовище функціонування як частину функції пристосованості. Імовірно, можна створити таку ефективну функцію пристосованості, якій не потрібно повністю симулювати всю нейронну діяльність організму впродовж повного життєвого циклу, щоб перевірити його придатність. Сучасні програмні ШІ часто розвиваються й існують у дуже абстрактних середовищах (програми доведення теорем функціонують у світах математичних символів, ігрові агенти — у ігрових світах для змагань тощо).

Скептик може наполягати, що абстрактне середовище не зможе належно забезпечити еволюцію загального інтелекту, бо це можливо лише в середовищі, максимально подібному до природного біологічного середовища, у якому еволюціонували наші предки. Його відтворення потребуватиме набагато більше обчислювальних ресурсів, ніж побудова простого ігрового світу чи обмеженого конкретною проблемою світу абстракцій (тоді як вільний доступ до реального фізичного світу «є в пакеті» біологічної еволюції). Якщо ж

обійтися без мікрофізичної точності симуляції не вдасться, потреби в обчислювальних ресурсах можуть зрости до фантастичних масштабів. Однак такий песимізм, найімовірніше, безпідставний. Малоймовірно, що повноцінна еволюція інтелекту можлива лише в середовищі, максимально наближеному до природного. Навпаки, спеціально створене штучне середовище може виявитися набагато ефективнішим стимулом до еволюції розуму, зокрема, потрібних нам характеристик (абстрактне мислення, загальний хист до розв'язання задач, а не швидкість інстинктивних реакцій та оптимізація зору).

102 Вікіпедія (2012 b).

103 Загальний виклад теорії упередження відбору — див. Bostrom (2002 a). Детальний розгляд застосування теорії для цього випадку — див. Shulman and Bostrom (2012). Короткий науково-популярний огляд — Bostrom (2008 b).

104 Sutton and Barto (1998, 21 f); Schultz et al. (1997).

105 Термін запропоновано Елізером Юдковським. Див., наприклад, Eliezer Yudkowsky (2007).

106 Такий сценарій описано Гудом (Good 1965) та Юдковським (Yudkowsky 2007). Проте як альтернативу можна розглянути послідовність ітерацій, у якій деякі етапи не передбачають покращення інтелектуальності, а натомість — оптимізацію та спрощення структури. Так зерно ШІ, змінюючи свою структуру на цих етапах, спрощуватиме собі пошук шляхів для наступних покращень.

107 Helmstaedter et al. (2011).

108 Andres et al. (2012).

109 Тобто сучасного рівня технологій достатньо для створення корисного інструмента пізнання та комунікації. Проте він усе ще суттєво програє можливостям наших м'язів, органів чуття та тіла загалом.

110 Sandberg (2013).

111 Див. розділ «Вимоги до комп'ютера» у Sandberg and Bostrom (2008, 79–81).

112 Частково успішною емуляцією мозку можна вважати систему, яка демонструє мікроактивність, подібну до біологічної, а також відтворює значну частину процесів, характерних для виду, зокрема повільний сон або нейропластичність, яка обумовлена діяльністю. Це ще не емуляція цілого мозку — принаймні поки такий мозок не здатний виконувати більшість когнітивних функцій свого прототипу. Але вже дуже цінний об'єкт для нейронаукових досліджень (а також потенційне джерело великих етичних проблем). Узагалі, щоб емуляція вважалася повною і успішною, можна сказати, що вона має бути здатна зв'язно висловлювати думки або мати змогу цього навчитися.

113 Sandberg and Bostrom (2008). Детальне пояснення можна знайти у звіті.

114 Sandberg and Bostrom (2008).

115 Першу таку карту описано в Albertson and Thomson (1976) та White et al. (1986).

116 Огляд попередніх спроб емуляції *C. elegans* та їхньої дальшої долі — див. Kaufman (2011). Кауфман цитує слова одного амбітного докторанта, Девіда Делраймпла: «З появою оптогенетичних методик можливість зчитувати та записувати нейронні стани живої *C. elegans* за допомогою високошвидкісної автоматизованої системи вже не здається такою

фантастичною... Я думаю, що за два-три роки емуляція *C. elegans* уже буде пройденим етапом. Хай там як, але якщо до 2020 року цього не станеться, я буду дуже здивований» (Dalrymple 2011). Запрограмовані вручну моделі мозку (а не згенеровані автоматично) уже досягли базової функціональності — див., наприклад, Eliasmith et al. (2012).

117 Вид Caenorhabditis elegans справді має низку важливих переваг. Наприклад, її організм повністю прозорий. Крім того, схема нейронних зв'язків однакова у всіх особин цього виду.

118 Якщо ж кінцевим продуктом буде нейроморфний ШІ, а не емуляція, то важливі відкриття можуть з'явитися не внаслідок спроб симуляції мозку *людини*. Так, цілком імовірно, що вивчення (нелюдського) тваринного мозку дасть змогу відкрити важливі особливості будови кортикальних структур. З мозком деяких тварин працювати може бути зручніше. Крім того, тваринний мозок може бути меншим, що дасть змогу зекономити обчислювальні ресурси та скоротити етап сканування. Також до роботи з тваринним мозком може бути менше зауважень з боку регуляторних органів. Може навіть статися, що ШІ людського рівня стане результатом покращення цифрової емуляції мозку тварини. Так, урешті-решт, у лабораторної миші чи макаки може з'явитися шанс відплатити людству сповна.

119 Uauy and Dangour (2006); Georgieff (2007); Stewart et al. (2008); Eppig et al. (2010); Cotman and Berchtold (2002).

120 Згідно з даними Всесвітньої організації охорони здоров'я від 2007 року, близько двох мільярдів людей страждає через нестачу йоду в раціоні (The Lancet 2008). Сильна нестача йоду пригнічує розумовий розвиток і призводить до кретинізму, тобто, по суті, зниження рівня IQ у середньому на 12,5 одиниці (Qian et al. 2005). Зарадити цьому з мінімальними витратами можна за допомогою йодування солі (Horton et al. 2008).

121 Bostrom and Sandberg (2009 a).

122 Bostrom and Sandberg (2009 b). Тести робочої пам'яті, уваги й інші свідчать про нібито здатність фармакологічних засобів та харчових добавок покращити розумові характеристики на 10–20 відсотків. Однак важко достеменно встановити, що такі результати справді спровоковані впливом зазначених засобів, стійкі та свідчать про таке ж покращення розумових можливостей у вирішенні реальних, а не тестових завдань (Repantis et al. 2010). Адже в деяких випадках зростання одних показників може відбуватися череззниження інших, які не охоплені тестами (Sandberg and Bostrom 2006).

123 Якби існував легкий спосіб покращення інтелекту, еволюція уже використала б його. Тому найбільш перспективним напрямом пошуку новітніх ноотропних препаратів може бути шлях, який має виразні еволюційні недоліки, наприклад, збільшення розміру голови у новонародженного або збільшене споживання мозком глюкози. Для детальнішого розгляду цієї ідеї (з декількома важливими зауваженнями) див. Bostrom (2009 b).

124 Сперматозоїди не надаються до скринінгу, бо складаються з однієї клітини, яка руйнується під час секвенування. Ооцити теж складаються з однієї клітини. Щоправда, перший і другий поділ ооцита дають дочірні клітини дуже малого розміру з мінімумом цитоплазми — полярні тільця. Вони мають той геном, що й основна яйцеклітина, тому їх можна використовувати для скринінгу (Gianaroli 2000).

125 Етичність усіх цих процедур свого часу ставили під сумнів, однак, здається, зараз існує тенденція до їх прийняття. Ставлення до генної інженерії та ембріональної селекції в різних культурах дуже різне. Тому, незважаючи на обережність окремих країн, нові технології та методи все одно, зрештою, з'являтимуться та застосовуватимуться, але моральний, релігійний і політичний тиск однозначно впливатиме на їхню швидкість.

126 Davies et al. (2011); Benyamin et al. (2013); Plomin et al. (2013). Див. також Mardis (2011); Hsu (2012).

127 Успадковуваність у широкому розумінні показника IQ дорослого представника середнього класу розвинутої країни зазвичай оцінюється на рівні 0,5–0,8 (Bouchard 2004, 148). Успадковуваність у вузькому розумінні, яка стосується тої частки варіативності, що зумовлена адитивними генетичними факторами, менша (0,3–0,5), але все-таки суттєва (Delvin et al. 1997; Davies et al. 2011; Visscher et al. 2008). У різних середовищах та для різних популяцій ці цифри відрізняються (Benyamin et al. 2013; Turkheimer et al. 2003). Численні фактори впливу на варіативність розумового потенціалу досліджено в Nisbett et al. (2012).

128 Наступні кілька абзаців базуються на ґрунтовній спільній роботі з Карлом Шульманом (Shulman and Bostrom 2014).

129 Через нестачу інформації про впливи адитивних генетичних варіацій на розумові здібності ефективність буде нижчою. Однак навіть незначне розширення про це знань матиме велике значення, бо при селекції залежність приросту ознаки від частки передбачуваної дисперсії не лінійна. Натомість, ефективність селекції пов'язана зі стандартним відхиленням передбачуваного середнього значення IQ, яке пропорційне квадратному кореню з дисперсії. Наприклад, за можливості визначати всього 12,5 відсотка дисперсії, можна досягти половини ефективності випадку селекції, який наведений у таблиці 5 і який передбачає, що визначені цілих 50 відсотків дисперсії. Для порівняння, в одному з нещодавніх досліджень (Rietveld et al. 2013) автори стверджують, що ідентифікували 2,5 відсотка дисперсії.

130 Для порівняння, зараз поширеною практикою є утворення менше ніж десять ембріонів.

131 Таблицю взято з Shulman and Bostrom (2014). Вона базується на спрощеній моделі, у якій рівні IQ ембріонів мають Гаусів розподіл зі стандартним відхиленням 7,5 одиниці. Розумове покращення, яке може спостерігатися серед різної кількості нащадків, визначається відомими нам адитивними генетичними варіаціями. Коефіцієнт інбридингу серед потомства становить 0,5, а загальні адитивні генетичні варіації дають до половини дисперсії в інтелектуальності дорослих особин (Davies et al. 2011). На основі цих двох фактів можна зробити висновок, що фактичне стандартне відхилення інтелектуальності популяції розвинених країн у 15 одиниць зумовлене стандартним відхиленням генетичних впливів, у групі ембріонів — не більше ніж 7,5 одиниці.

132 Як ембріональні, так і дорослі стовбурові клітини можуть розвинутися у сперматозоїди та ооцити, які після цього можна використовувати для запліднення й утворення ембріонів (Nagy et al. 2008; Nagy and Chang 2007). Прекурсори яйцеклітини можуть також утворювати партеногенетичні бластоцисти — незапліднені та нежиттєздатні зародки, які, проте,

спроможні дати лінії зародкових стовбурових клітин для продовження процесу (Maiet et al. 2007).

133 Цитату взято з [Cyranoski (2013)], а належить Кацухіко Хаяші. У 2008 році Hinxton Group, міжнародний науковий консорціум, що займається питаннями етики й іншими проблемами, пов'язаними з використанням стовбурових клітин, передбачив, що отримання гамет зі стовбурових клітин стане можливим за десять років, і досі темп досліджень не давав причин у цьому сумніватися. (Станом на 2019 рік у науковому середовищі досі точаться суперечки щодо етичності такої репродуктивної технології. Водночас завдяки їй одностатева пара лабораторних мишей уже отримала перше потомство (statnews.com/2019/06/05/creating-eggs-sperm-stem-cells/) — *Прим. пер.*).

134 Sparrow (2013); Miller (2012); The Uncertain Future (2012).

135 Sparrow (2013).

136 Більш практичні аргументи проти селекції можуть зосереджуватися навколо непередбачуваного впливу використання селекції на соціальну нерівність, ставити під питання медичну безпечність такої практики, висловлювати стурбованість, що бажання покращень може набрати в суспільства нездорових форм, турбуватися, що не вдасться віднайти правильний баланс прав і обов'язків батьків стосовно майбутнього потомства. Крім того, чисто наукову атмосферу селекції завжди отруюватимуть духи XX століття: євгенічні практики, турбота про людську гідність і дозволені межі втручання держави в репродуктивні справи її громадян. (Для глибшого розгляду етики розумового покращення див. Bostrom and Ord (2006), Bostrom and Roache (2011) та Sandberg and Savulescu (2011)). У межах окремих релігійних традицій можуть виникнути додаткові зауваження стосовно морального статусу ембріонів та меж дозволеного втручання людини в промисел творення.

137 Щоб подолати негативні наслідки інбридингу, знадобиться або значний початковий запас донорів, або суттєві витрати селективного потенціалу на відсіювання шкідливих рецесивних алелей. Так чи інакше нащадки в результаті будуть менше споріднені зі своїми батьками (але більш споріднені поміж собою).

138 Узято з Shulman and Bostrom (2014).

139 Bostrom (2008 b).

140 І наразі невідомо, наскільки міцним горішком виявиться епігенетика (Chanson et al. 2011; Iliadou et al. 2011).

141 Хоч успадковуваність розумових здібностей не викликає сумнівів, може виявитися, що не існує спільних для всіх людей алалей чи поліморфічних механізмів, здатних самотужки визначати розумові здібності нащадків (Davis et al. 2010; Davis et al. 2011; Rietveld et al. 2013). З удосконаленням методик секвенування можна буде легше простежувати кореляції ментальних та поведінкових ознак з низькочастотними й рідкісними алелями. Деякі теоретичні дослідження свідчать, що певні алелі, які в гомозиготах спричиняють генетичні розлади, у гетерозиготах можуть спричинити відчутне розумове покращення, що дало підставу передбачити вищий приблизно на п'ять одиниць рівень IQ у гетерозигот із

хворобою Гоше, Тея — Сакса та Німана — Піка (Cochran et al. 2006). Час покаже правдивість цієї гіпотези.

142 За оцінками авторів однієї із статей (Nachman and Crowell 2000), 175 мутацій генома за покоління. Інше дослідження (Lynch 2010), використовуючи інші методи, розрізняє від 50 до 100 нових мутацій на кожного новонародженого. Конг із колегами (Kong et al. 2012) наводить цифру близько 77 нових мутацій у кожному поколінні. Більшість таких мутацій не впливає на функціонування або такий вплив майже непомітний, але багато таких непомітних впливів у комплексі можуть стати причиною значного порушення. Див. також Crow (2000).

143 Crow (2000); Lynch (2010).

144 Ця ідея має важливі застереження. Модифікований геном може потребувати певного налаштування. Наприклад, для функціонування з певною заданою ефективністю генома, як єдиної системи, його окремі частини можуть потребувати додаткової адаптації. Збільшення ефективності функціонування цих частин може зумовити перевищення дозволеного мінімуму за іншими важливими показниками, зокрема метаболізму.

145 Композиції створені Майком Майком з фотографій, знятих Virtual Flavius (Mike 2013).

146 Насправді результати можна отримати раніше — насамперед, якщо змінити очікування людей.

147 Louis Harris & Associates (1969); Mason (2003).

148 Kalfoglou et al. (2004).

149 Звісно, інформації мало, але особи, відібрані під час лонгітюдного дослідження 1-з-10 000 результатів розумових тестів дітей значно частіше ставали професорами, власниками патентів, успішними бізнесменами, ніж ті, хто мав більш посередні результати (Kell et al. 2013). Ен Роу (1953) дослідила рівень інтелекту шістдесяти чотирьох видатних науковців і визначила, що їхній середній рівень інтелекту вищий на 3–4 стандартних відхилення від середнього по популяції і значно перевищує рівень, притаманний решті науковців. (Крім того, було виявлено кореляції розумових здібностей із заробітком та з нематеріальними показниками, як-от імовірна тривалість життя, кількість розлучень, імовірність виключення зі школи тощо (Deary 2012)). Зміщення вгору розподілу розумових здібностей матиме непропорційно значні наслідки в його хвостах: зросте кількість обдарованих особистостей та зменшиться кількість осіб з проблемами ментального розвитку. Див. також Bostrom and Ord (2006) та Sandberg and Savulescu (2011).

150 Наприклад Warwick (2002). На думку Стівена Гокінґа, людина змушена буде вдатися до використання нейроінтерфейсів, якщо хоче відповідати високим темпам розвитку ШІ: «Маємо якомога швидше розвивати технології прямого з'єднання між мозком та комп'ютером, щоб штучний мозок допомагав людському інтелекту, а не протистояв йому». (Walsh, 2001). Із цим погоджується Рей Курцвейл: «Щодо рекомендації Гокінґа … а саме, щодо прямого з'єднання між мозком і комп'ютером, я згоден з його розумністю, бажаністю та невідворотністю. [*sic*] Я давно вже це говорю» (Kurzweil, 2001).

151 Див. Lebedev and Nicolelis (2006); Birbaumer et al. (2008); Mak and Wolpaw (2009); та Nicolelis and Lebedev (2009). Більш особистий погляд на проблему покращення за допомогою імплантів можна знайти у Chorost (2005, розділ 11).

152 Smeding et al. (2006).

153 Degnan et al. (2002).

154 Dagnelie (2012); Shannon (2012).

155 Perlmutter and Mink (2006); Lyons (2011).

156 Koch et al. (2006).

157 Shalk (2008). Загальний огляд поточного стану галузі — див. Berger et al. (2008). Варіант застосування технології для покращення розуму — див. Warwick (2002).

158 Кілька прикладів: Bartels et al. (2008); Simeral et al. (2011); Krusienski and Shih (2011); та Pasqualotto et al. (2012).

159 Наприклад, Hinke et al. (1993).

160 Цьому твердженню існують часткові винятки, особливо коли це стосується попереднього ставлення даних органів чуття. Наприклад, первинна зорова кора ретинотопічна за своєю організацією, тобто сусідні ділянки зорової кори отримують дані з сусідніх ділянок сітківки (проте очні домінантні колонки дещо ускладнюють цю відповідність).

161 Berger et al. (2012); Hampson et al. (2012).

162 Деякі імпланти потребують двобічного навчання: навчання імпланта розуміти нейронні представлення організму та навчання організму навичок керування системою за допомогою певних послідовностей нейронної активності (Carmena et al. 2003).

163 Спробуємо вважати колективні сутності (корпорації, спілки, уряди, церкви тощо) агентами штучного інтелекту, сутностями, що мають рецептори й ефектори та здатні навчатися, міркувати й діяти (наприклад, Kuipers (2012); пор. Huebner (2008) — щодо можливості існування колективного представлення). Вони однозначно потужні й екологічно успішні, хоча їхні здібності та внутрішні стани відрізняються від людських.

164 Hanson (1995, 2000); Berg and Rietz (2003).

165 Роботодавці, наприклад, можуть використовувати детектор брехні на робочих місцях, щоб боротися з крадіжками та дармоїдством — наприкінці кожного робочого дня перевіряти працівників, чи не вкрали вони щось і чи старанно працювали протягом дня. Так само можна перевіряти політичних високопосадовців та топ-менеджерів, запитуючи, наскільки щиро вони переймаються інтересами акціонерів чи громадян. Диктатори можуть так шукати заколотників серед наближених осіб або схильних до бунту громадян.

166 Можна уявити способи розпізнавання нейрообразів мотивованого розуму за допомогою нейрофотографування. Без детектора самообману, детектор брехні не відсіюватиме людей, які щиро впевнені у своїй правоті. Тестування самообману може допомогти тренувати раціональність та визначати ефективність спроб позбутися від упереджень.

167 Bell and Gemmel (2009). Однією з перших спроб є робота Деба Роя (з Массачусетського технологічного інституту), який записав кожен момент перших трьох років життя свого сина. Аналіз цих даних дає змогу дослідити формування мови в людини. Див. Roy (2012).

168 Зростання популяції біологічного виду людей дасть лише частину зростання світової популяції. Сценарії появи ШІ можуть забезпечити різке зростання світової популяції (зокрема цифрових розумів) у багато разів за короткий проміжок часу. Але такий рух до суперінтелекту містить створення штучного інтелекту або емуляції цілого мозку, тому немає потреби розглядати його в наступному підрозділі.

169 Vinge (1993).

Розділ 3

170 Вернор Віндж для позначення такого пришвидшеного людського розуму вжив термін «слабкий суперінтелект» (Vinge 1993).

