AI Futures Project

AI 2040

План A — Угода

План розвитку
надлюдського ШІ · ai-2040.com

Дизайн —
Lightcone Infrastructure

Передмова

ШІ-компанії змагаються у створенні ШІ-систем, розумніших за людей у всьому. В AI 2027 ми передбачили, що це призведе або до вимирання, або до незворотної концентрації влади.[1]

План A — це наше позитивне бачення того, що має статися натомість.

У цьому сценарії людство відкладає створення суперінтелекту до 2040 року, робить усі дослідження ШІ публічними, дозволяє десяткам компаній по всьому світу наздогнати передовий рівень і свідомо входить у режим взаємно гарантованого знищення обчислювальних потужностей.

1 Сценарій AI 2027 досі приблизно відповідає тому, яким ми очікуємо побачити майбутнє: шалена гонитва до суперінтелекту, що веде або до захоплення влади ШІ, або до крайньої концентрації влади. Поки що реальність рухається ближче до AI 2027 , ніж очікували навіть ми. (2027 був нашим модальним роком на момент публікації, а не медіанним.) Більше про наші погляди на таймлайни можна прочитати тут і тут.

Що таке План A?

План A — це наше позитивне бачення того, як людство може уникнути екзистенційної катастрофи, спричиненої ШІ, і досягти квітучого майбутнього. Він ґрунтується на розмовах з експертами провідних американських передових ШІ-компаній, безпосередньому досвіді роботи в OpenAI, штабних іграх та обговореннях з політиками, експертами з національної безпеки й лідерами ШІ-політики. Ми рекомендуємо міжнародну угоду, щоб уникнути небезпечної гонки до суперінтелекту. Угода передбачає повну дослідницьку прозорість для AI R&D, що дозволяє країнам світу розуміти, що відбувається, і забезпечувати дотримання запобіжників. У результаті багато компаній у багатьох країнах разом повільно й безпечно масштабуються до суперінтелекту, замість того щоб таємно змагатися одна з одною.

План A — це насамперед рекомендація, а не прогноз. Цей сценарій — *не* наше найкраще припущення про те, яким насправді виявиться майбутнє. Радше це інструмент для донесення і стрес-тестування наших політичних рекомендацій. Хоча *впровадження* Плану A — це рекомендація, а не те, чого ми насправді очікуємо, *зображені подальші наслідки* — це прогнози.[2]

2 Тобто це прогнози того, що сталося б далі, якби наші рекомендації було впроваджено.

У цьому сценарії AI 2040 План A впроваджується успішно, хоча й недосконало і лише в останній момент.

Ми протиставляємо План A чотирьом альтернативним планам (B, C, D і S), які відповідають основним способам, у які США могли б відреагувати (або не відреагувати) на виклики суперінтелекту.[3]

3 Більше про ці чотири плани можна прочитати в точці розгалуження нижче та в нашому доповненні «Наскільки добрий кожен із планів?».

Навіщо ми це написали?

ШІ-компанії, ймовірно, досягнуть своєї заявленої мети — створити ШІ-системи, розумніші за людину, — протягом наступних 1–10 років.

Індустрія переконала себе, що контроль над суперінтелектуальним ШІ можна з'ясувати на ходу, і тому не має бодай віддалено адекватного плану. Ми вважаємо цю ситуацію жахливою: вона запросто може всіх нас убити.[4] Ми не очікуємо, що той, хто «виграє гонку», матиме великий відрив, і не очікуємо, що він односторонньо сповільниться, щоб зменшити екзистенційний ризик.[5] Якщо ця гонка триватиме,[6] ми не очікуємо, що люди збережуть реальний контроль у міру того, як їхні ШІ ставатимуть суперінтелектуальними.[7]

Ба більше, навіть якщо ШІ-компанії якимось чином узгодять свої ШІ, результатом стане безпрецедентна концентрація влади — тобто ситуація, коли крихітна група людей, а можливо, й одна-єдина особа, впродовж кількох місяців фактично контролюватиме єдину у світі армію суперінтелектів, і ці суперінтелекти пропонуватимуть їй різні варіанти подальших дій, частина з яких *де-факто* означатиме захоплення влади над світом.[8]