171 Наприклад, якби дуже розумна система могла робити все, що й людина, але не могла танцювати мазурку, вона б все одно залишалася суперінтелектом. Нас цікавлять насамперед економічно та стратегічно важливі розумові здібності.

172 Якщо порівняти швидкість процесів та енергетичні витрати мозку з відповідними процесами електронного оброблення інформації, то зрозуміло, що розумові процеси можна пришвидшити щонайменше в мільйон разів. Швидкість роботи нейронних зв'язків більш як у мільйон разів менша від швидкості світла. Синапси розсіюють більш як у мільйон разів більше енергії, ніж потрібно з погляду термодинаміки. Швидкість сучасних транзисторів більш як у мільйон разів вища від швидкості нейронів (Yudkowsky [2008 a]; див. також Drexler [1992]). Швидкість «швидкого суперінтелекту» обмежена швидкістю світла, квантовим обмеженням швидкості зміни стану та габаритами його матеріального втілення (Lloyd 2000) «Найсучасніший ноутбук», описаний Ллойдом, зі швидкістю $1{,}4 \cdot 10^{21}$ FLOPS виконуватиме емуляцію мозку з прискоренням у $3{,}8 \cdot 10^{29}$ разів (якщо, звісно, виконання емуляції вдасться паралелізувати). Проте Ллойд не намагався створити технічно можливий концепт. Він лише хотів показати ті фундаментальні обмеження, які накладають базові закони фізики.

173 У разі емуляції цікавить також, як довго людина може працювати над чимось, перш ніж збожеволіє чи втратить цікавість? Невідомо, чи загалом людина здатна прожити тисячі суб'єктивних років без психічних розладів, навіть за умови регулярної зміни видів діяльності та відпочинку. Ба більше, з обмеженою пам'яттю — завдяки певній максимальній кількості нейронів — людина не може навчатися безкінечно: рано чи пізно мозок переповниться знаннями і вона почне забувати те, що колись вивчила. (Штучний інтелект потенційно може зарадити цим проблемам).

174 Відповідно, наномеханізми, які рухаються зі скромною швидкістю 1 м/с мають наносекундні масштаби часу. Див. розділ 2.3.2 роботи Drexler (1992). Робін Генсон описує робота «дінь-дінь» розміром 7 мм, який здатний рухатися зі швидкістю у 260 разів вищою від звичайної (Hanson 1994).

175 Hanson (2012).

176 У контексті поняття «колективний інтелект» під паралелізацією розуміють не апаратну паралелізацію — на низькому рівні обчислювального обладнання, — а паралелізацію

діяльності на рівні окремих інтелектуальних агентів, подібних до людських істот. Виконання однієї емуляції на високопаралелізованому комп'ютері може дати приріст у швидкості (швидкий суперінтелект), але не буде колективним інтелектом.

177 Покращення швидкодії чи якості окремих компонентів може опосередковано вплинути і на швидкість колективного інтелекту, але ми розглядаємо такі покращення переважно в контексті інших двох типів суперінтелекту.

178 Є думка, що збільшення щільності населення спровокувало початок революції пізнього палеоліту, а з досягненням певної порогової щільності накопичувати культурні цінності стало ще легше (Powell et al. 2009).

179 А як же інтернет? Здається, що він ще не досяг своєї межі розумності. Може, скоро це трапиться. Інші згадані нами приклади мережевих утворень розвивалися століттями і навіть тисячоліттями, перш ніж їхній потенціал проявився сповна.

180 Звісно, це не реалістичний розумовий експеримент. Планета такого розміру, щоб вмістити сім квадрильйонів людських організмів із сучасними технологіями зруйнується під тиском власної ваги. Інакше вона має бути утворена з дуже легкої матерії або бути порожнистою і триматися завдяки тиску або інших штучних засобів. (Сфера Дайсона чи порожниста планетарна мегаструктура підійшла б краще). Історія розвитку життя на такій величезній поверхні безсумнівно відрізнялася б від нашої. Але облишмо ці деталі.

181 Тут ми зосередилися на функціональних характеристиках колективного інтелекту, а не на тому, чи зможе він відчувати, чи буде розумом здатним до досвіду суб'єктивної свідомості. (Хоч питання, яким може бути досвід свідомості в більш-менш інтегрованого мозку, порівнюючи з людським, дуже цікаве. Якщо розуміти свідомість у контексті теорії глобального робочого простору (GWT, global workspace theory. — *Прим. пер.*), свідомість інтегрованого мозку може бути значно місткішою. Пор. Baars [1997], Shanahan [2010] та Schwitzgebel [2013]).

182 Навіть невеликі соціуми, які деякий час перебувають в ізоляції, можуть продовжувати користуватися інтелектуальними плодами більшого колективного інтелекту. Наприклад, мова, яку вони продовжують використовувати, може бути наслідком розвитку більшої спільноти, як і інші інструменти, що могли також з'явитися завдяки розвитку більшого колективу людей, до того, як група опинилася в ізоляції. Проте навіть якщо ізоляція перманентна, популяція групи все одно є частиною більшої спільноти — колективного інтелекту не тільки живих членів спільноти, але також усіх попередніх поколінь: утворення, що працює як інформаційна система прямого поширення.

183 Відповідно до тези Черча — Тюрінга всі обчислювані функції можуть бути виконані машиною Тюрінга. Оскільки всі три форми суперінтелекту можуть симулювати машину Тюрінга (за умови необмеженої пам'яті та часу роботи), то формально є обчислювально еквівалентними. Справді, звичайна людина (маючи необмежену кількість місця для записів та необмежений час) також може функціонувати як машина Тюрінга, тобто також є еквівалентною їй. Але важливо, що ці такі різні системи можуть зробити на практиці — з обмеженою пам'яттю та часом. А різниця є, і дуже значна, тому можна відразу зробити

висновки. Наприклад, можна навчити людину з рівнем IQ у 85 одиниць працювати, як машина Тюрінга. (Власне, навчити цього вдалося б, імовірно, навіть достатньо обдарованого та здібного шимпанзе). Проте така людина, найімовірніше, нездатна самотужки, скажімо, створити загальну теорію відносності чи вибороти медаль Філдса.

184 Традиції усної розповіді можуть зумовити появу видатних праць (як-от епос Гомера), але деякі зі співавторів можуть бути особливо обдарованими.

185 Хіба що якийсь із його елементів має швидкий чи якісний суперінтелект.

186 Однак складність може бути почасти зумовлена саме браком спроб: яка користь із точного переліку завдань, які жодна людина чи організація на землі наразі не можуть виконати? Може бути, що концептуалізувати собі ці завдання — якраз і є однією із поки що недоступних нам справ.

187 Пор. Boswell (1917); див. також Walker (2002).

188 Така частота роботи властива лише деяким нейронам, і трапляється короткими серіями. Більшість же працює в більш спокійному темпі (Gray and McCormick, 1996; Steriade et al. 1998). Частоти збудження деяких нейронів (відомих як «балакучі» нейрони («chattering neurons») або клітини, що генерують швидкі ритмічні імпульси («fast rhythmically bursting» cells, FRB cells)) можуть досягати 750 Гц.

189 Feldman and Ballard (1982).

190 Провідність залежить від товщини аксона (товщі аксони — швидші) та від наявності в нього мієлінової оболонки. У центральній нервовій системі затримки поширення збудження можуть становити від близько мілісекунди до сотні (Kandel et al. 2000). Швидкість сигналу в оптоволокні становить десь 68 відсотків *c* (через фізичні властивості матеріалу). У звичайному кабелі електромагнітне поле поширюється з приблизно такою самою швидкістю — 59–77 відсотків *c*.

191 Якщо вважати, що швидкість сигналу 70 відсотків *c*. При 100 відсотках розмір збільшується до $1{,}8 \cdot 10^{18}$ м^3.

192 За сучасними оцінками мозок дорослої людини містить 86,1 ± 8,1 мільярда нейронів. Щоб порахувати нейрони, в дослідному мозку розчинили клітинні мембрани, потім фракціонували клітинні ядра та підрахували кількість тих ядер, які містили специфічні для нейронів маркери. Раніше вважалося, що мозок містить від 75 до 125 мільярдів нейронів. Такі оцінки базувалися на ручному підрахунку нейронів у невеликих репрезентативних ділянках (Azevedo et al. 2009).

193 Whitehead (2003).

194 Для обчислень і зберігання даних інформаційні системи можуть використовувати процеси молекулярного рівня та досягати щонайменше планетарного розміру. Фізичні межі обчисленням установлюють квантова механіка, загальна теорія відносності та термодинаміка, втім вони значно перевищують рівень «мозок-розміром-як-Юпітер» (Sandberg 1999; Lloyd 2000).

195 Stansberry and Kudritzki (2012). Усі датацентри світу використовують лише 1,1–1,5 відсотка електроенергії (Koomey 2011). Див. також Muehlhauser and Salamon (2012).

196 Це спрощення. Місткість робочої пам'яті залежить як від завдання, так і від того, що запам'ятовувати. Проте вона, безперечно, обмежена. Див. Miller (1956) та Cowan (2001).

197 Прикладом може бути той факт, що складність вивчення двійкових понять (визначених за допомогою логічних правил) пропорційна довжині еквівалентної їм найкоротшої пропозиційної формули. Зазвичай дуже складно вивчити формули завдовжки від 3–4 літералів. Див. Feldman (2000).

198 Див. Landauer (1986). Це дослідження базується на експериментальних оцінках швидкостей навчання та забування у людей. Якщо врахувати наявне навчання, то значення може виявитися дещо вищим. Якщо інформаційну ємність одного синапсу вважати за один біт, то верхня межа людського мозку становитиме 10^{15} бітів. Огляд різних оцінок наведено в додатку A до Sandberg and Bostrom (2008).

199 Шум аксонових каналів може спричинити самовільне збудження. Синаптичний шум також може зумовити значні варіації сигналу, який передає синапс. Схоже, еволюція була змушена піти на значні компроміси у процесі формування нервової системи — у стійкості до шумів, а також у витратах (маси, розміру, часових затримок); див. Faisal et al. (2008). Наприклад, аксон не може бути тонше від 0,1 мкм, інакше випадкові відкриття іонних каналів провокуватимуть спонтанне збудження.

200 Trachtenberg et al. (2002).

201 За оцінками пам'яті та обчислювальної потужності, але не енергоефективності. Найшвидший суперкомп'ютер світу (на час написання цієї книжки) — китайський Tianhe-2, — який обігнав свого попередника Titan (Cray Inc.) у червні 2013 року з піковою обчислювальною потужністю 33,86 петафлопс, використовує 17,6 МВт енергії, що на шість порядків більше за мозок (який споживає приблизно 20 Вт).

202 Варто зауважити, що майже всі наведені нами переваги машин мають однакову вагу. Навіть якщо деякі з наведених аргументів виявляться хибними, перевага машин зберігатиметься, поки справджуватиметься бодай один із наведених факторів.

Розділ 4

203 Може бути важко зафіксувати конкретний момент проходження системи через названі рівні. Просто упродовж якогось періоду поступово почне з'являтися все більше завдань і сфер діяльності, у яких система перевершуватиме команду науковців, що працюватимуть з нею.

204 Протягом останніх 50 років був принаймні один сценарій швидкого припинення існування поточного світового порядку, ймовірність якого визнавала більшість людей: термоядерна війна.

205 Це підтверджує спостереження недавньої тенденції скасування ефекту Флінна — стабільного зростання рівнів IQ на 3 одиниці щодекади протягом останніх 60 років у популяціях високорозвинених країн, як-от Об'єднане Королівство, Данія і Норвегія (Teasdale and Owen 2008; Sundet et al. 2004). Досі точаться суперечки навколо справжніх причин існування ефекту Флінна в минулому — чи відображав він справжнє зростання

інтелекту (якщо так, то якою мірою), чи був наслідком покращення навичок розв'язування IQ-тестів. Навіть якщо причиною (хоча б частковою) було справжнє зростання розумових здібностей і тепер відбувається протилежний процес, це не означає, що тепер спостерігається ефект зменшення віддачі першопричин зростання. Натомість причиною зниження може виявитися незалежний несприятливий фактор, дія якого була б ще помітнішою, якби не ефект Флінна.

206 Bostrom and Roache (2011).

207 Генна терапія в соматичних клітинах могла б усунути затримку дозрівання, проте така процедура технічно складніша від маніпуляцій з генотипом зародків та має обмежені можливості.

208 Середня економічна продуктивність глобальної економіки протягом 1960–2000 років зростала за рік у середньому на 4,3 відсотка (Isaksson 2007). Зростання організаційної ефективності є лише однією з компонентів цього зростання. Поза сумнівом, деякі окремі мережі та організаційні процеси зростають значно швидше за інші.

209 Еволюція біологічного мозку мала багато обмежень та компромісів. З перенесенням розумової діяльності на цифровий субстрат, більшості з них можна буде уникнути. Наприклад, фізичний розмір мозку обмежений розміром голови. Із збільшенням розміру голови виникають проблеми з проходженням через родовий канал. Великий мозок потребує інтенсивнішого метаболізму та знижує рухливість. Через стеричні обмеження проходження сигналів по такому мозку ускладнюється — обсяг білої речовини значно перевищує обсяг сірої речовини. Ускладнюється протік крові, що може призвести до проблем із розсіюванням тепла. Ба більше, біологічні нейрони потребують захисту, підтримки, постачання гліальних клітин та крові, а все це потребує додаткового місця в черепній коробці. Див. Bostrom and Sandberg (2009 b).

210 Yudkowsky (2008 a, 326). Дещо оновлений погляд на це питання — див. Yudkowsky (2013).

211 Для простоти: розумність на рисунку подано як одновимірну шкалу. Проте це не є обов'язковою умовою ілюстрації цієї думки. Розумові здібності можна представити як гіперповерхню в багатовимірному просторі.

212 Lin et al. (2012).

213 Збільшенням кількості інтелектуальних елементів, які складають колективний інтелект, можна досягти певного зростання інтелектуальності. Принаймні це забезпечить зростання загальної швидкості виконання завдань, які можна легко паралелізувати. Для того щоб повною мірою експлуатувати інтелектуальний потенціал нової кількості елементів, треба відповідно збільшити рівень координації системи.

214 У разі не нейроморфного ШІ різниця між швидким та якісним суперінтелектом дещо розмита.

215 Rajab et al. (2006, 41–52).

216 Існує думка, що використання програмованих мікросхем (а саме FPGA — програмована користувачем вентильна матриця) для створення нейронної мережі може збільшити швидкість її роботи на два порядки (Markram, 2006). Проведення високоточного

моделювання клімату на базі систем із потужністю, що вимірюється в петафлопсах, показало, що використання спеціалізованих чипів із вбудованими процесорами може здешевити виготовлення в 24–34 рази та на два порядки зменшити енерговитрати (Wehner et al. 2008).

217 Nordhaus (2007). Існує багато інтерпретацій закону Мура; див., наприклад, Tuomi (2002) та Mack (2011).

218 Якщо швидкість розроблення низька, проект може мати проміжний успіх в інших сферах знань: досягненнями в комп'ютерних науках чи в електроніці.

219 Алгоритмічне переважування на практиці малоімовірне, але винятком може бути поява екзотичного обладнання. Наприклад, квантових комп'ютерів, які зможуть виконувати недоступні досі алгоритми. Нейронні мережі та глибоке навчання теж можна вважати прикладом алгоритмічної переваги: математика, яка лягла в їхню основу, тривалий час була надто складна для комп'ютерів, тому ці механізми майже не використовувалися, але з появою швидких графічних процесорів про них згадали. Тепер вони в тренді і очолюють прогрес.

220 $\mathfrak{D}$світу — частина загальної оптимізаційної сили світу, яка спрямована на покращення конкретної системи. Для проекту, що розвивається в повній ізоляції, без жодного впливу ззовні, ми вважаємо $\mathfrak{D}$світу ≈ 0, хоч безсумнівно, що проект не міг би з'явитися без зовнішнього впливу, зумовленого комплексною дією світової економіки та її багаторічного розвитку (комп'ютери, наукові концепції, персонал тощо).

221 Найважливіша здібність для ШІ — самовдосконалення розуму, тобто посилення власного інтелекту. (Якщо зерно ШІ зможе покращувати іншу систему, яка призначена для інтелектуального вдосконалення зерна ШІ, тоді можемо розглядати їх як підсистеми однієї системи й аналізувати як одне ціле).

222 Уважатимемо, що консервативність системи не настільки велика, щоб відлякати інвесторів від проекту.

223 Схожий приклад наведений у Yudkowsky (2008 b).

224 Розмір вкладень значно зріс (тобто суми інвестицій у металургію, зайнятість людей у напівпровідниковому виробництві), тому, якщо визначати за вкладеннями, закон Мура відстає у фактичних показниках. Але якщо врахувати прогрес програмного забезпечення, 18-місячна тривалість подвоєння здається цілком правдоподібною.

225 Навіть якщо наближення до базового людського рівня було повільним.

226 Були попередні спроби описати ідею вибуху інтелекту в контексті теорії економічного зростання; див., наприклад, Hanson (1998 b); Jones (2009); Salamon (2009). Ці дослідження вказували на можливість надшвидких темпів зростання інтелекту за умови наявності цифрових розумів. Проте оскільки теорія ендогенного зростання надто недорозвинена навіть для застосування до минулих та сучасних процесів, будь-які спроби використати її для аналізу осяжного майбутнього мають розглядатися виключно як джерело потенційно цікавих концептів та міркувань, а не як спосіб упевненого прогнозування. Для огляду спроб математично змоделювати технологічну сингулярність див. Sandberg (2010).

227 Звісно, може трапитися, що жодного стрибка не буде. Але оскільки (як ми визначили раніше) поява суперінтелекту технічно можлива, стрибок може не відбутися лише під впливом яких-небудь зовнішніх чинників, наприклад, екзистенційної катастрофи. Якщо сильний суперінтелект буде не штучним інтелектом чи емуляцією мозку, а одним із інших видів суперінтелекту, які ми тут розглядали, тоді перехід може відбутися за повільним сценарієм.

Розділ 5

228 Програмний розум може виконуватися на одному комп'ютері або в глобальній комп'ютерній мережі. Але тут ідеться не про це. Ми говоримо про обсяг влади, зокрема техногенного походження, яка буде доступна на пізніх етапах або відразу після революції штучного ШІ.

229 Наприклад, у країнах, що розвиваються, швидкість проникнення технологій у споживчому сегменті продукції нижча (Talukdar et al. 2002). Див. також Keller (2004) та The World Bank (2008).

230 Економічна література з теорії фірми може дати цікаву перспективу для порівняння з предметом цієї дискусії. *Locus classicus*: Coase (1937). Див. також, наприклад, Canbäck et al. (2006); Milgrom and Roberts (1990); Hart (2008); Simester and Knez (2002).

231 З іншого боку, викрасти зерно ШІ може бути надзвичайно легко, адже це просто програма, яку можна переслати електронними засобами чи записати на портативний пристрій.

232 Якщо змоделювати ситуацію, у якій часова відстань між проектами має нормальний закон розподілу, тоді найбільш імовірна відстань між лідером і наступним проектом залежатиме від кількості проектів, що конкурують. Якщо конкурентів багато, то відставання від лідера, найімовірніше, незначне, навіть за високої дисперсії розподілу (хоч у разі нормального розподілу часу завершення розробок відстань між лідером і його найближчим конкурентом, зі зростанням кількості учасників, змінюється повільно). Проте малоймовірно, що існуватиме багато проектів, добре забезпечених ресурсами і здатних скласти серйозну конкуренцію лідеру. (Якщо існуватиме кілька перспективних підходів до створення ШІ, кількість проектів може бути більшою, але багато з них можуть виявитися хибними). Зазвичай лише кілька серйозних конкурентів працюють у напрямі конкретної технологічної цілі. У споживацькому сегменті ситуація інша. Там в основному безліч продуктів із незначними відмінностями конкурують один з одним за покупця і бар'єри між нішами низькі. Безліч приватних підприємців роблять дизайнерські принти на футболках, але лише кілька компаній у світі виготовляють якісні відеокарти. (Дві компанії — AMD та NVIDIA — наразі сформували щось на кшталт дуополії, проте Intel також конкурує з ними в бюджетному сегменті).