Наскільки ми можемо судити, CEO OpenAI, Anthropic, xAI та Google DeepMind це розуміють і все одно продовжують — можливо, тому що вважають себе меншим злом і думають, що використають свою величезну владу відповідально, на відміну від Сі Цзіньпіна чи CEO-суперників.[9]

Хоча ми згодні, що загалом правильно обирати менше зло, ми не вважаємо, що варто просувати стратегію з настільки лячно високим шансом призвести до вимирання людства чи глобальної диктатури. Натомість ми хочемо виступати за те, що є справді добрим. Якщо так робитиме достатньо людей, це може статися.

Тож ми написали сценарій, що окреслює такий можливий світ.

Чому саме сценарій?

«Плани нічого не варті, але планування — це все». — Дуайт Д. Ейзенхауер

Ми вважаємо, що більшість пропозицій ШІ-політики розсипаються під сценарною перевіркою — тобто якщо спробувати виписати детальний і правдоподібний сценарій, у якому ця пропозиція досягає успіху, зробити це виявиться складно, і ви зрозумієте, що план має менше шансів спрацювати, ніж здавалося, або більше неприємних побічних ефектів, ніж визнавали його прихильники.

4 Звісно, ми не знаємо точно, наскільки ймовірне буквальне вимирання людства через ШІ, але оцінки в нашій команді коливаються від 10% до 30%. Головна причина в тому, що є багато іншого, що неузгоджені ASI могли б зробити з нами після захоплення влади, окрім як убити всіх. Наприклад, вони можуть залишити частину людей живими, бо це дуже дешево, а їм не зовсім байдуже, або тримати нас як акаузальні козирі для торгу. Утім, ми не думаємо, що цей нюанс змінює підсумок: захоплення влади неузгодженим ШІ — це, найімовірніше, дуже погано.

5 Я (Daniel Kokotajlo) нещодавно виступав перед приблизно 100 людьми, з яких ~40% були з передових ШІ-компаній, і провів опитування підняттям рук: (a) скільки місяців відриву матиме провідна ШІ-компанія на момент, коли вона вперше повністю автоматизує AI R&D, і (b) яку частину свого відриву вона була б готова спалити. Медіанні відповіді — приблизно 3 місяці та 1/3 відповідно. Це збігається з моєю власною оцінкою.

6 А ще — умови жорсткої секретності й групового мислення! За замовчуванням у цей період компанії будуть ще закритішими, ніж у 2026-му. Коли вони кажуть «ми розв'язуватимемо проблеми узгодження в процесі», насправді це означає: «Перевантажена крихітна група експертів з узгодження в нашій компанії розв'язуватиме проблеми узгодження в процесі, без можливості винести свою роботу на зовнішнє рецензування».

Можливо, саме тому сценарна перевірка така рідкісна в ШІ-політиці. Кожен хоче казати, що його улюблені політики матимуть чудові наслідки, а політики суперників — жахливі. Сценарна перевірка власних улюблених політик може виявити незручні проблеми з ними; тим часом сценарна перевірка політик суперника — це багато роботи заради мізерного риторичного зиску.[10]

Ми вважаємо, що дискурс поліпшився б, якби більше пропозицій ШІ-політики проходили сценарну перевірку. Тож ми починаємо з власної — хоч це й відкриває нас для критики. Сподіваємося, критики судитимуть нас на тлі наявного state-of-the-art серед планів проходження ШІ-переходу (якщо зможуть такий знайти), а не на тлі туманної, але приємної фантазії, де нікому не доводиться робити важких виборів, і все, мабуть, буде добре.

А як щодо колосальної складності передбачити, який ефект наша політика матиме у світі, що наближається до надлюдських ШІ? Це як намагатися передбачити, як найкраще вести Третю світову війну, тільки з іще більшим відривом від минулих кейсів. Проте спробувати все одно цінно — так само, як американським військовим цінно до найдрібніших деталей програвати тайванські сценарії. Є й інші прецеденти: розвідувальні агентства, кліматичні органи та офіси готовності до пандемій — усі вони спираються на різні види сценарного планування.