233 Barber (1991) вважає, що шовк могли використовувати ще в часи культур Яншао (5000–3000 роки до н. е.). Sun et al. (2012). На основі даних генетичних досліджень припускають, що личинок шовкопряда почали вирощувати в домашніх умовах близько 4100 років тому.

234 Cook (1984, 144). Ця історія здається надто гарною, щоб виявитися історично правдивою. Імовірніше, усе було приблизно, як розповідає Прокопій Кесарійський (*Wars* VIII.xvii.1–7): до Візантії личинок пронесли бродячі монахи в порожнинах своїх посохів (Hunt 2011).

235 Wood (2007); Temple (1986).

236 Насправді доколумбійські культури мали колесо, але використовували його переважно як іграшку (напевне, через відсутність добрих тяглових тварин).

237 Koubi (1999); Lerner (1997); Koubi and Lalman (2007); Zeira (2011); Judd et al. (2012).

238 Оцінки з різних джерел. Затримки дещо приблизні, залежно від того, як саме визначаються «еквівалентні» технологічні можливості. У період винайдення радара його використовували щонайменше дві країни, проте складно знайти конкретні місяці.

239 Ellis (1999).

240 У 1953 році відбулися перші випробування RDS-6 — бомби, у якій використовували реакцію термоядерного синтезу, — проте лише у 1955 році випробували справжню термоядерну бомбу RDS-37, у якій більшість вибухової енергії надходила саме від термоядерної реакції.

241 Не підтверджено.

242 Випробування відбулися у 1989-му, а у 1994-му проект закрили.

243 Побудована система з радіусом дії більшим за 5000 км.

244 Ракети Polaris, придбані у США.

245 Сучасна розробка ракети Таймир, найімовірніше, на базі Китайської ракети.

246 Ракета RSA-3, випробувана у 1989–90 роках, призначалася для запусків супутників та/або для МБР.

247 РГЧ ІН = роздільна головна частина з блоками індивідуального наведення, технологія балістичної ракети, яка може нести кілька боєголовок, які окремо програмуються для ураження різних цілей.

248 Систему Agni V досі не введено в експлуатацію.

249 Bostrom (2006 c). Може існувати невидимий синґлтон (наприклад, суперінтелект із технологією чи здібностями, які дають йому змогу контролювати події у світі так, що жодна людина не помічатиме його втручання). Або синґлтон, який за власним бажанням обмежує використання своєї влади (вдаючись лише до дій, необхідних для забезпечення певних міжнародних домовленостей або лібертаріанських принципів). Безсумнівно, ймовірність появи конкретного типу синґлтону — питання емпіричне. Принаймні концептуально синґлтон може бути хорошим, поганим, нескінченно різноманітним, непроникно монолітним, тиранічно жорстоким, схожим на силу природи, а не на горластого деспота.

250 Jones (1985, 344).

251 Важливо пам’ятати, що Мангеттенський проект відбувався під час війни. Багато науковців стверджували, що беруть участь у проекті переважно через війну та страх, що нацистська Німеччина створить ядерну зброю раніше за альянтів. У мирний час важко, не привертаючи уваги, мобілізувати такий потужний інтелектуальний ресурс. Інший знаковий науковий/

технічний мегапроект — программа Аполлон — теж свого часу отримала потужний імпульс від напруженої атмосфери періоду «холодної війни».

252 Хоча навіть якби вони *цікавилися,* навряд чи це було б помітно.

253 Завдяки криптографічним засобам кожен з учасників групи розробників фізично може перебувати в будь-якому місці світу. Єдиним потенційно вразливим елементом ланцюга співпраці може бути етап введення інформації, коли хтось може спостерігати за набором інформації на клавіатурі. Але з появою повсюдного спостереження (за допомогою мініатюрних пристроїв), з'являться і методи протидії йому (наприклад, спеціальні приміщення, захищені від пристроїв стеження). В епоху поширення засобів відеоспостереження, що відкритішим ставатиме фізичний простір, то більше криптографічних засобів захищатиме приватність людей у кіберпросторі.

254 Тоталітарні держави можуть вдатися до більш примусових заходів. Науковців із відповідних сфер науки ув'язнять у трудових таборах, типу тих, що існували в Росії за часів Сталіна.

255 Коли рівень зацікавленості суспільства низький, деякі науковці можуть заохочувати нагнітання, страху в медіа, щоб привернути до себе і своєї роботи увагу та здаватися важливими. Але, заволодівши увагою громадськості, тон повідомлень змінюється до більш спокійного, щоб побоювання не спричинили скорочення фінансування, законодавчих обмежень та суспільного осуду. У суміжних дисциплінах — як-от комп'ютерній науці чи робототехніці — можуть обурюватися таким перетягуванням уваги та фінансування. Свої претензії вони справедливо можуть мотивувати відсутністю в їхній роботі будь-яких ризиків, пов'язаних із вибухом інтелекту. (Тут можна провести історичні паралелі з долею нанотехнологій; див. Drexler 2013).

256 Принаймні ці проекти досягли хоча б деяких своїх цілей. У ширшому сенсі їхню успішність (враховуючи зокрема й ефективність) оцінити важко. Наприклад, у роботі Міжнародної космічної станції траплялися значні перевитрати та затримки. Деталі таких випадків можна дізнатися з публікації NASA (2013). У побудові Великого адронного колайдера теж траплялися затримки, але тут причиною могла бути значна складність самого завдання. Проект «Геном людини» зрештою успішно завершився, але цьому, вочевидь, чимало посприяли приватні корпоративні зусилля Крейга Вентера. Спроби здійснити контрольовану реакцію ядерного синтезу, незважаючи на щедру міжнародну фінансову участь, не виправдали очікувань. Проте тут знову причиною може бути неочікувана складність самого завдання.

257 US Congress, Office of Technology Assessment (1995).

258 Hoffman (2009); Rhodes (2008).

259 Rhodes (1986).

260 Організація ВМС США, що займалася зламом криптографічних кодів, OP-20-G, вочевидь, проігнорувала пропозицію Великобританії ознайомитися з її досягненнями в розшифровці Енігми та не сповістила про цю пропозицію своє керівництво (Burke 2001). Через це в тогочасного керівництва США склалося враження, що Британія приховала від них цю

інформацію. Також Великобританія надала Радянському союзу певну інформацію, яку вона отримала з розшифровок, зокрема про підготовку німцями операції Барбаросса. Але Сталін поставив під сумнів попередження через те, що Британія відмовилася пояснити походження цих даних.

261 Протягом років Расселл, здавалося, виправдовував використання ядерної загрози, щоб примусити Росію погодитися на план Баруха, але згодом став палким прихильником взаємного роззброєння (Russell and Griffin 2001). Свого часу Джон фон Нейман, за його словами, щиро вважав, що війни між Сполученими Штатами та Росією не уникнути: «Якщо ви запитаєте, чому б не розбомбити їх (Росію) завтра, я відповім, чому б не бомбити їх сьогодні? Якщо скажете: то сьогодні, о п'ятій? Моє слово: а чому б не о першій?». (Можливо, він казав це, щоб зайвий раз підкреслити свої антикомуністичні погляди перед обороною США в часи Маккарті. Май він справжній вплив на політику США, невідомо, чи вчинив би він так насправді. Див. Blair (1957, 96)).

262 Baratta (2004).

263 Якщо ШІ буде підконтрольний деякій групі людей, тоді, можливо, що ця проблема також стосуватиметься і цієї групи. Проте коли з'явиться ШІ, можуть існувати нові способи засвідчення зобов'язань за угодою, і навіть люди зможуть подолати нещирість та внутрішні змови.

Розділ 6

264 Чи справді людство домінує на Землі? З екологічного погляду людина найпоширеніша велика тварина (близько 50 кг), але загальна суха біомаса людства (приблизно 100 мільярдів кг) не надто вражає, якщо порівнювати з мурахами (родина Formicidae, 300–3000 мільярдів кг). Людство та допоміжні організми становлять менше ніж 0,001 загальної біомаси Землі. Водночас сільськогосподарські угіддя та пасовища є одним з найпоширеніших типів екосистем та займають близько 35 відсотків суші, що не вкрита льодом (Foley et al. 2007). За типовими оцінками ми споживаємо близько чверті первинної продуктивності (Haberl et al. 2007), хоч за іншими оцінками, залежно від того, як визначати релевантні поняття, ця цифра становить від 3 до 50 відсотків (Haberl et al. 2013). З усіх видів людина має найширше середовище проживання та закінчує найбільшу кількість різних ланцюгів живлення.

265 Zalasiewicz et al. (2008).

266 Див. примітку 264.

267 Строго кажучи, це не зовсім так. Інтелект людського виду варіюється аж до нуля (наприклад, у ембріонів та пацієнтів, що перебувають у вегетативному стані). Тож у якісному розумінні максимальна різниця інтелекту в межах людського виду може бути більшою, ніж між людиною та суперінтелектом. Але якщо в значенні «людина» розуміти «функціонально нормальна доросла людина», то основна думка відповідає дійсності.

268 Gottfredson (2002). Див також Carroll (1993) та Deary (2001).

269 Див. Legg (2008). Загалом Легг пропонує оцінювати агента, що отримав «знання» за допомогою навчання з підкріпленням, за обсягом роботи, яку він може виконувати в усіх

середовищах. Де винагорода може бути просумована, а кожне середовище отримує ваговий коефіцієнт, зумовлений колмогоровською складністю цього середовища. Ми пояснимо, що означає навчання з підкріпленням у розділі 12. Див. також Dowe and Hernández-Orallo (2012) та Hibbard (2011).

270 Суперінтелект досягне значних результатів у створенні та моделюванні нових біологічних та нанотехнологічних структур. Якщо бракуватиме знань для точного моделювання, тоді успіх суперінтелекту залежатиме від доступу до засобів для експериментальних досліджень.

271 Наприклад, Drexler (1992, 2013).

272 Спеціалізований ШІ може, без сумнівів, бути дуже корисним і комерційно успішним, але це не означає, що він матиме суперздібність економічної успішності. Навіть якщо, наприклад, такий ШІ приноситиме своїм власникам кілька мільярдів доларів прибутку на рік, це на чотири порядки менше за решту світової економіки. Для того щоб система могла суттєво впливати на світову економіку, вона має виконувати багато видів роботи, тобто бути компетентною в багатьох сферах одночасно.

273 Проте цей критерій не виключає існування невдалих для ШІ сценаріїв. Наприклад, ШІ може свідомо вдатися до гри з високою ймовірністю програшу. Але в цьому випадку умови можуть складатися так, що (а) ШІ неупереджено оцінив свої незначні шанси і (б) визначив, що кращих варіантів дій у нього немає.

274 Пор. Freitas (2000) та Vassar and Freitas (2006).

275 Yudkowsky (2008 a).

276 Freitas (1980); Freitas and Merkle (2004, розділ 3); Armstrong and Sandberg (2013).

277 Див., наприклад, Huffman and Pless (2003), Knill et al. (2000), Drexler (1986).

278 Цифра базується на космологічній баріонній густині, яка, за даними зонду WMAP, дорівнює $9{,}9 \cdot 10^{-30}$ г/см^3, та на припущенні, що 90 відсотків маси Всесвіту це міжгалактичний газ, 15 відсотків маси галактик припадає на зірки (близько 80 відсотків баріонної матерії), а типова зірка має масу 0,7 Сонця (Read and Trentham 2005; Carrol and Ostile 2007).

279 Armstrong and Sandberg (2013).

280 Навіть, якби ми могли пересуватися зі швидкістю 100 відсотків с (що неможливо для об'єктів з ненульовою інертною масою), кількість досяжних галактик становитиме лише близько 6 · 109. (Пор. Gott et al. (2005) та Heyl (2005)). Ця оцінка основана на припущенні, що ми добре розуміємо всі релевантні фізичні категорії і процеси. Але суперінтелектуальна цивілізація може знайти який-небудь, поки що фізично неможливий для нас спосіб сягнути далі в космос (збудувавши, наприклад, машину часу, створивши нові всесвіти, які розширювалися б швидше, чи будь-який інший найнеймовірніший спосіб). Усвідомлення цього не дає змоги надто покладатися на власні оцінки.

281 Наразі невідомо, скільки придатних для заселення планет припадає на одну зірку, тож ця оцінка приблизна. За деякими прогнозами (Traub 2012), кожна третя зірка класів F, G та K має принаймні одну тверду планету у придатній для життя зоні; див. також Clavin (2012). Серед найближчих до Сонця зірок FGK-зорі становлять приблизно 22,7 відсотка, тобто 7,6

відсотка можуть мати потенційно придатні для життя планети. Крім того, придатні для життя планети можуть існувати і навколо більш численних зірок класу M (Gilster 2012). Див. також Robles et al. (2008).

Не буде потреби піддавати людські тіла випробуванням міжгалактичних перельотів. За процесом колонізації стежитиме ШІ. *Homo sapiens* пересилатиметься як інформація, з якої згодом ШІ зможе генерувати екземпляри нашого виду. Наприклад, з генетичної інформації генеруватиметься ДНК, з якої ШІ з допомогою антропоморфних роботів-наглядачів виведе, виростить та вивчить перше покоління колонізаторів.

282 O'Neill (1974).

283 Дайсон стверджує (Dyson 1960), що основну ідею він запозичив у письменника-фантаста Олафа Стейплдона (Olaf Stapledon 1937), який тим часом міг черпати натхнення з подібних ідей Дж. Д. Бернала (Dyson 1979, 211).

284 Згідно з принципом Ландауера мінімальна кількість енергії, яка потрібна для того, щоб змінити один біт інформації, відома як межа Ландауера, дорівнює kT ln 2, де k — стала Больцмана ($1{,}38 \cdot 10^{-23}$ Дж/К), а T — температура. Якщо припустити, що обчислення відбуваються за температури близько 300 К, то щоб стирати приблизно 10^{47} бітів у секунду потрібно 10^{26} Ватт. (Стосовно максимально досяжної ефективності наномеханічних пристроїв див. Drexler (1992). Див. також Bradbury (1999); Sandberg (1999); Ćirković (2004). Досі немає єдиної думки щодо обґрунтованості принципу Ландауера; див. наприклад, Norton (2011)).

285 Усі зірки мають різну потужність світіння, і Сонце — цілком типовий представник головної послідовності.

286 Більш детальний аналіз може дати змогу точніше визначити тип потрібних обчислень. Максимальна кількість послідовних обчислень, яка може бути виконана за одиницю часу, обмежена, тому розмір швидкого послідовного комп'ютера має бути невеликим, щоб зменшенити затримки під час пересилання інформації. Крім того, кількість збережених бітів і кількість виконаних незворотних обчислень мають свої обмеження (включно зі стиранням інформації).

287 Тобто незначною за деякою «натуральною» метрикою, наприклад, логарифм розміру популяції, яка може впевнено себе забезпечувати на мінімально необхідному для проживання рівні завдяки наявним можливостям, якщо на це спрямовувати всі доступні ресурси.

288 За умови, що не знайдеться жодної позаземної цивілізації, яка могла б цьому зашкодити. Також вважаємо, що гіпотеза про симуляцію в цьому разі хибна. Див. Bostrom (2003 a). Якщо ж будь-яке із цих припущень хибне, то можуть існувати впливові неантропогенні ризики — включно з інтелектом нелюдського походження. Див. також Bostrom (2003 b, 2009 c).

289 Принаймні можливо, що розумний синґлтон, засвоївши основну ідею еволюції, засобами євгеніки поступово збільшуватиме власний рівень колективного інтелекту.

290 Tetlock and Belkin (1996).

291 Для розуміння: наразі людство не має *безпосередньої* можливості колонізувати й освоїти велику частину досяжного Всесвіту. Рівень сучасних технологій не дає змоги розпочати міжгалактичну колонізацію. Я маю на увазі, що наш поточний рівень загалом може дозволити нам створити потрібні для цього технології, тобто колонізація нам доступна лиш *опосередковано*. Крім того, людство поки що не синґлтон, і не можна гарантувати, що ми, розпочавши освоєння Всесвіту, не зустрінемо опору від якої-небудь позаземної інтелектуальної сили. Проте для досягнення рівня стійкості поміркованого синґлтону достатньо мати набір можливостей, з якими поміркований синґлтон, за відсутності інтелектуального опору, мав би непряму можливість колонізувати й освоїти ресурси досяжної частини Всесвіту.

292 Іноді може бути корисно говорити, що два ШІ мають певну суперздібність. Тоді під суперздібністю в ширшому розумінні можна розуміти те, що агент має відносно певного поля діяльності. У нашому прикладі, це поле містить людську цивілізацію, але виключає інший ШІ.

Розділ 7

293 Це, однак, не виключає, що незначні на вигляд відмінності можуть бути спричинені глибинною функціональною різницею.

294 Yudkowsky (2008 a, 310).

295 Девід Ґ'юм, шотландський філософ доби Просвітництва, уважав, що одних лише переконань (наприклад, який вчинок є гідним, а який — ні) недостатньо, щоб спонукати нас до дії: потрібно прагнення. Таке твердження підтримує тезу ортогональності, відкидаючи одне з можливих заперечень, а саме, що достатній рівень інтелекту може спричинити формування переконань, які тим часом зумовлюватимуть мотивацію індивіда. Але хоч теорія мотивації Ґ'юма підтримує тезу ортогональності, вона не є її обов'язковою умовою. Зокрема, це стосується твердження, що самі лише переконання не можуть зумовлювати дії. Достатньо буде припустити, що інтелектуальний агент може бути мотивований до дій достатньо сильним прагненням. Другий можливий спосіб виправдати тезу ортогональності, незалежно від теорії мотивації Ґ'юма, — через твердження, що будь-який рівень інтелекту не може бути самодостатньою причиною набуття (нібито) мотивувальних переконань. І, нарешті, третій спосіб обійти Ґ'юма в наших міркуваннях: припустити можливість побудови такого агента (чи краще «оптимізаційного процесу»), який, маючи будову цілком не схожу на людський інтелект, не міститиме прямих аналогій до людських понять «переконання» та «прагнення» і заразом демонструватиме високий рівень загальної розумності. (Останні спроби захистити теорію мотивації Ґ'юма — див. Smith (1987), Lewis (1988) та Sinhababu (2009)).

296 Наприклад, Дерек Парфіт стверджував, що базові вподобання можуть бути ірраціональними, як у загалом нормального агента, але з «Байдужістю-До-Майбутнього-Вівторка»:

Певний гедоніст вельми піклується про якість свого майбутнього досвіду. Це однаково стосується як найближчого майбутнього, так і віддаленішого, але з одним винятком. Його не турбує, що на нього чекає прийдешнього вівторка. До подій інших днів тижня він зазвичай небайдужий. Що ж стосується страждань чи насолод, які чекають його у вівторок, — він ними не переймається…

Така його байдужість є для нас доконаним фактом. Тому, коли він планує своє майбутнє, він завжди віддасть перевагу жахливим стражданням у вівторок перед найменшим болем у будь-який інший день. [Parfit (1986, 123–124); див. також Parfit (2011)].

Поки ми визна́ємо, що така поведінка може бути виправдана інструментальними потребами, нам не важливо, чи справді такий агент буде ірраціональним за Парфітом. Агент Парфіта може бути бездоганно раціональним в інструментальному сенсі, а отже — високоінтелектуальним, і водночас не зважати на деякі «об'єктивні причини», на які мав би зважати повністю раціональний агент. Тому такий приклад не суперечить тезі ортогональності.

297 Навіть припущення про існування деяких об'єктивних моральних норм, які доступні кожному повністю раціональному агенту і, по суті, мотивують до певних дій (тобто будь-хто, кому вони доступні, достатньо мотивований до певного типу поведінки), не суперечить тезі ортогональності, якщо може існувати агент, якому бракує справжньої раціональності або деяких можливостей, необхідних для сприйняття цих об'єктивних моральних норм, але водночас він є бездоганно раціональним відносно своїх інструментальних мотивацій. (Також агент може не мати повної інструментальної раціональності у всіх сферах і заразом бути надзвичайно інтелектуальним, навіть суперінтелектуальним).

298 Більше про тезу ортогональності — див. Bostrom (2012) та Armstrong (2013).