Таймлайни ШІ?

План А — наша амбітна пропозиція щодо того, що робити, і ми хотіли б, щоб щось подібне впровадили якнайшвидше, бо ми не впевнені, скільки часу залишилося.[11] Але щоб написати конкретний сценарій, нам потрібен конкретний таймлайн.

Таймлайн цього сценарію такий:

- У 2029 році США та Китай домовляються уникнути безрозсудної гонки до суперінтелекту.
- **У 2030-му ми *мали б* повністю автоматизувати AI R&D**, що вело б до суперінтелекту до кінця року. **Завдяки Угоді ми цього уникаємо**.
- Між 2030 і 2035 роками ми масштабуємося в межах людського діапазону — до ШІ-систем, приблизно настільки ж здібних, як топові людські експерти.
- У 2035-му ми ставимо паузу на рівні ШІ топового людського експерта, щоб зберегти людський контроль.
- У 2040-му ми знімаємо паузу й масштабуємося до суперінтелекту.[12] (**Звідси й назва**: AI 2040)

7 Ми очікуємо, що вони збережуть **видимий** контроль — уточнимо, принаймні на якийсь час, — і почасти саме тому ця проблема така загрозлива. Ситуація була б набагато кращою, якби ми мали надійний спосіб визначити, чи ШІ-система справді стійко слухняна/лояльна/узгоджена тощо, чи лише тимчасово така, чи просто прикидається.

8 Фінал «Сповільнення» в AI 2027 зображує саме це. Його доволі легко уявити, бо там немає нічого карикатурно чи відверто злого — лише низка кроків, кожен з яких можна виправдати окремо (консолідація американських обчислювальних потужностей в єдиний пул, перемога над Китаєм, інтеграція ШІ в уряд та економіку, недопущення того, щоб терористи здобули власні AGI, тощо).

9 Наприклад, ці старі імейли OpenAI показують, що OpenAI було засновано значною мірою через страх, що Demis Hassabis, CEO DeepMind, використає AGI, щоб стати диктатором. Див. також відповідні матеріали в Empire of AI. Див. також це і це і це.

10 По-перше, важко застосувати сценарну перевірку до політичної пропозиції, якщо розумієш її лише поверхово, а люди зазвичай розуміють власні улюблені ідеї набагато краще, ніж ідеї, які ненавидять. По-друге, доказова сила сценарної перевірки за своєю природою більша для сценаріїв, написаних прихильниками, ніж опонентами: якщо прихильник політики щосили намагається зобразити її успіх — і в нього не виходить, це набагато вагоміший доказ, ніж коли опонент політики зображує її провал.

11 Тобто ми не впевнені, наскільки швидко прогресуватимуть можливості ШІ — наприклад, скільки часу лишилося до того, як ШІ-компанії зможуть автоматизувати власні дослідження й почнуть прискорюватися до суперінтелекту.

У нашому попередньому сценарії, AI 2027, ШІ повністю автоматизував процес створення розумніших ШІ у 2027 році, що призвело до вибуху інтелекту та суперінтелекту в межах року. Дві відмінності цього сценарію: (1) дефолтний таймлайн тепер — 2030 рік, і (2) завдяки діям у сфері врядування загалом надлюдські ШІ вперше з'являються у 2040-му.

Ми змінили дефолтний таймлайн, **бо хочемо**, **щоб наш портфель сценаріїв відображав нашу невизначеність щодо таймлайнів ШІ**. Рік у назві AI 2027 обрали тому, що на момент, коли ми почали писати, Daniel вважав: є приблизно 50% шансів, що все піде так швидко або швидше.[13] На момент, коли ми почали писати План А, відповідним роком для Thomas був 2030-й. **Daniel наразі вважає**, **що все**, **ймовірно**, **піде дещо швидше**, **ніж зображено в цьому сценарії**; більше про погляди нашої команди на таймлайни можна прочитати тут і тут.

Ми змінили дії у сфері врядування, бо **цей сценарій — насамперед рекомендація**, **а не прогноз**. Проводити вибух інтелекту на повній швидкості — дико безрозсудно, і це концентрує владу до крайньої міри.

12 Суперінтелект означає ШІ-системи, які значно кращі за найкращих людей у всьому, а до того ж швидші й дешевші.

13 На момент завершення роботи над AI 2027 медіана Daniel щодо AGI була 2028 рік. Медіани інших авторів були між 2031 і 2035. Докладніше про те, як погляди Daniel та Eli змінювалися з часом, див. тут. Ми регулярно оновлюємо наші прогнози тут, з відповідними дописами в блозі на нашому Substack.

2027: Письмена на стіні

В Америці тепер дві робочі сили. Перша — це люди, їх 165 мільйонів. Друга — ШІ-агенти: мільйони копій, що запускаються і вимикаються щогодини, працюючи цілодобово на надлюдських швидкостях.

Більшість їхньої роботи — слоп. Але достатня її частина добра — настільки, що люди платять десять мільярдів доларів на місяць за ШІ, здатні, принаймні в теорії, робити на комп'ютері будь-що з того, що може працівник.

Є одна робота, яку ШІ-компанії хочуть автоматизувати понад усі інші, — їхня власна. Поки що їм це не вдалося; жодного рекурсивного самовдосконалення наразі немає.[14] Але, схоже, вони наближаються — і затягують драбину за собою: найсильніші ШІ для кодингу відмовляються допомагати конкурентам з AI R&D.[15] Навіть коли найоптимістичніші співробітники визнають, що все триває трохи довше, ніж планувалося, скептики помічають, що їхні звичні відмахування починають звучати порожньо. Чому саме ШІ ніколи не зможе виконувати мою роботу? Який там, ще раз, бар'єр?

Конгрес починає звертати більше уваги. Про ШІ там чують уже давно: дата-центри споживають забагато води,[16] чат-боти підштовхують до самогубства, Mythos зламує системи АНБ — і, звісно, лобісти технологічної індустрії застерігають, що найменший натяк на регулювання змусить Америку миттєво програти гонку Китаю і провести решту історії як держава-данник КПК.[17]

Тепер вони роблять крок назад і питають: куди ми з цим ідемо? Яким буде світ через п'ять, десять чи п'ятнадцять років? Чи будуть іще робочі місця? А що, як ні?

Одне питання тисне на них особливо сильно: **Хто контролюватиме всі ці ШІ?**

Конгрес доходить важливої частини відповіді: **Мабуть, не ми**.[18]

Вони проводять серію напружених слухань щодо ШІ. Читають імейли OpenAI 2016 року, де обговорюється, що OpenAI було засновано, аби не дати Demis Hassabis стати диктатором.[19] Але хто не дає стати диктатором Сему чи Ілону? Наявні відповіді Конгрес не задовольняють.

Результатом цього пробудження стає AI Transparency Act 2027 року — всеосяжний закон, який робить багато різного, щось добре, а щось погане, але фундаментально ситуації не змінює.[20]

14 ШІ допомагають робити майбутні ШІ здібнішими, але розробка ШІ досі потребує людей.

15 Ця політика егоїстично вигідна провідним ШІ-компаніям, оскільки ускладнює відсталим компаніям наздоганяння, але вона також корисна для безпеки, бо сповільнює незворотне поширення алгоритмічних інсайтів.

16 Ми вважаємо, що занепокоєння щодо поточного споживання води ШІ здебільшого перебільшені. Див. цю статтю , щоб дізнатися більше. Утім, як стане видно далі в сценарії, ми справді вважаємо, що вплив ШІ на довкілля протягом наступного десятиліття-двох буде величезним і що нестримне зростання буквально закип'ятило б океани. Див. додаток про енергію в нашому економічному доповненні , де про це докладніше.