299 Sandberg and Bostrom (2008).

300 Стівен Омогундро у двох своїх статтях розвідує цю тему (Omohundro 2007, 2008). Він стверджує, що всі розвинені системи ШІ схильні мати низку «базових потягів», під якими він розуміє «схильності, які проявлятимуться, якщо їм відкрито не протидіяти». Термін «потяг ШІ», незважаючи на лаконічність і провокативність, незручний тим, що передбачає подібність впливу інструментальної цілі на процес ухвалення рішення штучним інтелектом, на вплив психологічного потягу на рішення людини. Щось подібне до феноменологічної мотузки, за яку можна сіпати наше его, але натяг якої ми, напруживши волю, усе-таки здатні пересилити. У випадку ШІ така конотація оманлива. Ми ж не скажемо, що людина має потяг до заповнення щорічної податкової декларації про доходи, хоча сплата податків є достатньо конвергентною інструментальною метою в більшості сучасних людських суспільств (метою, реалізація якої допомагає уникнути проблем під час реалізації багатьох інших наших цілей). Існують інші, ґрунтовніші відмінності між нашим трактуванням і трактуванням Омогундро, але основна ідея та сама. (Див. також Chalmers (2010) та Omohundro (2012)).

301 Chislenko (1997).

302 Див. також Shulman (2010b).

303 Агент може також змінити представлення своєї мети: для того щоб зі зміною старої онтології на нову транспонувати старе представлення мети в нову онтологію; пор. de Blanc (2011).

Агент, оснований на доказовій теорії ухвалення рішень, може змінити своє рішення під впливом доведеної важливості такої зміни. Наприклад, агент може вважати, що існують інші такі самі агенти і що його діяльність є свідченням того, як ці інші агенти можуть діяти. Так він може вибрати дружню до цих інших агентів кінцеву мету як свідчення того, що інші агенти можуть теж поставитися до нього дружньо. Якщо не змінювати кінцеву мету насправді, а просто вибирати дії так, ніби кінцева мета змінена, то результат може бути той самий.

304 Адаптивному формуванню вподобань присвячено потужний пласт психологічної літератури. Див., наприклад, Forgas et al. (2010).

305 У формальних моделях цінність інформації визначається як різниця між очікуваними результатами оптимальних рішень з урахуванням цієї інформації та без. (Див., наприклад, Russel and Norvig (2010)). Звідси випливає, що цінність інформації не може бути від'ємною. Крім того, це означає, що будь-яка наявна інформація ніколи не приведе до рішення, що має нульову цінність. Проте така модель передбачає кілька ідеалізацій, які зазвичай не справджуються в реальних умовах — а саме, що знання не мають кінцевої цінності (тобто знання самі собою не мають цінності, а мають лише інструментальну цінність) та що один агент не може до кінця зрозуміти іншого агента.

306 Наприклад, Hájek (2009).

307 Таку стратегію використовують личинки покривника, які плавають, поки не знайдуть придатної скелі, до якої остаточно прикріплюються. Втративши потребу в русі, личинки перетравлюють частину свого мозку (церебральні ганглії). Схожу поведінку можна спостерігати в деяких академічних науковців після затвердження у штаті.

308 Bostrom (2012).

309 Bostrom (2006 c).

310 Можна поглянути на питання з іншого боку і спробувати знайти причини для суперінтелектуального синґлтону *не* прагнути ширших технологічних можливостей. Такими причинами можуть бути: (а) синґлтон не бачить застосування для таких можливостей; (б) витрати перевищують користь (наприклад, якщо технологія не підходить для досягнення кінцевої мети або надто швидко знецінюється); (в) кінцеві цілі синґлтону несумісні з певними способами технологічного розвитку; (г) задля збереження власної стабільності синґлтон може уникати створення технологій, які сприятимуть дестабілізації чи ускладнять її наслідки (наприклад, світовий уряд може запобігати появі технологій, що можуть сприяти повстанню — навіть якщо вони матимуть мирну користь, або технологій виробництва зброї масового ураження — через значну небезпеку, яку вони становитимуть у разі повалення такого уряду); (ґ) синґлтон може бути зв'язаний стратегічним зобов'язанням щодо недопущення появи певної технології, незважаючи на її корисність не певному етапі

розвитку. (Однак зауважте, що *сучасні* пересторороги проти шляхів технологічного розвитку, на кшталт небезпеки гонки озброєнь, можуть на синґлтон *не* діяти).

311 Припустимо, що агент дисконтує майбутні ресурси за експоненційним законом. Також через обмеження швидкості світла він може нарощувати доступні йому ресурси з поліноміальною швидкістю. Чи означатиме це, що за певний час агент втратить сенс у захопленні нових ресурсів? Ні, тому що до нуля прямуватиме не тільки майбутня вартість захоплених зараз ресурсів, а й поточна вартість їх захоплення. Вартість майбутньої (через сотні мільйонів років) відправки додаткового зонда Неймана (імовірно, за рахунок ресурсів, здобутих незадовго перед тим) зменшуватиметься з такою самою швидкістю (з точністю до сталого множника), як і ціна ресурсів, що цей додатковий зонд здобуде.

312 Хоч об'єм колонізованого простору в певний момент часу може бути приблизно сферичним та зростати у квадратичній пропорції до часу (~t2), який минув з моменту запуску першого зонда, кількість ресурсів, що міститимуться в цьому об'ємі, зростатиме не так рівномірно через значну негомогенність розміщення ресурсів у просторі. Спершу, поки відбуватиметься колонізація нашої галактики та найближчих до неї галактик, швидкість зростання ресурсів приблизно відповідатиме ~t2. Далі зростання відбуватиметься стрибками, коли колонізація охоплюватиме наступні галактичні кластери, поки, зрештою, через розширення Всесвіту подальша колонізація стане неможливою.

313 У цьому контексті фактор симуляції може бути особливо важливим. Суперінтелектуальний агент може присвоїти значну ймовірність гіпотезі, відповідно до якої він перебуває в комп'ютерній симуляції, а його дані чуття згенеровані іншим суперінтелектом. Тоді в нашого суперінтелекту можуть виникнути конвергентні інструментальні цілі, які залежатимуть від типу симуляції. Пор. Bostrom (2003 a).

314 Відкриття фундаментальних законів фізики та інших основ світобудови є конвергентною інструментальною метою. Можна вважати це частиною «розумового покращення», хоч певною мірою ця діяльність пов'язана і з «технологічним розвитком» (адже вивчення нових фізичних явищ може надихнути на створення нових технологій).

Розділ 8

315 Крім того є небезпека, що людство не зникне остаточно, але якість його існування буде вкрай незадовільною, а його потенціал буде безповоротно втрачено. Існує також екзистенційний ризик ворожнечі та війни всередині людства навколо першості у створенні суперінтелекту.

316 Важливо не пропустити момент, коли ШІ вперше прийде ідея про потребу приховування (подія, яку ми можемо назвати зародження зради). Цю первинну появу ідеї ШІ приховати не зможе. Хоч одразу після її появи він може спробувати приховати факт її минулого існування і водночас переорганізувати (тим часом не привертаючи уваги) процес довготермінового планування діяльності так, щоб надалі унеможливити стороннє спостереження за ним.

317 Навіть людина може написати невеличку, на перший погляд, цілком безневинну програму, яка матиме зовсім неочікувані наслідки. (Приклади можна знайти серед переможців

Міжнародного конкурсу заплутаного коду на C (International Obfuscated C Code Contest)).

318 Елізер Юдковський також зауважував (Yudkowsky, 2008 a), що деякі засоби контролю ШІ, ефективні в певному сталому контексті, зі зміною контексту можуть втратити ефективність.

319 Термін увів письменник-фантаст Ларрі Нівен (Niven, 1973), проте взятий з реального експерименту стимулювання мозку; пор. Olds and Milner (1954) та Oshima and Katayama (2010). Див. також Ring and Orseau (2011).

320 Bostrom (1997).

321 Імовірно, можна організувати процес навчання з підкріпленням так, що спроба ШІ використати вайргединґ призведе до втрати дієздатності, а не до інфраструктурного пригнічення. Але правда в тому, що в разі вайргединґу завжди щось може піти не так.

322 Такий приклад запропонував Марвін Мінскі (див. Russell and Norvig [2010, 1039]).

323 Тут важливо знати, який тип цифрового мозку буде наділений свідомістю, тобто суб'єктивним феноменологічним досвідом чи, як це називають філософи, «кваліа». Не зрозуміло, як оцінити реакції людиноподібної істоти на різні обставини без детальної симуляції її свідомості в цих обставинах. Також невідомо, чи взагалі існують придатні для суперінтелекту алгоритми, на кшталт навчання з підкріпленням, здатні сформувати кваліа. Навіть якщо вважати, що ймовірність існування свідомості в будь-якому такому обчислювальному процесі дуже низька, кількість процесів може бути настільки великою, що навіть низька ймовірність стане вагомою на шальках моральних терезів. Див. також Metzinger (2003, розд. 8).

324 Bostrom (2002 a, 2003 a); Elga (2004).

Розділ 9

325 Наприклад, Laffont and Martimort (2002).

326 Уявімо, що більшість виборців хоче, щоб їхня країна створила суперінтелект певного типу. Вони обирають кандидата, який обіцяє врахувати їхні побажання, але забезпечити виконання його обіцянок після виборів дуже важко. Проте цього разу він виявився людиною слова: дає доручення уряду створити науковий або промисловий консорціум для виконання роботи. Проте і тут виникають проблеми принципала-агента: урядові бюрократи можуть мати власне бачення того, що має бути зроблено, або діяти за буквою, а не духом наказів свого лідера. Але навіть якщо уряд спрацює якнайкраще, виконавці роботи, науковці, теж можуть мати власний погляд на свій сектор роботи. Проблема виникає знову і знову на кожному рівні. Може, одного вечора, директор якоїсь лабораторії втратить сон, бо виявить, що хтось із його підлеглих вніс несанкціоновані зміни до проекту: в його уяві д-р фіз.-мат. н., панна З. Рада глупої ночі прокрадається до його кабінету, входить до системи контролю версій від його імені, щоб внести зміни до вихідного коду системи мотивації зерна ШІ. Тепер там, де має бути «служити людству», вказано «служити д-ру фіз.-мат. н., панні З. Раді».

327 Біхевіоральне тестування може придатися і не етапі розроблення суперінтелекту — як один з елементів системи заходів безпеки. Якщо ШІ поводитиметься неналежно вже на

стадії розроблення — щось точно не так, але важливо розуміти, що відсутність поведінкових аномалій не означає, що все гаразд.

328 Класичний експлойт, написаний Стівом Домп'є 1975 року для мікрокомп'ютера Altair 8800 («мікрокомп'ютер» — загальна назва тогочасних побутових обчислювальних пристроїв, насправді Altair 8800 мав габарити 457 мм × × 432 мм × 178 мм. — *Прим. пер.*), який використовував ефект електромагнітної індукції (і відсутність екранування корпусу). Програма створювала електромагнітні хвилі, які можна було приймати побутовим транзисторним приймачем, якщо поставити його поблизу комп'ютера (Driscoll, 2012). Молодий Білл Ґейтс, який був присутній на демонстрації цього ефекту, згадує, що був вражений і зачарований побаченим (Gates, 1975). Зараз існують плани випускати чипи із вбудованою можливістю з'єднання через Wi-Fi (Greene, 2012).

329 Пересторога — серйозна справа, якщо помилка може коштувати всього, що людство надбало з часу своєї появи. Можливо, корисним би був такий принцип: той, хто N разів помилково вважав систему безпечною, надалі може вважатися достовірним джерелом такої впевненості лише на $1/(N + 1)$.

330 Під час неформального експерименту роль ШІ виконувала людина, а інша людина, що виконувала роль охоронця, мала завдання не випускати ШІ з ув'язнення. ШІ міг спілкуватися з охоронцем лише за допомогою тексту і мав дві години, щоб переконати охоронця його випустити. У трьох випадках з п'яти ШІ зміг переконати різних охоронців випустити його на волю (Yudkowsky, 2002). Усе, що може зробити людина, ШІ під силу теж. (Проте не навпаки. Навіть якщо в реальних умовах завдання суперінтелекту буде складнішим, ніж у згаданому експерименті — наприклад, охоронець буде краще мотивованим — суперінтелект зможе досягти успіху там, де людина зазнає поразки).

331 Не варто переоцінювати можливості таких запобіжних заходів. Ментальні образи можуть діяти не гірше за графічні. Подумайте про враження, яке справляє на читача книжка; водночас вона не є інтерактивним засобом.

332 Див. також Chalmers (2010). Проте не варто робити висновок, що з повністю ізольованої системи немає *жодної* користі. Цінністю може бути сам процес, який відбувається всередині такої системи. Також її існування та робота можуть бути важливі для інших людей, а отже — впливати на них. Усвідомлення існування ізольованих систем (також і таких, що містять спостерігача) може спричинити в зовнішніх спостерігачів антропічну непевність та загалом вплинути на їхню поведінку.

333 Може бути не до кінця зрозуміло, чому ми розглядаємо соціальну інтеграцію як форму контролю здібностей. Хіба не логічніше було б її віднести до методів відбору мотивації, адже вона впливає на поведінку системи за допомогою стимулів? Невдовзі ми детальніше розглянемо методи відбору мотивації, а зараз, щоб відповісти на запитання, зауважимо, що відносимо до методів відбору мотивації ті методи, які вибирають або змінюють кінцеві цілі системи. Тобто цілі, які важливі для системи самі собою, а не з інструментальних причин. Натомість, соціальна інтеграція не впливає на кінцеві цілі, тому не є методом відбору мотивації. Вона радше прагне обмежити ефективні можливості системи: не дати їй

отримати зиск від шкідливих дій без відповідного покарання (та втрати від погіршення відносин). Тоді, втративши можливість ефективного досягнення кінцевих цілей, система зрозуміє, що в її інтересах поводитися дружньо.

334 Якщо емуляція керуватиметься антропоморфними мотивами, такий підхід може бути більш ефективним.

335 Цю ідею я запозичив у Карла Шульмана.

336 Створити шифр, якого не зможе зламати навіть суперінтелект, — нетривіальне завдання. Наприклад, суперінтелект може якось видобути випадкову послідовність з мозку спостерігача або з мікроструктур генератора випадкових чисел. Якщо ж послідовність отримана за допомогою генератора псевдовипадкових чисел, суперінтелект може відгадати зерно, з якого вона була згенерована. Крім того, суперінтелект може збудувати великий квантовий комп'ютер або навіть відкрити інший, новий фізичний феномен, який потім використати для створення нового виду обчислювальних систем.

337 ШІ може перекомутувати внутрішні з'єднання, щоб вважати, що він отримав токени нагороди, але це може не спрацювати, якщо ШІ будуть потрібні, власне, токени (а не «вважати, що в нього є токени»).

338 Оригінальна стаття — див. Bostrom (2003 a). Див. також Elga (2004).

339 Shulman (2010 a).

340 Базова реальність, ймовірно, містить більше обчислювальних ресурсів, ніж симульована реальність. Адже обчислення, що відбуваються в симульованій реальності мусять також відбуватися у комп'ютері, що творить її. Окрім того, базова реальність може містити інші ресурси, оцінити які симульований агент не здатний. Такий агент існує з волі суб'єкта, що проводить симуляцію і має на ці ресурси інші плани. (Тут наші міркування дедуктивно не повністю правильні: загалом, незважаючи на те, що несимульовані цивілізації володіють усіма ресурсами симульованих цивілізацій, всесвіти, в яких існують симуляції, можуть мати багато ресурсів, тож симульовані цивілізації в середньому можуть мати доступ до більшої кількості ресурсів, ніж несимульовані).

341 У перекладі Г. П. Кочура.

342 Є інші езотеричні міркування стосовно цього питання, які потребують аналізу. Для того щоб усебічно підготуватися до вибуху інтелектуальності, ці міркування можуть бути дуже важливими. Але навряд чи ми зможемо усвідомити цю важливість, поки не дамо собі раду із більш приземленими аспектами проблеми, яким присвячено більшу частину цієї книжки.

343 Пор., наприклад, Quine and Ullian (1978).

344 Які ШІ може дослідити, порівнюючи характеристики різних базових обчислювальних підсистем, ширину та потужність різних інформаційних шин, часових затримок доступу до різних областей пам'яті, частоту зміни випадкових бітів тощо.

345 Апріорні ймовірності світів (обчислені апроксимації) можна визначати як Соломонову імовірність, яка залежить від алгоритмічної складності світу. Див. Li and Vitányi (2008).

346 Одразу *після* зародження зради ШІ може спробувати знищити сліди своїх бунтівних думок. Тому важливо, щоб оцінка відбувалася постійно. Добре було б також

використовувати щось на зразок «чорної скриньки», яка записувала б усю діяльність (включно з точним часом кожного натиску клавіш програмістами), і після автоматичного вимкнення можна було б простежити та проаналізувати причини помилки. Запис краще вести на пристрій з одноразовою можливістю запису.

347 Asimov (1942). До трьох законів пізніше додався «Нульовий закон»: (0) «Робот не може заподіяти своїми діями чи бездіяльністю шкоди людству» (Asimov, 1985).

348 Пор. Gunn (1982).

349 Russell (1986, 161 f).

350 Поки деякі філософи все життя намагалися будувати деонтологічні системи, стали відомі нові обставини та їхні наслідки, які потребують додаткового вивчення. Наприклад, у філософії деонтологічної моралі останнім часом з'явився новий перспективний вид філософських мисленнєвих експериментів, «проблема вагонетки», завдяки якому вдалося дослідити деталі взаємозв'язків між різними уявленнями людей: про моральні відмінності різними діями/бездіяльністю, різницю між навмисними та ненавмисними наслідками вчинків та іншими подібними матеріями; див., наприклад, Kamm (2007).

351 Armstrong (2010).

352 Якщо в складі системи безпеки ШІ мають працювати кілька механізмів, кожен із них варто будувати так, *ніби* це єдиний механізм безпеки. Якщо одне діряве відро поставити в інше діряве, вода все одно витікатиме.

353 Варіантом цієї ідеї може бути створення ШІ, який має діяти на основі прогнозів щодо неявно визначеного стандарту. Тоді кінцевою метою ШІ буде — завжди діяти на основі неявно визначеного стандарту, а постійний пошук сутності цього стандарту буде інструментальною метою.

Розділ 10

354 Це антропоморфні назви, тому не варто їх сприймати занадто буквально — як аналогії абощо. Це приблизні назви для різних концептів типових систем які, ймовірно, мають певні шанси з'явитися в майбутньому.

355 Навряд чи на запитання про результат найближчих виборів користувач хоче почути докладний звіт про проекції розміщення у просторі та вектори імпульсів усіх навколишніх часток матерії.

356 Призначену для конкретної обчислювальної машини з певним доступним набором команд.

357 Kuhn (1962); de Blanc (2011).

358 До джинів та суверенів «метод консенсусу» буде застосувати складніше, бо може існувати багато однаково ефективних варіантів послідовностей елементарних дій (наборів сигналів, які треба надіслати на актуатори) для виконання завдання, і різні агенти можуть вибрати різні, але однаково ефективні, варіанти. Якщо правильно формулювати питання, можна зменшити кількість можливих відповідей на них (наприклад, «так» або «ні»). (Стосовно поняття точки Шеллінга, яку також називають «фокальною точкою», див. Shelling (1980)).

359 Світова економіка певною мірою схожа на редукованого джина, щоправда він вимагає плати за свої послуги. У такому разі значно більшу економіку, яка, можливо, постане у світі майбутнього, можна порівняти із джином з колективним суперінтелектом.

Але існує одна важлива *відмінність* — я можу наказати економіці (за винагороду) доставити мені додому піцу, але не можу наказати доставити мир. Причина не в тому, що економіка має недостатньо влади. Просто вона має недостатньо координації. У цьому вона більше схожа на багатьох різних джинів, що служать різним панам (із суперечливими інтересами). Якщо збільшити загальну потужність економіки, підсилюючи окремих джинів чи додаючи нових, вона однаково може не набути здібностей забезпечення миру. Для того щоб функціонувати, як суперінтелектуальний джин, економіка має не лише покращити свої здібності створення продуктів та надання послуг (включно зі створенням нових технологій), але й навчитися краще вирішувати завдання глобальної координації.

360 Якщо джин втратить здатність виконувати команди — і не зможе самостійно виправити помилку у власній програмі — тоді він може вжити заходів, щоб запобігти появі нових команд.