➤ *Див.* Додаток А — Список побажань щодо поступової ШІ-політики *, де більше деталей.*

2028: ШІ у виборчому бюлетені

Виборчий цикл 2028 року, як завжди, розпечений. ШІ — найбільша тема. Дата-центри, що зараз будуються, коштують удвічі більше за весь військовий бюджет США.[21]

Більшість офісних професій переживають дизрапшн, подібний до того, що спіткав розробку ПЗ у 2026-му; така робота тепер значною мірою полягає в керуванні ШІ-агентами. ШІ-компанії поставили процес навчання на індустріальні рейки: керівники кажуть «цього року заходимо в [професію]» — і компанія проводить інтерв'ю з фахівцями, купує дані, створює навчальні середовища тощо, поки їхні ШІ не наберуть обертів. Далі ШІ швидко вдосконалюються в міру того, як їх ширше використовують у полі й вони накопичують більше даних з реального світу.

Інші країни починають лякатися і злитися. Схоже, жменька американських і китайських компаній ось-ось автоматизує всі офісні робочі місця. Влада концентрується у США — зокрема в руках Президента та кількох CEO технологічних компаній.

Експерти з ШІ попереджають, що вибух інтелекту вже близько. Прискорюючи дослідження ШІ, ШІ-системи ставатимуть ще компетентнішими, прискорюючи дослідження ще швидше, що робитиме їх ще *більш* компетентними, і так далі. Є складна динаміка з вузькими місцями та апаратними обмеженнями, що визначає, наскільки швидко йде цей процес і де він завершується, але, схоже, він може піти дуже швидко й завершитися десь дуже далеко.

На дефолтному шляху вже в наступну президентську каденцію з'являться ШІ, що далеко перевершують людський рівень, — створені повністю ШІ, які самі створені повністю іншими ШІ, без жодної людини в циклі вже кілька поколінь поспіль. Чи будуть ті ШІ слухняними, узгодженими тощо? Чому? І якщо так — хто їх контролюватиме? Як саме все це має добре закінчитися?

Поставивши людство на цей шлях, ШІ-компанії вважають його прийнятним. Але більшість людей — ні. Годі вже про його *спадщину* — Президент починає думати про те, що буде з *ним самим* після того, як він піде з посади, а світ трансформується.[22] Обох кандидатів у президенти постійно питають, що вони робитимуть із ШІ, і під час кампанії ті випробовують дедалі драматичніші ідеї. Дискурс кидається туди-сюди між усіма варіантами, показаними нижче, і не тільки.

17 Чи справді ми в такій гонці? Певною мірою, але не зовсім. На нашій дефолтній траєкторії ШІ незабаром стане дуже потужним — аж настільки, що буде стратегічно важливішим за ядерну зброю; це правда. Багато людей, включно з CEO передових ШІ-компаній, щосили намагаються створювати дедалі розумніші ШІ-системи раніше за конкурентів. Однак надмірна параноя щодо конкурентів — добре відоме упередження. Припущення, що ми в гонці, де переможець забирає все, відкидає багато можливих шляхів співпраці, створюючи самоспрaвджувану пастку. Ми очікуємо, що масштаб і швидкість прогресу ШІ робитимуть дедалі очевиднішим для осіб, які ухвалюють рішення, і в США, і в Китаї, що поточний темп прогресу та брак довіри між ними надзвичайно небезпечні. Ба більше, двом країнам, можливо, вдалося б скоординувати сповільнення й без жодної юридично зобов'язальної угоди — якщо вони збудують довіру, поступово деескалуючи кожна зі свого боку. Запропонований нами План А слід тлумачити як підхід, що міг би спрацювати, *навіть якщо* США й Китай мають абсолютно нульову довіру одне до одного, а не як твердження, що суворі протоколи верифікації точно необхідні.

18 Тобто не Конгрес. За замовчуванням тими, хто контролюватиме ШІ — якщо взагалі хтось по-справжньому контролюватиме, — будуть ШІ-компанії або, можливо, Білий дім.

19 Від Ilya та Greg: «Мета OpenAI — зробити майбутнє добрим і уникнути диктатури AGI. Ви занепокоєні, що Demis міг би створити диктатуру AGI. Ми теж. Тому створювати структуру, в якій ви могли б стати диктатором, якби захотіли, — погана ідея, особливо з огляду на те, що ми можемо створити іншу структуру, яка виключає таку можливість».