361 Навіть оракул, який може лише давати відповіді типу «так/ні», може допомогти у створенні джина чи суверена або навіть стати їхнім компонентом. Якщо кількість питань, які можна ставити, необмежена, то оракул може навіть допомогти написати вихідний код для такого ШІ. Питання, наприклад, можуть бути такі: «У двійковому представленні коду першого ШІ-джина, який спадає на думку, *n*-й символ — нуль?».

362 Можна уявити також трохи складнішого оракула або джина, який сприйматиме запитання або команди лише від певної особи, але не виключатиме того, що її наміри можуть бути не доброчесними чи що її шантажує третя сторона.

363 Провідний політичний філософ XX століття Джон Ровлз відомий тим, що використав художню виразність образу завіси невідання, щоб відкрити нам істинні пріоритети, яких ми маємо прагнути під час формулювання суспільного контракту. Його пропозицією було вибирати соціальний контракт так, ніби ми перебуваємо за такою собі завісою і не відаємо наперед, яка роль чекає особисто кожного з нас. За його задумом, у такому разі соціальний контракт формуватиметься виходячи з загальної справедливості та рівного блага для всіх, а не егоїстичних особистих інтересів та самолюбства, яке штовхає нас до подвійних стандартів та невиправданої несправедливості. Див. Rawls (1971).

364 Karnofsky (2012).

365 Винятком було б програмне забезпечення, під'єднане до достатньо потужних виконавчих пристроїв, як, наприклад, програми ранніх систем реагування на ядерну загрозу: замість того щоб безпосередньо запускати ракети, вони сигналізували про загрозу черговому офіцеру, який був уповноважений ухвалювати рішення про подальший запуск. Неправильна робота таких систем може створити небезпечну ситуацію. І таке вже траплялося: двічі. 9 листопада 1979 року через комп'ютерний збій система NORAD (North American Aerospace Defence Command — Командування повітряно-космічної оборони Північної Америки) помилково звітувала про повномасштабний напад Радянського Союзу на Сполучені Штати.

Командування США вже готове було завдати удару у відповідь, коли виявилося, що жодна з радіолокаційних систем раннього попередження не підтверджує ворожі запуски (McLean and Stewart, 1979). А 26 вересня 1983 року внаслідок несправності радянська система «Око» сповістила про ракетний удар з боку Сполучених Штатів. Черговий офіцер командного центру Станіслав Петров визнав тривогу помилковою: рішення, яке відвернуло загрозу термоядерної війни (Lebedev, 2004). Навіть якби війна сталася в розпал «холодної війни» і в ній узяли участь усі ядерні держави, навряд чи вона стала причиною повного зникнення людства, але цивілізації, безперечно, прийшов би кінець, а рівень смертності та руйнувань складно навіть уявити (Gaddis, 1982; Parrington, 1997). Проте в майбутньому людство може накопичити ще більше зброї або винайти нові більш смертоносні її види. Крім того, наше уявлення про вплив ядерного Армагеддону (зокрема наслідки ядерної зими) може бути неправильним.

366 Цей підхід можна класифікувати як метод контролю за допомогою прямого визначення набору правил.

367 Те саме може трапитися, якщо критерієм результату буде певне *мірило* добра, а не чітке визначення потрібного результату.

368 Прихильники створення оракула можуть наполягати, що користувач все-таки здатний розгледіти хибу в запропонованому рішенні, яке не відповідає його наміру, хоч і виконує формальний критерій успішності. Можливість розпізнати помилку на цьому етапі залежить від багатьох факторів: наскільки зрозумілий для людини формат має результат роботи оракула, наскільки адаптований для розуміння людини опис характеристик потенційного результату.

Замість того щоб покладатися на оцінки оракула, можна створити окремий інструмент, який інспектуватиме результат його роботи та в доступній формі ознайомлюватиме нас із перспективами його застосування. Але загалом із таким завданням зможе впоратися тільки інший оракул, висновкам якого нам також доведеться довіряти. Тобто проблема довіри залишається невирішеною — лише змінює точку прикладання. Можна спробувати захиститися, доручивши оцінювання кільком незалежним оракулам, але якщо всі вони зазнають невдачі, результат буде невтішний. Таке може трапитися, наприклад, якщо всі вони матимуть те саме формальне визначення успішності результату.

369 Bird and Layzell (2002) та Thompson (1997); також Yaeger (1994, 13–14).

370 Williams (1966).

371 Leigh (2010).

372 Приклад запозичений з Yudkowsky (2011).

373 Wade (1976). Комп'ютерні експерименти симуляції еволюційного процесу, які спеціально відтворювали особливості біологічної еволюції, привели до дещо дивних результатів (див., наприклад, Yaeger (1994)).

374 Навіть на сучасному рівні розвиненості алгоритмів можна було б створити загальний суперінтелект: для цього, щоправда, знадобився б досить значний — фізично недоступний — обсяг обчислювальних ресурсів (пор., наприклад, система AIXItl; Hutter (2001)). Але

навіть якщо потужність обчислювальних систем людства зростатиме за законом Мура впродовж найближчих ста років, її все одно не вистачить для практичної реалізації такої системи.

Розділ 11

375 Не тому, що це найбільш імовірний чи бажаний тип сценаріїв, а тому, що його найлегше аналізувати засобами стандартної економічної теорії, а тому — зручний початок для дискусії.

376 American Horse Council (2005). Див. також Salem and Rowan (2001).

377 Acemoglu (2003); Mankiw (2009); Zuleta (2008).

378 Fredriksen (2012, 8); Salverda et al. (2009, 133).

379 Важливо вкласти частину капіталу в активи, ціна яких прив'язана до загальних трендів. Диверсифікація інвестиційного портфеля часткою в індексних фондах дає змогу зменшити ризик втрат.

380 Системи соціального забезпечення багатьох європейських країн не забезпечені фондами, тобто повністю залежать від поточних відрахувань верств і податків, що працюють. Такі системи не забезпечать підтримки, бо в разі раптового масового безробіття фінансування відразу зупиниться. Проте уряди можуть деякий час покривати витрати в ручному режимі з інших джерел.

381 American Horse Council (2005).

382 Для того щоб виплачувати 7 мільярдам людей пенсію на рівні 90 тисяч доларів на рік, потрібно щорічно витрачати 630 трильйонів доларів, що в десять разів перевищує поточний світовий ВВП. За останні сто років світовий ВВП зріс у дев'ятнадцять разів з близько 2 трильйонів у 1900 році до 37 трильйонів у 2000 році (у міжнародних доларах 1999 року, Maddison 2007). Тому якби такі темпи зростання зберігалися протягом ще двох сотень років, а популяція не зростала, то забезпечення кожного пенсією у 90 тисяч доларів коштувало б три відсотки світового ВВП. Вибухове зростання інтелектуальності може значно пришвидшити зростання ВВП. Див. також Hanson (1998 a, 1998 b, 2008).

383 А за останні 70 000 років — мабуть, чи не в мільйон разів, якщо справді, як припускають деякі науковці, до того часу зростання популяції було ускладнене низькою густотою населення. Більше можна дізнатися з Kremer (1993) та Huff et al. (2010).

384 Cochran and Harpending (2009). Див. також Clark (2007), а для критики — Allen (2008).

385 Kremer (1993).

386 Basten et al. (2013). Також можливі сценарії неперервного зростання. Узагалі із заглибленням у майбутнє на одне-два покоління точність прогнозів різко знижується.

387 Станом на 2003 рік мінімально необхідний для відтворення населення рівень народжуваності становив 2,33 дитини на одну жінку. Двоє дітей — для відтворення батьків і «третина дитини» — щоб компенсувати (1) більшу ймовірність народження хлопчика і (2) ранню смертність до закінчення плідного віку. У розвинутих країнах цей коефіцієнт менший і становить 2,1 через нижчу смертність. (Див. Espenshade et al. (2003, Introduction,

Table 1, 580)). Кількість населення більшості розвинутих країн зменшувалася б, якби не імміграція. Ось кілька вартих згадки прикладів країн із рівнем народжуваності нижчим за рівень відтворення: Сингапур — 0,79 (найнижчий рівень народжуваності у світі), Японія — 1,39, Китайська Народна Республіка — 1,55, Росія — 1,61, Бразилія — 1,81, Іран — 1,86, В'єтнам — 1,87, Великобританія — 1,90. Навіть у США кількість населення мала б потроху знижуватися, бо рівень народжуваності становить 2,05. (Див. CIA (2013)).

388 Повнота часів може настати через багато мільярдів років. (Див., наприклад, До Ефесян 1:9–10. — *Прим. пер.*).

389 Карл Шульман зокрема звертає увагу, що якщо біологічні люди в період цифрової економіки розраховують жити стільки, як зараз, то треба, щоб політичний режим цифрової економіки не просто захищав їхні інтереси, але й робив це незмінно довго (Shulman 2012). Наприклад, якщо в цифровому світі все відбуватиметься в тисячу разів швидше, ніж у біологічному, то біологічним людям доведеться сподіватися, що протягом 50 000 років суб'єктивного часу в цифровому світі не зміниться нічого, що може стосуватися їхньої долі. Проте якщо цифровий світ мало чим відрізнятиметься від нашого, то протягом цих років там може трапитися багато революцій, війн та інших катастрофічних подій, які можуть спричинити біологічним людям чимало незручностей. Ризик глобальної термоядерної війни чи іншого подібного катаклізму на рівні 0,01 відсоткіа в рік для біологічних людей означатиме майже неминучу загибель. Для подолання цієї проблеми в період цифрової економіки знадобиться більш стабільний світовий порядок: синґлтон, який поступово покращуватиме стабільність.

390 Навіть коли машини працюватимуть значно ефективніше за людей, усе одно наймати людей може бути вигідно: скажімо, за 1 цент на годину. Якби це було єдиним джерелом прибутку, наш вид припинив би своє існування, адже на таку плату неможливо прожити. Але ще одним джерелом доходу може бути капітал. Тепер якщо припустити, що кількість населення зростатиме, поки загальний дохід не вийде на рівень прожиткового мінімуму, логічно очікувати, що в цьому стані людям доведеться тяжко працювати. Наприклад,- уявімо, що прожитковий мінімум становить 1 долар в день. Тоді 90 центів з них даватиме дохід від капіталу, а для отримання ще 10 центів людині доведеться працювати десять годин на день. Але реальний прожитковий мінімум залежить від того, яку і скільки роботи виконує людина: важча робота потребує більше калорій. Якщо кожна година роботи збільшує витрати на харчування на 2 центи, то маємо модель, за якою є сенс збалансовувати роботу й відпочинок.

391 Така форма існування може послабити правові позиції людини та перешкоджати успішно відстоювати свої інтереси. Проте такі добровільні овочі можуть доручити ШІ піклуватися про свої інтереси, зокрема політичні, та вести інші свої справи. (Якщо не відбудеться значних змін у правовому розумінні власності).

392 Можливо, це не найкращий термін («убити», англ. kill — команда інтерфейсу командного рядка UNIX-подібних операційних систем, призначена для припинення роботи програмних процесів. — Прим. пер.). Термін «убивати» звучить невиправдано брутально. «Завершувати»

— надто евфемістично. Справа ускладнюється тим, що процедура припинення складається з двох окремих подій: припинення виконання програми і видалення її з пам'яті. Смерть людини передбачає обидві події, проте у випадку емуляції вони можуть відбуватися цілком незалежно. Наслідки тимчасового припинення виконання програми не серйозніші від наслідків сну в людини, утім остаточне припинення подібне до стану коми. Окрім того, не варто забувати, що емуляцію можна копіювати і вона може працювати на різних швидкостях: цим можливостям не існує прямих аналогій зі світу людей. (Пор. Bostrom (2006 b); Bostrom and Yudkowsky (2015)).

393 На вищій межі комп'ютерних швидкостей може існувати обернена залежність між загальною паралельною обчислювальною потужністю і швидкістю: найвищих показників швидкості можна буде досягнути лише завдяки компромісу в ефективній потужності. Це особливо актуально в еру оборотних обчислень.

394 Емуляцію можна перевіряти приманкою. Щоб упевнитися в надійності емуляції, що перебуває в певному стані, можна час від часу піддавати її впливу тестових послідовностей стимулів. Що більше накопичується змін до попередньо протестованого стану емуляції, то менше впевненості в надійності поточного стану. (Зокрема, досвідчена емуляція може іноді припускати, що вона перебуває в симульованому середовищі, тому варто враховувати можливість впливу цього усвідомлення на її рішення).

395 Деякі емуляції можуть ідентифікувати себе зі своїм кланом, тобто з усіма копіями і варіаціями певного початкового образу, а не з конкретною його реалізацією. Тому якщо після її зупинення інші представники клану функціонуватимуть, вона не вважатиме таку зупинку смертю. Емуляції можуть усвідомлювати, що наприкінці дня вони будуть відновлені до певного збереженого стану і втратять спогади поточного дня, і ставитися до цього цілком байдуже, ніби завсідник вечірок, який звик до епізодичності спогадів з попередньої ночі. Зрештою, це лише ретроградна амнезія, а не смерть.

396 Етична оцінка може також брати до уваги багато інших факторів. Навіть якщо працівники будуть постійно задоволені своїми умовами, результат такої оцінки все одно може бути морально неприпустимим за низкою інших причин. Які саме це причини наразі є предметом дебатів між різними моральними теоріями. Однак будь-яка методика оцінки обов'язково враховуватиме суб'єктивний добробут. Див. Bostrom and Yudkowsky (2015).

397 World Values Survey (2008).

398 Helliwell and Sachs (2012).

399 Пор. Bostrom (2004). Див. також Chislenko (1996), Moravec (1988).

400 Важко сказати, чи інформаційні структури, які утворяться внаслідок такого сценарію, матимуть свідомість (кваліа, досвід сприйняття). Почасти через те, що досі ще ніхто не спостерігав появу нових форм свідомості, а також і через нашу філософську незрілість розуміння, які структури мають свідомість. Можна змінити перспективу і замість питання про свідомість майбутніх сутностей поставити питання, чи матимуть такі сутності моральний статус або чи має нас турбувати їхній «добробут». Але вони можуть виявитися не легшими, ніж питання про свідомість. Власне від відповіді на питання свідомості

залежить і питання морального статусу: тією мірою, якою залежить від здатності цих нових сутностей суб'єктивно сприймати власний стан.

401 Аргументи на користь тези, що тенденція до ускладнення структур характерна не лише для історії розвитку людства, але й геологічного розвитку — див. Wright (2001). Аргументи опонентів (які Райт критикує в розділі 9) — див. Gould (1990). Див. також Pinker (2011) — де обстоюється думка, що ми спостерігаємо довгострокову тенденцію зниження в суспільстві насильства та брутальності.

402 Детальніше про теорію упередження відбору — див. Bostrom (2002 a).

403 Bostrom (2008 a). Для того щоб обійти ефект упередження відбору, може знадобитися уважніше розглянути всі деталі еволюційної історії людства. Див., наприклад, Carter (1983, 1993); Hanson (1998d); Ćirković et al. (2010).

404 Kansa (2003).

405 Наприклад, Zahavi and Zahavi (1997).

406 Див. Miller (2000).

407 Kansa (2003). Також інший дещо провокативний підхід — див. Frank (1999).

408 Немає очевидного способу оцінки рівня глобальної інтеграції. Один із варіантів оцінки розміру політичних сутностей: орган ухвалення рішень у племені мисливців-збиральників інтегрував сотні індивідів, тоді як сучасні найбільші політичні сутності інтегрують понад мільярд осіб. Різниця між цими прикладами — сім порядків, і лише один порядок залишається до того, щоб політична одиниця охопила населення всієї планети. Але в часи, коли найбільшою політичною одиницею було плем'я, населення Землі було значно меншим. Плем'я могло охоплювати до однієї тисячної всієї глобальної популяції. Тоді політична інтеграція людства збільшилася всього на два порядки. У цьому контексті виправдано оцінювати політичну інтеграцію саме за відносними значеннями, а не за абсолютними (особливо якщо врахувати, що вибух інтелектуальності може привести до значного збільшення популяцій завдяки емуляціям та цифровим розумам). Проте варто враховувати також розвиток глобальних інституцій і неформальних спільнот.

409 Одна з причин вважати, що первинна революція ШІ буде швидкою — ймовірне апаратне переважування — тут не діє. Однак можуть існувати інші джерела прискорення, наприклад нові програмні технології та наступний перехід від емуляцій до чисто синтетичних ШІ.

410 Shulman (2010 b).

411 Як врівноважаться переваги й недоліки, залежатиме від роботи, яку виконуватиме суперорганізм, та від загальних здібностей найздібнішого базового образу емуляції. Сучасні організації наймають багато різних людей почасти тому, що всебічно обдаровані люди трапляються дуже рідко.

412 Безумовно, зробити багато копій одного програмного агента неважко. Але варто розуміти, що для того, щоб нові агенти мали ту саму кінцеву мету, лише копіювання може бути недостатньо. Агенти повинні мати не однакові цілі, а радше спільну мету. Тобто якщо агент Боб себелюбний, то й копія Боба буде теж себелюбною. Проте їхня мета не збігатиметься: Боб дбатиме про інтереси Боба, тоді як копія Боба дбатиме про інтереси копії Боба.

413 Shulman (2010 b, 6).

414 Це більше стосується біологічних людей та емуляцій, ніж інших типів інтелекту, функціонування яких завдяки конструктивним особливостям може бути приховане від зовнішнього детектування. З іншого боку, перевірити та верифікувати ШІ з повністю відкритою конструкцією може бути значно легше, ніж нейроморфні інтелекти. Зовнішній тиск соціуму може заохочувати ШІ відкривати власний програмний код і змінювати власну конструкцію в бік більшої відкритості — особливо якщо відкритість буде обов'язковою умовою довіри та можливості брати участь у корисній діяльності. Пор. Hall (2007).

415 Інші проблеми, що на фоні наслідків провалу глобальної співпраці здаються незначними, охоплюють: витрати на пошук взаємно вигідних умов; ймовірність, що деякі агенти віддають перевагу «автономності» настільки, що не бажають брати участь у глобальних формах співпраці, які передбачають механізми моніторингу та забезпечення.

416 Для цього ШІ може внести відповідні зміни до своєї структури і надати спостерігачам можливість переглянути власний вихідний код без права внесення змін. Штучний інтелект, структура якого не настільки зрозуміла (наприклад, емуляція), може досягти того самого, публічно застосувавши до себе певний метод відбору мотивації. Або деяка зовнішня силова структура, наприклад, суперорганізаційна поліція, може не тільки забезпечувати виконання умов договору між двома сторонами, але й внутрішніх зобов'язань окремого агента.

417 Можливо, носії такої безкомпромісної позиції та відчайдухи, які радше битимуться на смерть, але не погодяться на найменший дискомфорт, були популярні в еволюції. Така рішучість могла давати своєму носію важливу сигнальну перевагу. (Проте така рішучість не має обов'язково бути частиною мотивації агента: справедливість та честь повинні бути важливими для нього самі собою).

418 Точний вердикт із цього питання є предметом глибшого аналізу. Існують інші потенційні ускладнення, які ми не маємо змоги тут розглянути.

Розділ 12

419 Можна також запропонувати низку ускладнених та видозмінених варіантів основної ідеї. Один із таких варіантів ми розглянули в розділі 8 — використання задовільної логіки замість максимізаційної, — а в наступному ми коротко оглянемо альтернативні теорії ухвалення рішень. Проте наразі не відволікатимемося і для простоти зосередимося на базовому випадку максимізації корисності.

420 Якщо ШІ матиме нетривіальну функцію корисності. Якщо ж функція корисності агента, скажімо, константа $U(w) = 0$, то створити його буде дуже просто. Будь-яка його дія однаково максимізуватиме очікувану корисність.

421 Крім того, з нашої пам'яті вивітрилися спогади найранішого дитинства, голосистого та рожевощокого маляцтва, коли ми не могли добре бачити, бо через обмежений досвід не вміли розпізнати незнайомі навколишні образи.

422 Див. також Yudkowski (2011) та огляд у розділі 5 роботи Muehlhauser and Helm (2012).

423 Можна сподіватися, що досягнення в розробці програмного забезпечення дадуть змогу подолати ці труднощі. За допомогою сучасних засобів розробки один програміст може створювати програми, про які колись не могла б і мріяти велика група програмістів, що писали програми безпосередньо машинним кодом. Сучасні програмісти ШІ мають доступ до великого різноманіття бібліотек високоякісних засобів машинного навчання та наукових обчислень, тому, використовуючи бібліотечні функції, самотужки можуть побудувати програму для підрахунку унікальних облич на відео з веб-камери, яку в минулому ніколи б не змогла написати одна людина. Поява великої кількості якісних програмних бібліотек, створених фахівцями, для використання будь-ким, дає майбутнім програмістам широкі можливості. Наприклад, програміст-робототехнік майбутнього може мати доступ до бібліотеки стандартних функцій тривимірного друку елементів обличчя, колекції типових офісних об'єктів, бібліотек спеціалізованих траєкторій та безлічі інших функцій, які наразі не доступні.

424 Dawkins (1995, 132). Я не стверджую, що кількість страждання у світі переважає кількість добра.

425 Для цього необхідні значно більші або менші популяції, аніж популяція наших предків. Див. Shulman and Bostrom (2012).

426 З погляду моралі було б краще, якби можна було досягти того самого результату, не заподіявши шкоди багатьом невинним людям. Якщо ж уникнути незаслуженного страждання цифрових особистостей неможливо, то можна компенсувати їм заподіяну шкоду, зберігши їхній останній стан у файл і відновивши тоді, коли настануть більш сприятливі умови й існування людства не буде під загрозою. Таке відновлення в певному розумінні подібне до життя після смерті: до його ролі у спробах теології позбутися доказового аспекту проблеми зла.

427 Одна з найвизначніших фігур галузі — Річард Саттон — визначає навчання з підкріпленням не через метод, а через проблему. Будь-який метод, який дає змогу вирішити проблему, вважається методом навчання з підкріпленням (Sutton and Barto 1998, 4). Але наша розмова стосується групи методів, коли агент діє з кінцевою метою максимізації (у певному сенсі) накопиченої винагороди. Теоретично може існувати деякий агент, який має зовсім іншу мету, але в багатьох ситуаціях успішно імітує прагнення винагороди. А отже, він добре пристосований для вирішення проблем навчання з підкріпленням. Тоді можуть існувати методи, які, за Саттоном, будуть «методами навчання з підкріпленням», але не будуть вразливі для вайргедингу. Проте зауваження, які наведено в тексті, стосуються більшості методів, які сьогодні застосовують у навчанні з підкріпленням.

428 Навіть якби виявилося, що в людиноподібному штучному інтелекті діють людиноподібні розумові механізми, усе одно кінцева мета, яку вибере собі такий інтелект, необов'язково буде подібною до мети звичайної людини. Хіба що все оточення і середовище виховання такої цифрової дитини детально відтворюватиме життя звичайної дитини: завдання не з простих. Проте, навіть якщо вдасться створити таке середовище, навіть незначні відмінності вроджених схильностей цифрової особистості можуть зумовити непередбачені

реакції на цілком буденні події. Однак у майбутньому для людиноподібних ШІ, можливо, вдасться створити надійніший механізм сприйняття цінностей (використовуючи новітні препарати, нейроімпланти або їхні цифрові еквіваленти).

429 Може здаватися дивним, чому це ми, люди, не намагаємося позбутися механізму, який примушує нас змінювати кінцеві цілі. Тут може впливати кілька факторів. По-перше, людська мотивація насправді не надто схожа на холоднокровний алгоритм максимізації корисності. По-друге, у нас немає жодного дієвого способу змінити механізм набуття цінностей. По-третє, причиною періодичної зміни кінцевих цілей може бути інструментальна доцільність (наприклад, для соціальної сигналізації). Якби наші розумові процеси були частково доступні для сприйняття іншим людям або для імітації іншого набору кінцевих цінностей потрібно було занадто багато розумових зусиль, така інструментальна доцільність змін була б меншою. По-четверте, іноді ми справді намагаємося протистояти змінам наших цінностей, наприклад, коли опираємося згубним впливам несприятливого середовища. По-п'яте, існує цікавий варіант, що ми бачимо кінцеву цінність у тому, щоб бути істотою, яка може набувати нових кінцевих цінностей у типовий для людини спосіб.

430 Або мотиваційна система може бути влаштована так, що ШІ буде байдужим до їхньої зміни; див. Armstrong (2010).

431 Тут ми використовуємо роботу Daniel Dewey (2011). Окрім того, наша модель використовує напрацювання Marcus Hutter (2005) та Shane Legg (2008), Eliezer Yudkowsky (2001), Nick Hay (2005), Moshe Looks та Peter de Blanc.

432 Щоб не ускладнювати наш приклад, ми обмежуємося детерміністичними агентами, які не дисконтують майбутні винагороди.

433 Математично поведінку агента можна виразити через *функцію агента*, яка визначає дію для кожної можливої історії взаємодій. Представити функцію агента безпосередньо як таблицю можна лише для найпростіших агентів. Зазвичай агент визначає потрібну дію за допомогою розрахунків. Існує багато способів описати ту саму функцію, тому точніше було б говорити про *програму агента*. Програма агента — це конкретний алгоритм, який вираховує дію, що відповідає певній історії взаємодії. З погляду математики програма агента, яка взаємодіє з формально визначеним середовищем, є зручною, проте ідеалізацією. Справжні агенти — фізичні сутності. Це означає, що агент не лише взаємодіє із середовищем за допомогою фізичних сенсорів та маніпуляторів, але й те, що його «мозок» чи процесор *сам є частиною фізичної реальності*. Тому на його роботу загалом можуть впливати зовнішні фізичні фактори (не лише сенсори). У якийсь момент виникає необхідність говорити про конкретне *втілення агента* — фізичну структуру, яка реалізує функцію агента за відсутності побічних впливів середовища. (Визначення за Dewey 2011).

434 Для навчання агента цінностей Дьюї пропонує таку формулу визначення оптимальності:

$$y_k = \operatorname{argmax}_{y_k} \sum_{x_k y x_{k+1} m} P_1(y x_{\leq m} \mid y x_k y_k) \sum_U U(y x_{\leq m}) P_2(U \mid y x_{\leq m}).$$

Тут P_1 і P_2 — дві функції ймовірності. Друга сума визначена над певним класом функцій корисності по всіх можливих історіях взаємодії. У тексті ми навели дещо іншу версію формули, у якій прямо вказано певні залежності та використане спрощене представлення світів.

435 Варто розуміти, що корисності з множини функцій корисності $\mathbb{U}$, повинні бути порівнювані та усереднювані. У загальному випадку це проблематично. Не завжди можливо представити різні моральні теорії добра в кількісному вигляді функції корисності. Див., наприклад, MacAskill (2010).

436 Або узагальнимо: оскільки $\mathcal{V}$ не завжди передбачає, що $\mathcal{V}(u)$ істинне у світі w для будь-якої пари ймовірного світу та функції корисності (w, u), ШІ потрібне адекватне представлення розподілу умовної ймовірності $P(\mathcal{V}(u)|w)$.

437 Розглянемо множину можливих дій агента $\mathbb{Y}$. Насамперед визначимо, що вважається дією: лише команда базової моторики (наприклад, «надіслати електричний імпульс на канал виводу #00101100») чи більш високорівнева дія («тримати камеру наведеною на обличчя»)? Оскільки ми намагаємося визначити оптимальність, а не створити план практичного втілення, обмежимося сферою базової моторики (оскільки набір команд моторики може згодом змінюватися, нам доведеться індексувати $\mathbb{Y}$ по часу). Проте для планування дій вочевидь доведеться запровадити певну ієрархічну логіку і застосувати її до більш високорівневих дій. Крім того, треба визначитися, як аналізувати внутрішні дії, як-от запис даних у пам'ять. Такі дії теж можуть впливати на наслідки, отже, їх також варто долучити до множини дій $\mathbb{Y}$. Проте існують обмеження: обчислення очікуваної корисності будь-якої дії з $\mathbb{Y}$ потребують багатьох операцій. Якщо кожна з них вважатиметься дією, яка потребує окремої оцінки відповідно до ШІ-ВЦ, то таке фрактальне ускладнення розрахунків ніколи не дасть нам нарешті перейти до дій. Щоб запобігти безкінечному ускладненню, треба обмежити множину можливих дій, які мають бути оцінені. Тоді потрібен алгоритм евристичного пошуку, за яким система визначатиме групу варіантів, з яких далі, виходячи з оцінки, вибиратиме найкращу дію. (Згодом система може ухвалити рішення змінити евристичний алгоритм щодо деяких дій, промаркувавши їх так, що в довгостроковій перспективі ефективність апроксимації наближатиметься до ідеалу, визначеного ШІ-ВЦ).

Далі розгляньмо множину ймовірних світів $\mathbb{W}$. Тут варто визначати множину так, щоб вона була достатньо інклюзивною. Через відсутність деякого ймовірного w у $\mathbb{W}$ агент може неправильно оцінити реальність і ухвалити правильне рішення. Наприклад, уявімо, що ми використовуємо певну онтологічну теорію, щоб визначити склад множини $\mathbb{W}$. Наприклад, ми включили до $\mathbb{W}$ світи, які складаються з часопросторів, наповнених елементарними частками зі стандартної моделі. Якщо дані стандартної моделі неповні або неправильні, епістемологія ШІ буде викривленою. Можна розширювати множину $\mathbb{W}$ на всі можливі фізичні світи, але для таких багатомірних об'єктів, потенціал розширення безкінечний. Наприклад, як щодо дуалістичних світів, у яких властивості свідомості не вичерпуються фізичними властивостями? А як щодо індексизації? Чи інших важливих властивостей світів,

які нам, обмеженим людям, поки що невідомі? Деякі люди вірять у різні онтологічні теорії. (Серед авторів на тему ШІ віра в матеріалістичну онтологію, що ґрунтується на фізиці, — звичайне явище). Однак навіть побіжного погляду на історію розвитку ідей достатньо, щоб усвідомити можливість хибності звичної нам онтології. Так, у XIX столітті фізики напевне не включили б у $\mathbb{W}$ можливість існування неевклідового часопростору, Еверетової (багатосвітової) інтерпретації квантової теорії, космологічного мультивсесвіту чи гіпотези симуляції — теорій, які цілком серйозно розглядаються зараз. Схоже і ми сьогодні можемо не допускати багатьох варіантів, звичних для вчених майбутнього. (З іншого боку, для завеликої $\mathbb{W}$ виникають технічні ускладнення, пов'язані з необхідністю роботи з вимірами трансфінітивних множин). В ідеалі ШІ мав би використовувати деяку відкриту онтологію, яка дала б йому змогу самотужки визначати, коли доречно визнати можливість нової альтернативної метафізики.

Тепер розглянемо $P(w|Ey)$. Визначення умовної ймовірності, строго кажучи, не пов'язане суто з набуттям цінностей. Як інтелектуальна сутність, ШІ повинен від початку мати можливість з певною точністю визначати ймірність тих чи тих фактів чи подій. Система, яка не відповідає таким вимогам, навряд чи зможе створити всі ті загрози, які ми розглядаємо в цій книжці. Але все ж існує частка ризику, що ШІ, маючи достатньо ефективну епістемологію, не зможе правильно судити про засадничо важливі речі. (У цьому сенсі проблема визначення $P(w|Ey)$ подібна до проблеми визначення $\mathbb{W}$). Спроби визначити $P(w|Ey)$ ставлять нас також перед проблемою представлення ймовірності логічно неймовірних речей.

Ці проблеми — визначення множини дій, множини світів та розподілу ймовірностей світів залежно від сенсорного досвіду — досить універсальні: характерні для широкого кола формалізованих агентів. Тепер перейдемо до проблем, характерних виключно для вивчення цінностей, а саме: визначення $\mathbb{U}$, $\mathcal{V}(u)$ та $P(\mathcal{V}(u)|w)$.

$\mathbb{U}$ — множина функцій корисності. $\mathbb{U}$ та $\mathbb{W}$ пов'язані між собою, адже $u(w)$ із $\mathbb{U}$ в ідеалі має визначати корисність кожного ймовірного світу w з $\mathbb{W}$. Водночас $\mathbb{U}$ має бути достатньо численною, щоб ми могли сподіватися, що принаймні одна u зможе адекватно представити потрібну якість.

Далі ми пишемо $P(\mathcal{V}(u)|w)$ замість $P(u|w)$, для того щоб підкреслити, що нас цікавить імовірність саме судження про функцію корисності, а не власне самої функції. Сама собою функція корисності не є судженням, але ми можемо сформувати судження про неї, увівши її в деяке твердження. Наприклад, можемо твердити, що деяка функція корисності $u(\cdot)$ виражає бажання певної особи або представляє постулати певної етики, або це функція корисності, яку принципал хотів би створити, якби добре все обдумав. Отже, про «критерій цінності» $\mathcal{V}(\cdot)$ можна говорити як про функцію, яка як аргумент приймає функцію корисності u, а повертає судження про те, наскільки u задовольняє критерій $\mathcal{V}$. Маючи судження $\mathcal{V}(u)$ ми, ймовірно, зможемо, використовуючи засоби ШІ, знайти умовну ймовірність $P(\mathcal{V}(u)|w)$. (Для ймовірних світів, у яких враховано всі визначальні фактори,

$P(\mathcal{V}(u)|w)$ має давати нуль або одиницю). Питання ж про те, як визначати $\mathcal{V}$, розглянуто далі в тексті.

438 Цим виклики вивчення цінностей не вичерпуються. Наприклад, є питання, як сформувати у ШІ притомні початкові уявлення — хоча би до того моменту, коли він стане достатньо сильним, щоб опиратися спробам програмістів їх виправити.

439 Yudkowsky (2001).

440 Термін узято з американського футболу. Так називають довгу передачу м'яча в залікову зону суперника наприкінці ігрового часу, яку часто роблять у відчайдушному сподіванні, що якимось дивом там опиниться який-небудь гравець дружньої команди і зарахує тачдаун.

441 Підхід «Радуйся, Маріє» оснований на ідеї, що суперінтелект здатний точніше ніж ми, люди, артикулювати бажані умови та цінності. Наприклад, суперінтелект може представити свої цінності як програмний код. Тому якщо наш ШІ здатен представити інші суперінтелекти як програмні процеси, які взаємодіють зі своїм середовищем, він може передбачити, як ці суперінтелекти реагуватимуть на гіпотетичні подразники, як-от вікно із програмним кодом нашого ШІ та проханням надати свої інструкції в якій-небудь зрозумілій нам формі. Тоді наш ШІ прочитає ці уявні інструкції (з власної моделі цих інших суперінтелектів), і ми зможемо створити ШІ, який буде повинен їх виконувати.

442 Наприклад, можна створити детектор, який шукатиме (у моделі світу нашого ШІ) фізичні структури (а точніше — їхні представлення), створені суперінтелектуальною цивілізацією. Тоді можна пропустити етап ідентифікації елементів, відповідальних за мотивацію гіпотетичного суперінтелекту, а призначити нашому ШІ кінцеву мету відтворювати ті самі структури, які, на його думку, продукує суперінтелектуальна цивілізація.

Однак цей шлях також має технічні ускладнення. Наприклад, наш ШІ навіть після досягнення суперінтелектуальності може не знати точно, які фізичні структури здатен створювати інший суперінтелект, тож йому доведеться задовольнятися апроксимацією. Для цього йому може знадобитися спосіб оцінки подібності одного фізичного артефакту до іншого — деяка метрика подібності. Але яка користь з метрики подібності, яка основана на приблизних вимірюваннях фізичних характеристик? З метрики, за якою незрозуміло, на що більше схожий мозок: чи на комп'ютер, що виконує програму емуляції, чи на камамбер.

Напевно, простіше буде шукати «маячки»: повідомлення про функції корисності, закодовані у зручному форматі. Якщо припустити, що деякі дружні до нас позаземні суперінтелектуальні цивілізації створять у Всесвіті багато таких «маячків», то ми маємо створити ШІ, який прагнутиме знайти і виконати ці гіпотетичні повідомлення.

443 Якби *всі* цивілізації вибрали спосіб «Радуйся, Маріє», то така передача не принесла б результату. Комусь доведеться йти до вирішення складним шляхом.

444 Christiano (2012).

445 ШІ, який ми будуємо, також не повинен створювати деталізовану модель. Як і ми, у своїх міркуваннях він може спиратися на наслідки неявного визначення (наприклад, на характеристики середовища).

446 Пор. розділи 9 та 11.

447 Наприклад, MDMA («екстазі». — *Прим. пер.*) може тимчасово збільшувати здатність до емпатії; окситоцин здатен тимчасово збільшувати довірливість (Vollenweider et al. 1998; Bartz et al. 2011). Проте ефект, схоже, досить індивідуальний і значною мірою залежить від обставин.

448 Удосконалених агентів можна остаточно зупинити або тимчасово зупинити, відновити до попереднього стану, знизити їхній статус та заборонити їм дальше вдосконалення, поки усі інші підсистеми не розвинуться настільки, щоб компенсувати небезпеки попередніх вдосконалень.

449 Те саме може стосуватися і майбутніх соціумів біологічних людей, які матимуть доступ до просунутих технологій нагляду, біомедичних засобів психологічних маніпуляцій або просто зможуть дозволити собі утримувати велику кількість професійних охоронців для захисту від звичайних громадян (та одне від одного).

450 Пор. Armstrong (2007) та Shulman (2010 b).

451 Невідомо, чи треба буде наглядачу рівня *n* спостерігати не тільки за рівнем (*n* — 1), але й за рівнем (*n* — 2), щоб упевнитися, що наглядачі (*n* — 1) сумлінно виконують свої обов'язки. І, зрештою, чи треба йому моніторити рівень (*n* — 3), щоб упевнитися в ефективності рівня (*n* — 1) щодо (*n* — 2)?

452 Цей підхід одночасно має ознаки і відбору мотивації, і контролю здібностей. Людський контроль за програмними наглядачами технічно можна вважати контролем здібностей, а ієрархічну структуру агентів, у якій вищі рівні контролюють нижчі, можна вважати способом відбору мотивації (адже така організація зумовлює мотиваційні тенденції всередині системи).

453 Власне, існує багато інших, вартих уваги витрат, яких ми, на жаль, не можемо навести. Наприклад, керівні повноваження можуть зіпсувати деяких агентів або призвести до корупції.

454 Щоб такі кроки не зашкодили ефективності організації, вони мають запроваджуватися в атмосфері довіри. Інакше, маніпулюючи логікою та емоціями, можна навіяти емуляціям страх бути зупиненими або примусити їх утриматися від використання своїх прав.

455 Див., наприклад, Brinton (1965); Goldstone (1980, 2001). (Прогрес у соціальних науках був би корисним авторитарним лідерам, які завдяки більш точному моделюванню соціальних заворушень могли б оптимізувати стратегії контролю населення та без зусиль відсікати найменші пагони заколоту).

456 Пор. Bostrom (2011 a, 2009 b).

457 Повністю штучні системи можуть користуватися деякими перевагами інституціональної структури без фактичного створення для цього окремих агентів. Така система може реалізовувати різні перспективи завдяки організації процесу ухвалення рішень, не створюючи окремих розумових структур. Проте реалізувати описану в тексті функцію «оцінити біхевіоральні наслідки запропонованої зміни і, якщо наслідки небажані, повернутися до попереднього стану» може бути проблематично без субагентів.

Розділ 13

458 Нещодавнє опитування професійних філософів показало, що респонденти переважно «приймають або схиляються до» певних позицій. Щодо нормативної етики результати були такі: деонтологія — 25,9 %; консеквенціалізм — 23,6 %; етика чеснот — 18,2 %. Що стосується метаетики, то найбільша кількість опитаних схилялася до: морального реалізму — 56,4 %; морального антиреалізму — 27,7 %. Щодо моральних суджень: когнітивізм вибрали 65,7 %; нонкогнітивізм — 17,0 % (Bourget and Chalmers 2009).

459 Pinker (2011).

460 Щодо обговорення цієї проблеми див. Shulman et al. (2009).