20 Ми не маємо на меті *рекомендувати* погане; ми *прогнозуємо* , що у 2027 році відбудеться суміш добрих і поганих поступових реформ. Більше про те, як ми балансуємо між прогнозами й рекомендаціями в цьому сценарії, див. тут.

Зрештою Президент і його протеже сходяться на одному плані; кандидат від опозиції — на іншому. А тоді настає день виборів.

2029: Оберіть шлях

«Довіряй, але перевіряй» — Рональд Рейган

Президент оголошує, що США прагнутимуть міжнародної співпраці, аби уникнути вибуху інтелекту, що насувається.

«Ця божевільна гонка до суперінтелекту має закінчитися. Надто довго ми обирали менше зло та найменш погані рішення. Нам потрібен План А. Ми можемо продовжувати розвивати ШІ, але мусимо робити це *обережніше, прозоріше* і залучаючи *набагато більше країн і компаній*».

На подив багатьох у Вашингтоні, Китай виявляється сприйнятливим. По свій бік Тихого океану вони обговорювали ті самі проблеми — соціальну дестабілізацію, втрату робочих місць, некерований суперінтелект. Вони очікували на «китайське століття» і думали, що ШІ може зірвати їхні плани. І мали ще одну причину сісти за стіл переговорів: США й далі мають більше обчислювальних потужностей, більше дата-центрів і кращі моделі. Їх непокоїло, що США можуть із ними зробити, якщо першими досягнуть суперінтелекту.[23]

➤ *Див.* Додаток B — Чому Китай був би зацікавлений в угоді? Чому взагалі хтось був би? *, де більше деталей.*

США й Китай не довіряють одне одному. На щастя, їм і не треба: План А містить положення про верифікацію дотримання. Але його розгортання потребуватиме часу.

Тож поки що вони починають із чогось грубого. На решту 2029 року вони тимчасово зупиняють навчання ШІ, бо це відносно легко верифікувати. У 2030-му вони вже матимуть готову інфраструктуру, щоб перейти до Плану А.

2029: Квапливі переговори

США й Китай узгоджують серію домовленостей, що зрештою виливаються в План А.

- Декларація обчислювальних потужностей
- Пауза в навчанні
- Заручитися світовою підтримкою

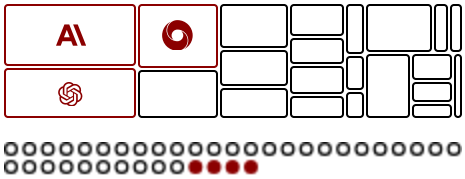
2030: План А запроваджено

Консорціум узгоджує керівні принципи.

- Виграти час
- Повна дослідницька прозорість
- Широко поширити ШІ
- Зворотність

2028

Рівень зайнятості	62 %
Медіанний дохід	US$49k
Дослідники узгодження	1k
Сумарне сповільнення	0 міс



○ **3.5bn** Праця людей ····· ×1 швидкість

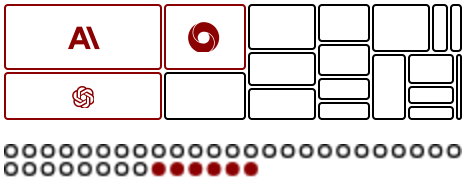
● **23m** Надійні агенти ····· до ×130 швидкості

21 Тобто ШІ-індустрія витрачає $2.4T капітальних видатків (CapEx) за 2028 рік — утричі більше, ніж було витрачено у 2026-му. Військовий бюджет США у 2025 році становив близько $1T. Найбільша ШІ-компанія на початку 2028 року має $360B ARR і далі зростає з річним темпом приблизно 150%, що виводить її на траєкторію стати водночас найдорожчою компанією світу та найбільшою компанією світу за виручкою протягом 1-2 років.

22 Зрештою, після відходу з посади він усе ще буде живий. Як і всім нам, йому доведеться пережити всі кризи, які принесе суперінтелект, і сподіватися, що новий Президент дасть їм раду й не стане диктатором. У його інтересах і підготувати нового Президента до успіху щодо проблеми втрати контролю, і встановити стримування та противаги, щоб запобігти крайній концентрації влади.