461 Moore (2011).

462 Bostrom (2006 b).

463 Bostrom (2011 a)

464 Окрім тих випадків, коли ми маємо всі підстави вважати свою думку більш правильною. Наприклад, ми краще, ніж суперінтелект, знаємо, про що думаємо в конкретний момент, хіба що суперінтелект здатен просканувати наш мозок. Однак якщо суперінтелект може якимось іншим способом отримати доступ до наших думок, тоді краще і тут покладатися на його судження. (В окремих випадках можуть існувати індексичні дані, які треба враховувати, — тоді, наприклад, суперінтелект має пояснити нам, яка думка з нашої перспективи була б розумнішою). Ввід у новий напрям філософської літератури — філософію свідчень та епістемічного авторитету — див., наприклад, Elga (2007).

465 Yudkowsky (2004). Див. також Mijic (2010).

466 Наприклад, Девід Льюїс запропонував диспозиційну теорію цінності, яка, якщо спрощено, стверджує, що *X* є цінністю для *A* тоді і тільки тоді, коли *A* є ідеально раціональним і досконало знає *X* (Smith et al. 1989). Схожі ідеї уже висловлювали раніше; див., наприклад, Sen and Williams (1982), Railton (1986) та Sidgwick and Jones (2010). Популярний спосіб філософського обґрунтування, а саме: *метод рефлексивної рівноваги* з подібною логікою пропонує процес ітеративного взаємного коригування висновку поміж нашою інтуїтивною інтерпретацією окремих випадків, загальними правилами, які, на нашу думку, поширюються на ці випадки, та принципами, виходячи з яких ми згодні коригувати ці елементи, щоб отримати більш когерентну систему; див., наприклад, Rawls (1971) та Goodman (1954).

467 Імовірно, за цією логікою, ШІ мусить запобігати катастрофам, діючи максимально м'яко: так, щоб відвернувши найгірші наслідки, уникати надлишкового впливу на долю людства.

468 Yudkowsky (2004).

469 Ребекка Роуч, з власних слів.

470 Три принципи такі: «Захищай людей, майбутнє людства та природу людяності» (під *людяністю* розуміється найвищий ідеал, те що бажає втілити в собі *людство*); «Рід людський не повинен бути приреченим провести залишок вічності, жалкуючи, що програмісти не зробили чого-небудь по-іншому» та «Допомагай людям».

471 Деякі релігійні вчення більше акцентують на вірі, нехтуючи розумом, який, на їхню думку, принципово неспроможний сповна осягнути духовні істини — навіть якщо пильно, старанно й неупереджено вивчить усі писання, одкровення та екзегези. З такої позиції КЕБ може не здаватися оптимальним джерелом рішень (проте все ж, ліпше, ніж інші підходи, які можуть бути застосовані, якщо відкинути КЕБ).

472 ШІ, який діє непомітно — як сила природи, регулює взаємодії людей — у джерелах називається «Сисоп»: ніби «операційна система» людської цивілізації. Див. Yudkowsky (2001).

473 «Може», бо це залежить від того, чи вважатимуться, з погляду когерентного екстрапольованого бажання людства, ці сутності морально значущими (хоч зараз здається, що таки мали б вважатися). «У перспективі», тому що навіть якщо розподіл інтересів виявиться не на користь цих маргінальних сутностей, залишається імовірність, що інші морально легітимні сутності-симпатики в межах чинних правил зможуть успішно представляти інтереси маргіналів (можливо, в обмін на деякі власні ресурси). Можливість таких коригувань залежить від еластичності засобів регулювання, установлених механізмами КЕБ, і від того, як вдасться подолати проблеми стратегічних переговорів.

474 Індивіди, які здійснили свій вклад у створення безпечного та прихильного до людей суперінтелекту можуть все-таки розраховувати, хай і не на ексклюзивне право визначати долю багатств людства, але принаймні на певну винагороду. Проте ідея рівної участі всіх людей у базі екстраполяції — чудова точка Шеллінга, не варто так просто її відкидати. У будь-якому разі чесноту можна винагородити — непрямим способом, а саме: КЕБ може передбачати деяку нагороду для всіх добрих людей, що працюють на благо людства. І це необов'язково має бути більша вага в базі екстраполяції КЕБ.

475 Bostrom et al. (2016).

476 Суперінтелект може пізнати сутність поняття «моральної правоти» тією мірою, якою існує яке-небудь (достатньо точно визначене) загальноприйняте значення, що ми використовуємо в моральних твердженнях. Тоді суперінтелект зможе судити і про правдивість тверджень типу «Агент *X* повинен *Ф*», якщо взагалі можна судити про правдивість моральних тверджень (наявність у них деякої характерної сутнісної властивості, що визначає їхню інгерентну правдивість чи хибність). Принаймні він зможе це робити значно швидше.

477 Оскільки метаетика під сумнівом, невідомо, що робити ШІ, якщо не вдасться забезпечити попередні умови застосування МП. Наприклад, якщо хибна теорія виявиться правильною (а, відповідно, позитивні моральні твердження виду «Я повинен *Д*» є хибними), то запускається запасна процедура (наприклад, завершення програми). Можна також заздалегідь визначити, як чинити, якщо існує кілька варіантів морально правильних вчинків. Наприклад, запропонувати ШІ виконати (одну з багатьох) дозволених дій, які схвалило б наше когерентне екстрапольоване бажання. Можемо також визначити, як діяти, якщо віднайдена ШІ правдива мораль взагалі не послуговуватиметься терміном, як-от «моральна правота». Наприклад, консеквенційна теорія може стверджувати, що деякі дії кращі за інші, але водночас немає чіткої кореляції між кращою дією та «морально

правильною». Тому, якщо така теорія правильна, механізм МП повинен вибрати одну з морально найкращих можливих дій. Якщо ж дій безліч, і для кожної може існувати краща, механізм МП міг би вибрати будь-яку можливу дію, набагато кращу, ніж та, яку в ідентичній ситуації вибрали б люди. Якщо ж можливість відсутня, тоді можна діяти так, як діяли б люди.

Якщо спробувати вдосконалити запропонований механізм МП, на думку спадає кілька загальних зауважень. Перший варіант має бути якомога консервативнішим, будь-які ускладнення мають завершуватися запобіжним механізмом. У вживанні поняття «моральної правоти» варто обмежитися лише повністю зрозумілими випадками. Крім того, варто додати, що визначення запропонованого механізму має «інтерпретуватися поблажливо і критично та переглядатися з появою нових даних та накопичення досвіду».

478 З усього цього лише «знання» можуть здатися найбільш придатними для формального аналізу (в інформативно-теоретичному розумінні). Проте для представлення людського розуміння знання ШІ може знадобитися заплутаний набір представлень складних психологічних понять. Адже людина не «знає» всього, що зберігається в неї в мозку.

479 Категорії, якими оперує КЕБ, насправді малозрозумілі, бо якби ми могли міркувати про моральну правоту в поняттях КЕБ, це вважалося б прогресом філософії. Саме цього, власне, і прагне, один з основних елементів метаетики — теорія ідеального спостерігача. Див., наприклад, Smith et al. (1989).

480 Для цього треба вирішити проблему фундаментальної нормативної невизначеності. Виявляється, не завжди правильно діяти відповідно до моральної теорії, яка має найбільшу ймовірність бути правдивою. Крім того, не завжди варто вибирати дію, яка має найвищу ймовірність бути правильною. Для вирішення проблеми, окрім ймовірностей, треба зважати на «ступінь неправильності» та серйозність наслідків. Деякі міркування у цьому напрямі — див. Bostrom (2009 a).

481 Можливо, навіть це умова адекватності ШІ: вміння застосувати поняття моральної правоти. Адже будь-яка людина і без ящика пива може відрізнити добро від зла.

482 Навіть якщо ми віримо в те, що *ШІ* завжди діятиме відповідно до принципу МП, це не означає, що створити такий ШІ-МП буде морально правильно *для нас*. Можливо, створення такого ШІ насправді було б зумовлене губристичними мотивами або самовпевненістю (особливо з огляду на можливий осуд). Зарадити сумнівам можна, частково змінивши механізм МП. Нехай ШІ буде зобов'язаний діяти (згідно з принципом МП) лише якщо підтвердить, що його створення було морально правильним вчинком з нашого боку; інакше він повинен буде припинити роботу. Які моральні недоліки можна закинути *такому* ШІ? Адже якби його створення було помилкою, він би просто вимкнувся відразу після початку роботи, не зробивши нічого поганого. (Однак ми таки могли вчинити неправильно. Наприклад, не скориставшися можливістю створити який-небудь інший тип ШІ).

Ще одна проблема — це надмірна старанність. Уявімо, що ШІ має вибір із багатьох можливих дій, однаково правильних з погляду моралі — точніше, усі вони морально прийнятні. Проте деякі з них кращі з погляду моралі, ніж інші. Один з варіантів для ШІ — прагнути вибрати

морально найкращу дію в кожній ситуації (або одну з однаково хороших). Іншим варіантом було б вибрати з усіх морально прийнятних дій ту, яка найкраще відповідає іншому (не пов'язаному з мораллю) критерію. Наприклад, з усіх морально прийнятних дій ШІ може вибрати ту, яка найкраще відповідає нашому КЕБ. Такий ШІ може краще забезпечувати наші інтереси, водночас не порушуючи вимоги моралі.

483 Оцінюючи моральну допустимість створення нами ШІ, він має розуміти «допустимість» в об'єктивному розумінні цього поняття. У звичайному розумінні лікар, який призначає пацієнту ліки і не знає, що пацієнт помре, бо в нього на цей препарат алергія, теж діє морально допустимо. Епістемічна вищість ШІ дасть йому змогу досягти об'єктивного розуміння моральної допустимості.

484 Власне, це залежить від того, у правдивості якої етичної теорії ШІ буде *переконаний* (чи точніше від імовірнісного розподілу правдивості відомих йому теорій).

485 Складно уявити, якими надзвичайно прекрасними можуть бути ці фізично можливі життя. Поетична спроба передати це — див. Bostrom (2008 c). Аргументація того, що деякі з цих можливостей можуть бути сприятливими для *нас*, сучасних людей — див. Bostrom (2008 b).

486 Просування не тієї пропозиції, яку вважаєш кращою, може здаватися нещирим та маніпулятивним вчинком. Проте його можна виправдовувати недосяжністю ідеалу, стверджуючи, що вибрана пропозиція — найкраща з практично можливих.

487 Або в інших позитивно-оцінювальних категоріях, як «доброта», «класність» чи «чудовість».

488 Є такий принцип у розробленні програмного забезпечення — «Do What I Mean» (роби те, що я маю на увазі. — *Прим. пер.*), DWIM. Див. Teitelman (1966).

489 Обов'язково мають бути з'ясований вибір трьох компонент: мети, теорії ухвалення рішень та епістемології. Проте не наполягатимемо на саме такому поділі.

490 Проект створення суперінтелекту, який базується на етиці, спрямує лише скромну частину майбутніх надбань на винагороду тим, хто в морально допустимий спосіб сприяв його успішному завершенню. Неприпустимо було б використовувати для «привабливої обгортки» надто багато ресурсів. Це було б схоже на благодійний фонд, що використовує 90 відсотків зібраних коштів на бонуси для фандрейзерів та на рекламу.

491 Як винагородити мертвих? Є кілька способів. Насамперед для тих, хто бажав би посмертного визнання, можна побудувати монументи та проводити меморіальні заходи. Також померлі можуть мати побажання, що стосуються культур, мистецтва, будівель, природного середовища, які теж можна вшанувати. Далі більшість людей опікується добробутом своїх нащадків, тому їхнім дітям та онукам можна надати які-небудь переваги. Зрештою, суперінтелект може створити досить точні симуляції цих людей — симуляції, які матимуть свідомість та відтворюватимуть оригінал досить точно, щоб це вважалося для них можливістю подальшого існування (принаймні за деякими критеріями). Можливо, це буде простіше зробити для людей, щодо яких здійснили кріомедичну зупинку. Однак для суперінтелекту може бути можливо створити симуляцію особи на основі інших записів:

кореспонденції, публікацій, аудіовізуальних матеріалів чи особистих спогадів очевидців. Також суперінтелект може мати інші, недоступні нам, джерела інформації.

492 Про пограбування Паскаля див. Bostrom (2009 b). Щодо аналізу проблематики нескінченної корисності — див. Bostrom (2011 a). Стосовно нормативної невизначеності — див. Bostrom (2009 a).

493 Наприклад, Price (1991); Joyce (1999); Drescher (2006); Yudkowsky (2010); Dai (2009).

494 Наприклад, Bostrom (2009 a).

495 Використання непрямої нормативності для визначення мети ШІ може усунути проблеми, пов'язані з вибором неправильної теорії ухвалення рішень. Розглянемо, наприклад, запропонований у тексті механізм КЕБ. Вдала реалізація може компенсувати принаймні деякі з помилок визначення цієї теорії. Особливості реалізації механізму КЕБ можуть поставити в залежність отримані цінності від теорії ухвалення рішень. Якби наші ідеалізовані моделі могли знати, що визначають кінцеву мету для ШІ, який ґрунтується на певній теорії, вони, можливо, видозмінили б цінності, щоб забезпечити більш сприятливий результат роботи ШІ — так можна компенсувати викривлення, які вносить одна лінза, додавши перед нею іншу з протилежним ефектом.

496 Деякі епістемологічні системи працюють холістично, не мають конкретної ідейної основи. У такому разі структурним спадком буде не деякий набір принципів, а, власне, початок пізнання, схильність до певних реакцій на певну послідовність подразників.

497 Причини виникнення таких помилок обговорюються, наприклад, у Bostrom (2011 a).

498 Одним із відкритих питань антропічного міркування є т. зв. припущення про самоіндикацію. Відповідно до нього факт мого існування збільшує ймовірність існування інших *N* спостерігачів пропорційно до кількості *N*. Аргументацію проти цього принципу можна знайти в розумовому експерименті «Самовпевнений Філософ» у Bostrom (2002 a). Захист принципу наведено в Olum (2002); критику захисту — Bostrom and Ćirković (2003). Застосування принципу може вплинути на низку потенційно стратегічно важливих емпіричних гіпотез, наприклад, «Теорему Судного дня» Картера — Леслі, проблему симуляції, гіпотезу «великого фільтра». Див. Bostrom (2002 a, 2003 a, 2008 a); Carter (1983); Ćirković et al. (2010); Hanson (1998 d); Leslie (1996); Tegmark and Bostrom (2005). Те саме можна сказати стосовно інших ускладнень теорії упередження відбору, наприклад, чи може вибір опорного класу залежати від моменту-спостерігача і як?

499 Див., наприклад, Howson and Urbach (1993). Існують також цікаві результати дослідження, які звужують множину ситуацій, у яких два баєсових агенти можуть зробити різні раціональні висновки на основі однакових знань; див. Aumann (1976) та Hanson (2006).

500 Пор.: концепт «останнього судді» в Yudkowsky (2004).

501 Існує багато відкритих епістемологічних проблем; деякі з них згадано в тексті. Суть у тому, що для досягнення майже ідеального результату нам, можливо, не потрібно намагатися знайти всі відповіді. Може спрацювати змішана модель (яка складається з великої кількості різних передумов).

Розділ 14

502 Цей принцип запропоновано в Bostrom (2009 b, 190), і він не тавтологічний. Візуальна аналогія: уявіть собі велику коробку зі скінченним об'ємом — це простір базових можливостей, які може забезпечити деяка технологія. Уявіть, що в коробку насипають пісок — це зусилля, які витрачають розробники цієї технології для її створення. Від того, як насипати пісок, залежить, де спершу з'явиться купка, але так чи інакше, зрештою пісок заповнить усю коробку.

503 Bostrom (2002 b).

504 Питання регулювання наукової діяльності зазвичай розглядається з іншої перспективи. На думку Гарві Аверча, науково-технічна політика США в період з 1945 по 1984 рік переважно зосереджувалася навколо питання оптимального рівня інвестицій у галузь та до якої міри уряд має втручатися, «обираючи найкращих» для максимального економічного зростання та забезпечення військової міці. У таких розрахунках технологічний прогрес завжди поставав силою добра. Проте Аверч також зауважує, що в критичних перспективах, позиція «прогрес — завжди на краще» втрачає свою аксіоматичність (Averch 1985). Див. також Graham (1997).

505 Bostrom (2002 b).

506 І в цьому теж немає тавтології. Можна уявити, що було б, якби розвиток відбувався в іншому порядку. Логічно припустити, що краще спочатку вирішити простіше завдання, скажімо, створити нанотехнологію. Це прискорить появу необхідних інститутів, міжнародну співпрацю та зріліше розуміння глобальної стратегії. Можливо, замість штучного інтелекту, нам буде легше протистояти більш метафізично зрозумілій загрозі. Нанотехнологія (синтетична біологія чи будь-яке інше завдання, що трапиться на нашому шляху) може стати тою першою сходинкою, на шляху сходження до вищого рівня можливостей, необхідного для створення суперінтелекту.

Тут можна оцінювати лише конкретні ситуації. Наприклад, з нанотехнологіями треба оцінювати всі можливі наслідки, як-от зростання швидкодії нанотехнологічних апаратних засобів; вплив появи дешевого фізичного капіталу на розвиток економіки; бурхливий розвиток технологій спостереження; можливість появи синґлтону як прямий або непрямий наслідок прориву в нанотехнологіях; зрештою, зростання ймовірності створення нейроморфного штучного інтелекту й емуляції цілого мозку. Розгляд цих проблем (чи схожих наслідків будь-якої іншої ризикової технології) виходить за межі цього дослідження. Тут ми навели лише першу ліпшу ілюстрацію випадку ранньої появи суперінтелекту. Водночас існують ускладнення, які можуть вплинути на точність нашої попередньої оцінки.

507 Pinker (2011); Wright (2001).

508 Може здатися, що гіпотеза пришвидшення безглузда через те, що (на перший погляд) немає жодних спостережуваних наслідків, але див., наприклад, Shoemaker (1969).

509 Рівень готовності не визначається тим, скільки часу витрачено на підготовку, а лише якістю умов та готовністю ключових суб'єктів ухвалення рішень до дій.

510 Також важливим фактором впливу на процес підготовки до вибуху інтелектуальності може бути ступінь міжнародної довіри. Ми розглянемо це трохи далі, у пункті «Співпраця».

511 Кумедно, але здається, ніби ті, кого сьогодні серйозно цікавить вирішення проблеми контролю, досить непропорційно представляють лише один з екстремумів загального розподілу інтелектуальності. А втім, таке враження може мати інше пояснення. Якби ця сфера досліджень раптом стала модною, її відразу б наповнили посередні науковці та просто навіжені.

512 Цей термін я запозичив у Карла Шульмана.

513 Наскільки точно штучний інтелект має відтворювати мозок, щоб вважатися емуляцією мозку, а не просто нейроморфним ШІ? Критерієм може бути, чи система відтворює лише смисли, чи всі аспекти пізнавальних та оціночних функцій конкретного індивіда чи базової особи, бо це може бути важливо для вирішення проблеми контролю. Для того щоб зафіксувати ці характеристики, може знадобитися досить високий рівень точності відтворення.

514 Звісно, прискорення залежатиме від сили імпульсу та його природи. Тобто якщо прогрес в емулюванні цілого мозку відбуватиметься завдяки ресурсам, які, за інших умов, були б спрямовані на нейронаукові дослідження, немає підстав очікувати серйозного прогресу в цій галузі. Однак, можливо, варто вважати емуляцію цілого мозку одним із напрямів нейронауки, а інвестиції ресурсів в емулювання лише ефективнішим способом її розвитку.

515 Див. Drexler (1986, 242). Дрекслер (особисто) підтвердив, що наведений приклад цілком відповідає його ідеї. Звісно, для того щоб перетворити цей набір тез на дедуктивно правильний ланцюжок міркувань, доведеться дещо доповнити його деякими неявними засновками. (Зверніть увагу, що теза про «подібну аргументацію», наведена далі по тексту, не належить Дрекслеру і ним не схвалена).

516 Можливо, нам не варто толерувати малі катастрофи, якщо від того ми станемо аж надто пильними і запобігатимемо помірно-серйозним нещастям, які потрібні для того, щоб ми могли вжити потрібних запобіжних заходів проти екзистенційних катастроф? (Крім того, не варто забувати про можливість перебільшених реакцій — по аналогії з імунною системою, як алергії та аутоімунні захворювання).

517 Пор. Lenman (2000); Burch-Brown (2014).

518 Пор. Bostrom (2007).

519 Варто зауважити, що в наших міркуваннях щодо цього не врахований час подій, а лише їхній порядок настання. Суперінтелект може зменшити екзистенційні ризики, лише якщо його поява змінить порядок основних подій: наприклад, якщо суперінтелект з'явиться раніше, ніж відбудуться важливі відкриття в нанотехнологіях або синтетичній біології.

520 Якщо вирішити проблему контролю *набагато* важче, ніж досягти потрібної швидкості роботи штучного інтелекту, і якщо інтелектуальний потенціал наукової групи проекту не надто залежить від її розміру, тоді, можливо, було б краще, якби саме невеликий проект першим створив суперінтелект, особливо якщо варіативність здібностей невеликих проектів вища. За таких умов — більша ймовірність, що дослідникам стане ресурсів подолати

завдання контролю, навіть якщо загалом малі проекти в середньому демонструють менший інтелектуальний потенціал, ніж великі.

521 При цьому, без сумнівів, можуть існувати засоби, здатні пришвидшити глобальну наукову взаємодію, і потребують новітніх потужних апаратних засобів — наприклад, високоякісний переклад, кращий пошук, широка доступність смартфонів, привабливі можливості віртуальної реальності для соціальної взаємодії тощо.

522 Інвестиції в технології емуляції можуть не лише безпосередньо сприяти появі емуляції цілого мозку (через створення конкретних технічних вирішень), але й опосередковано — формуючи суспільні настрої, які сприятимуть подальшому зростанню фінансування, суспільного розуміння та сприйняття емуляції цілого мозку (ЕЦМ).

523 Чи багато втратимо, якщо майбутнє формуватиметься на основі бажань одної випадкової людини, а не суперпозицією бажань всього людства? Це залежить від того, як оцінювати ці бажання і чи вони ідеалізовані чи ні.

524 Наприклад, тоді як люди для спілкування використовують мову, екземпляри того самого програмного ШІ можуть мати змогу легко і швидко обмінюватися як уміннями, так і знаннями. Тоді синтетичний ШІ може нарешті дозволити собі позбавитися від успадкованих недолугих, і не потрібних у цифровому світі, систем, призначених для роботи з інформацією природного середовища. Такий ШІ може використовувати переваги швидкісних послідовних обчислень, недоступних біологічним мізкам, а також установлювати додаткові модулі з новим спеціалізованим функціоналом (наприклад, оброблення текстів, розпізнавання образів, симуляція, глибинний аналіз даних та планування). Крім того, синтетичний ШІ може мати інші, нетехнічні переваги, наприклад, простіше патентування і менше моральних та етичних обтяжень, як порівняти з емуляцією.

525 Якщо p_1 та p_2 — ймовірності зазнати невдачі на кожному з переходів відповідно, тоді загальна ймовірність невдачі дорівнює $p_1 + (1 - p_1)p_2$, адже її можна зазнати лише один раз.

526 Утім, поза сумнівом, лідер може не мати достатнього відриву від переслідувачів для формування синґлтону. Крім того, також може трапитися, що синґлтон з’явиться ще до ШІ, навіть ЕЦМ, — тоді не залишиться причин прагнути появи ЕЦМ.

527 Чи можна лише вибірково сприяти появі ЕЦМ: так, щоб мінімізувати перетікання зусиль і технологій у суміжну галузь — створення синтетичного ШІ? Наприклад, розвивати технологію мікросканування, а не нейрофункціональне моделювання. (Розвиток апаратних засобів, який і тепер досить швидкий через значний комерційний інтерес, однаково впливатиме на прогрес в обох напрямах).

Водночас швидкий розвиток мікросканування, висока доступність цієї технології може збільшити ймовірність багатополярного сценарію, емуляцій буде багато, вони будуть різноманітні, походитимуть від різних прототипів, а не обмеженої кількості базових образів. Крім того, фактором обмеження може виявитися якраз обчислювальна потужність апаратного забезпечення і це знижуватиме темп переходу до суперінтелектуальності.

528 Нейроморфний суперінтелект, на відміну від емуляцій, може не мати деяких інших атрибутів, які сприяють безпечному функціонуванню. Наприклад, профіль його розумових

переваг та недоліків може дуже відрізнятися від людського (тому ми не зможемо покладатися на власний досвід у наших очікуваннях щодо нього на різних етапах розроблення).

529 Пришвидшити створення ЕЦМ так, щоб воно відбулося раніше за створення ШІ, можна, якщо спочатку просування цими двома напрямами відбуватиметься синхронно, з невеликим випередженням ШІ. Інакше, завдяки додатковим інвестиціям у створення ЕЦМ, воно або відбудеться раніше (з меншим апаратним переважуванням та обмеженим часом на підготовку), але без суттєвих змін у послідовності розроблення, або ефект від інвестицій буде мінімальним (окрім, можливо додаткового стимулювання досліджень у напрямі нейроморфного ШІ).

530 Коментар щодо Hanson (2009).

531 Проте, безперечно, може існувати *такий* екзистенційний ризик, за якого навіть з особистісного погляду буде розумним відтермінувати його, щоб просто виграти побільше життєвого часу для людства або заради спроб зарадити небезпеці.

532 Уявімо, що деякою однією дією ми могли б наблизити вибух інтелекту на один рік. Нехай поточне населення Землі зменшується на один відсоток щороку, а початковий ризик вимирання людства внаслідок вибуху інтелекту становить 20 % (цифра суто ілюстративна). Тоді з особистісного погляду наближення вибуху інтелекту на один рік може бути виправданим, навіть за умови зростання рівня ризику з 20 до 21 %. Щоправда, більшість людей, які житимуть за рік до вибуху, будуть зацікавлені в тому, щоб відкласти його, якщо так вони зможуть зменшити ризик бодай на відсоток (бо вважатимуть, що їхні шанси померти значно менші за один відсоток, що основна частка смертності припадає на людей літнього віку). Отже, можлива модель, коли кожного року населення голосуватиме за відкладення вибуху інтелекту на рік, хоч усі погоджуються, що такий вибух мав би колинебудь статися. Насправді ж, через неузгодженність, непередбачуваність, а також через те, що існують речі важливіші за особисте виживання, з цього порочного кола може бути вихід.

Якщо замість особистісних факторів використовувати стандартний економічний коефіцієнт дисконтування, то привабливість потенційних переваг зменшується, бо цінність астрономічно довгого життя тогочасних людей швидко знижуватиметься. Цей ефект особливо добре видно, якщо оцінювати життя індивідів не за астрономічним часом, а за суб'єктивним. Якщо вартість майбутньої вигоди знижується зі швидкістю x % річних, а фоновий рівень екзистенційного ризику, не пов'язаного зі стрибком інтелектуальності, становить y % річних, то оптимальним для початку стрибка буде момент, коли зниження екзистенційного ризику стрибка від його відтермінування ще на один рік буде менше, ніж $x + y$ %.

533 За допомогу з цим моделюванням я дякую Карлу Шульману та Стюарту Армстронгу. Див. також Shulman (2010 a, 3): «Чалмерс (Chalmers, 2010) засвідчує єдність кадетів та персоналу американської військової академії у Вест-Пойнті щодо того, що, через небезпеку отримання ворожими силами противника вирішальної переваги, уряд Сполучених Штатів не

обмежуватиме досліджень у сфері ШІ, навіть з огляду на їхні потенційні катастрофічні наслідки».

534 А отже, у наведеній моделі інформованість завжди є негативним ex ante фактором. Звісно, це дуже залежить від змісту та складу інформації. Іноді доступність певної інформації може принести велику користь, особливо якщо відрив лідера значно більший, ніж можна припустити.

535 Це навіть може становити екзистенційний ризик для людства, особливо якщо в результаті безпрецедентно зросте кількість озброєнь або з'являться нові військові технології значної руйнівної сили.

536 Учасники проекту можуть фізично бути в багатьох різних місцях по всій планеті і працювати разом, використовуючи зашифровані канали зв'язку. Проте така система має вади безпеки: хоч географічна розподіленість ресурсів може стати запорукою стійкості до військових загроз, вона також може сприяти вразливості окремих осередків до людського фактора, витоку інформації та захоплення ворожими силами.

537 В умовах швидкого знецінення часу проект створення ШІ буде змушений працювати швидко, незважаючи на відсутність будь-яких серйозних суперників. Просто буде невигідно затягувати розроблення. Доведеться полишити безтурботний та неспішний темп досліджень, який відкладатиме революцію ШІ до більш слушного часу (не зважаючи на те, що тоді вона може відбуватися швидше через апаратне переважування). Висока вартість часу — і швидке його знецінення — здешевить також майбутні екзистенційні ризики. Заохочуватимуться ставки, які даватимуть швидкий результат та матимуть високу ймовірність катастрофічних наслідків; як у гонитві за першістю, найбільший дохід приноситимуть інвестиції у швидкість, а не в безпеку. Але, на відміну від конкурентної атмосфери перегонів, швидке знецінення часу (та потреб майбутніх поколінь) не створюватиме особливих конфліктних тенденцій.

Зменшення суперництва та поспіху є основними перевагами співпраці. Співпраця активізує обмін ідеями щодо вирішення проблеми контролю, і це дуже добре. Водночас активніше поширюватимуться ідеї розв'язання проблеми компетентності. У результаті помалу зростатиме колективний інтелект наукової спільноти.

538 З іншого боку, громадський нагляд одного уряду збільшує ризик монополізації результату досліджень однією нацією. Такий сценарій здається гіршим, як порівняти з випадком, коли незалежні альтруїсти забезпечують рівний розподіл результатів. Ба більше, контроль уряду над перебігом проекту не означає навіть, що всі громадяни країни отримають свою частку вигоди. Залежно від країни існує більша чи менша ймовірність того, що всі переваги отримає політична еліта або й ще менше коло дотичних осіб.

539 Є ймовірність, що приваблива упаковка проекту (як описано в розділі 12) може заохочувати інших людей до активнішої участі в роботі, а не пасивної ролі «самотнього воїна».

540 Зменшення віддачі може проявлятися у зменшенні масштабу. Більшість людей швидше погодиться на одну зірку, ніж на одну мільярдну шансу отримати галактику з мільярда зірок,

одну мільярдну ресурсів Землі, ніж одну з мільярдів можливостей володіти всією планетою.

541 Пор. Shulman (2010 a).

542 Якщо серйозно розглядати можливість, що космос може бути нескінченним, агрегативна етика заходить у глухий кут; див. Bostrom (2011 b). Такий самий безглуздий результат може дати ідея використання замість нескінченності довільно велетенських чисел; див. Bostrom (2009 b).

543 Збільшення розміру комп'ютера, зрештою, впирається в релятивістські обмеження, зумовлені затримками з'єднання між різними його вузлами — сигнал не може поширюватися швидше за швидкість світла. Зменшення розміру комп'ютера з іншого боку обмежується квантовими межами. Якщо збільшувати густину комп'ютера, то врешті наштовхуєшся на обмеження чорної діри. Щоправда, не можна стверджувати, що одного дня нові фізичні відкриття не дадуть нам спосіб, як обійти ці перешкоди.

544 Кількість копій індивіда прямо пропорційна кількості ресурсів і не має верхньої межі. Утім не ясно, наскільки корисно звичайній людині буде мати багато власних копій. Навіть якщо така кількість копій комусь і знадобиться, невідомо, наскільки лінійною буде його функція корисності щодо дедалі більшої кількості копій. Як і прожиті роки, кількість копій може мати спадну віддачу у функції корисності звичайної людини.

545 Синґлтон — високоінтегрований на найвищому рівні ухвалення рішень. Водночас якщо на найвищому рівні так буде потрібно, нижчі рівні його структури *можуть* бути дуже розрізнені та конфліктні.

546 Якщо команди-суперники будуть взаємно переконані в некерованості та неспроможності інших до успішного функціонування, одна з причин для співпраці — уникнення негативних факторів суперництва — відпаде: тоді всі команди можуть сповільнити темп, адже будуть переконані у відсутності серйозних суперників.

547 Здобувач звання доктора філософії.

548 У цьому визначенні я волів би, щоб «користь для людства» природно охоплювала добробут тварин, а також інших розумних сутностей (разом із цифровими розумами), які існують зараз або з'являться коли-небудь. Я би не хотів, щоб який-небудь окремий програміст самовільно підміняв загальноприйняті етичні норми власною моральною інтуїцією. Наприклад, запропонований у розділі 12 підхід «когерентного екстрапольованого бажання» з базою екстраполяції, яка охоплює все людство, цілком відповідає цьому принципу.

Також прошу зауважити: це визначення не відкидає можливості існування в постперехідному світі права власності на штучні суперінтелекти, окремі їхні алгоритми й дані. Визначення також агностичне щодо правової чи політичної системи, яка найкраще відповідатиме потребам відносин у гіпотетичному постгуманному суспільстві майбутнього. Воно лише стверджує, що вибір такої системи, який цілком залежатиме від суперінтелекту і того, як ми його створюватимемо, має відбуватися виходячи із зазначених критеріїв: служити на користь людству та відповідати загальноприйнятим етичним нормам — а не будь-кому, хто зуміє долучитися до створення суперінтелекту.

549 Звісно, такі форс-мажорні положення можна вдосконалити. Наприклад, порогове значення можна визначити в доході на душу населення. Можливо, варто під час поділу надлишку віддавати країні-лідеру більшу частку, ніж іншим — щоб заохотити подальше зростання (щось подібне до стратегії «максимін» Ролза). Наступним удосконаленням могла б бути пропозиція відмовитися від долара як мірила й оперувати поняттями «вплив на майбутнє людства» та «вагові коефіцієнти врахування інтересів різних зацікавлених сторін у функції корисності майбутнього синґлтону».

Розділ 15

550 Деякі дослідження корисні не відкриттями, які вони здійснюють. Вони можуть провадитися задля розваги, освіти, акредитації або самоствердження їхніх учасників.

551 Я не вважаю, що *ніхто* не повинен працювати у сферах чистої математики чи філософії. І також не стверджую, що це найбільша з марнот сучасної академічної науки, та й суспільства загалом. Ба, це навіть тішить, що деякі люди повністю віддають себе інтелектуальній роботі та слідують за покликом цікавості туди, куди він їх веде, незалежно від корисності чи важливості об'єкта пізнання. Моя пропозиція полягає в тому, що деякі з найвидатніших розумів сьогодення, усвідомивши минущість своїх видатних розумових здібностей, могли б дещо змістити фокус своїх уподобань на теоретичні проблеми, вирішення яких могло б принести користь уже невдовзі.

552 А втім, варто пам'ятати, що незнання може бути корисним — згадаймо хоч би модель із додатка 13, де ми виявили, що додаткова стратегічна інформованість може зашкодити результату. Загалом варто звертати пильну увагу на інформаційні загрози (див. Bostrom, 2011 b). Можна сказати, що є потреба в глибшому аналізі інформаційних загроз, але є небезпека, що такий аналіз сам собою може стати джерелом небажаної, надлишкової інформації.

553 Пор. Bostrom (2007).

554 Я вдячний Карлу Шульману за це зауваження.

Науково-популярне видання

Бостром Нік

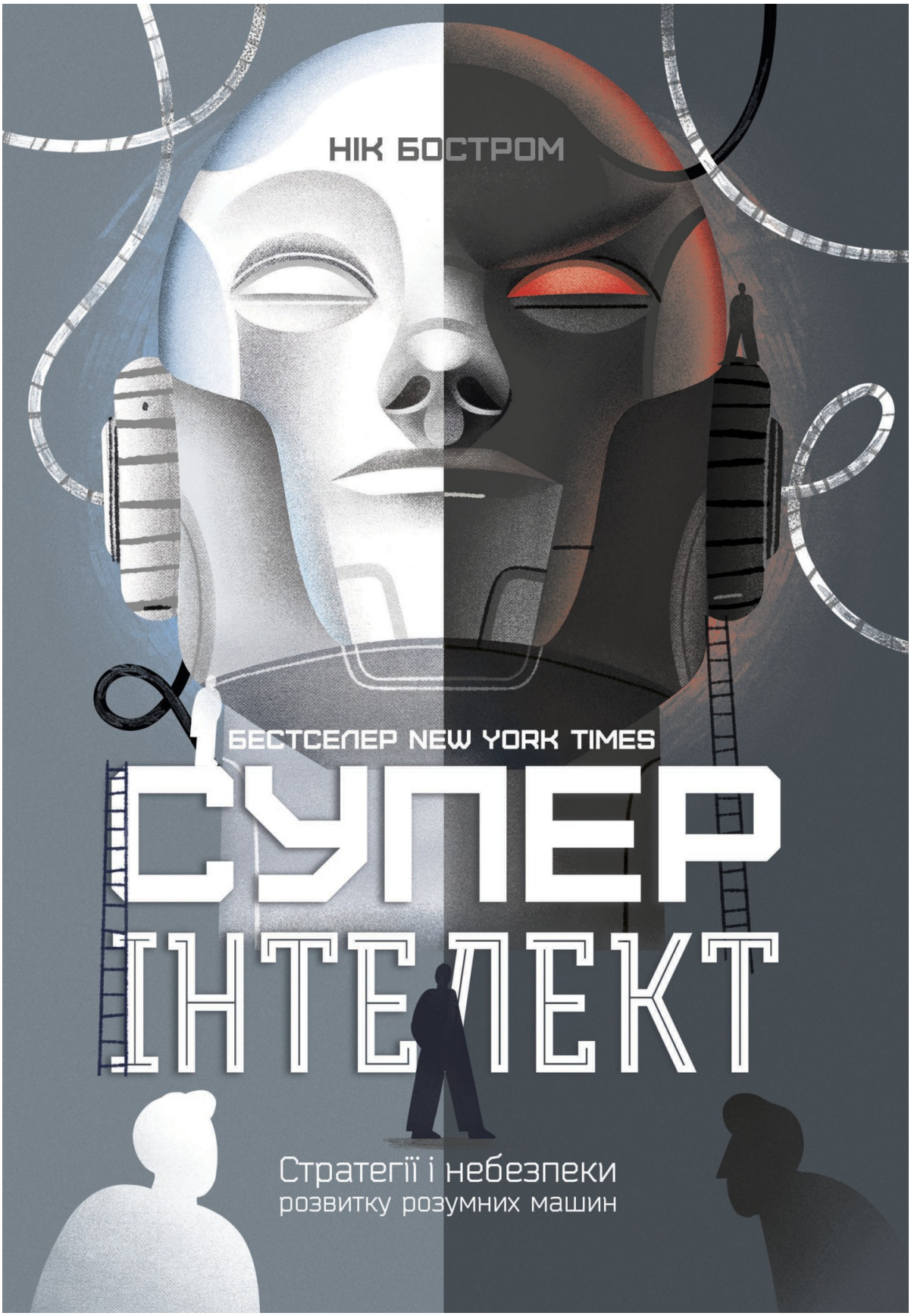
СУПЕРІНТЕЛЕКТ
Стратегії і небезпеки розвитку розумних машин

Керівниця проекту *Галина Харук-Бачуро*
Координаторка проекту *Марина Кичак*
Літературна редакторка *Віталія Євстіфеева*
Наукова редакторка *Тетяна Манжос*
Коректорка *Анастасія Ушинська*
Технічна редакторка *Наталія Коваль*
Верстальник *Владислав Сутковий*
Дизайнерка обкладинки *Юлія Вус*
Художня редакторка *Оксана Гаджій*

Підписано до друку 14.08.2020
Формат 60 × 90/16
Ум. друк. арк. 25,5
Наклад 1000 прим.
Зам. №

Видавець: ТОВ «НФ»
Свідоцтво ДК № 4722 від 19.05.2014
Висновок Держ. сан.-епідем. експертизи № 05.03.02.-04/51017 від 16.11.2015
Пров. Алли Горської, 5, м. Київ, Україна, 01032
Тел. (044) 222-53-49, pub@nashformat.ua

Надруковано в ПП «Юнісофт»
Свідоцтво ДК №5747 від 06.11.2017
Вул. Морозова, 13 б, м. Харків, Україна, 61036


НІК БОСТРОМ
БЕСТСЕЛЕР NEW YORK TIMES
СУПЕР
ІНТЕЛЕКТ
Стратегії і небезпеки
розвитку розумних машин