2029

Рівень зайнятості	62 %
Медіанний дохід	US$50k
Дослідники узгодження	2k
Сумарне сповільнення	0 міс



○ **3.4bn** Праця людей ····· ×1 швидкість

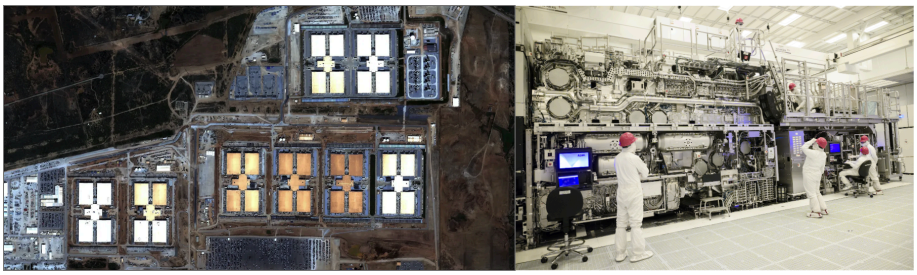
● **29m** Надійні агенти ····· до ×150 швидкості

Фондовий ринок дико гойдається вгору-вниз у відповідь на новини й коментарі про епохальні кроки, що їх ухвалюють. Усі кричать, що це заходить надто далеко або недостатньо далеко. До кінця року ринок заспокоїться, але крики триватимуть.[24]

Крок 1: **Декларація обчислювальних потужностей**

Жодна з країн не хоче зупиняти власне навчання ШІ, якщо не бачитиме, що інша сторона теж зупиняється.

На щастя, навчання ШІ потребує великої кількості ШІ-чипів. Більшість ШІ-чипів перебувають у гігантських дата-центрах.[25] ШІ-дата-центри зазвичай достатньо великі, щоб їх було видно з космосу, і достатньо енергоненажерливі, щоб потребувати помітної інфраструктури. Нові ШІ-чипи можна виготовляти лише на жменьці заводів із виробництва чипів (фабів), розташованих переважно на Тайвані, у Південній Кореї, США та Китаї.



Ліворуч: дата-центр Stargate компанії OpenAI. Праворуч: машина екстремальної ультрафіолетової літографії (EUV) — імовірно, найскладніша машина, яку будь-коли створило людство. EUV-машини необхідні для виробництва передових ШІ-чипів, і їх виробляє одна-єдина нідерландська компанія — ASML.

США й Китай ведуть переговори з країнами, які відіграють значну роль у ланцюжку постачання чипів, і вимагають від кожного великого власника дата-центрів (та їхніх постачальників вище по ланцюжку, включно з фабами) публічно декларувати свої великі закупівлі та продажі.[26] Аналітики з численних країн та інституцій можуть прискіпливо вивчати список, ставити запитання й оскаржувати аномалії. Держави-суперниці також надсилають інспекторів на інфраструктуру одна одної, переконуючись, що заявлені числа точні. До кінця року кожна сторона впевнена, що інша не приховує більше ніж приблизно 1% ШІ-обчислень і що всі потенційні джерела нових обчислювальних потужностей відстежуються та обліковані.[27]

Крок 2: **Пауза в навчанні**

Щоб відрізняти небезпечні ШІ від безпечних, знадобиться більше часу й розуміння, тож наразі вони обирають просте рішення: тимчасову паузу на всі нові навчальні запуски.[28]

23 Якщо розгорнути: вони побоювалися, що США використають свою величезну перевагу в ШІ, щоб підірвати китайські ШІ-проєкти кібератаками, фізичними диверсіями чи іншими засобами, а потім скористаються ще більшою і тривалішою перевагою в ШІ, щоб диктувати умови КПК або й узагалі її повалити. Запобігання ризику втрати контролю було лише додатковим бонусом.

24 Геополітично ситуація стабілізується за кілька років, але на той момент прогрес і поширення ШІ вже відбудуться й породять нові проблеми та кризи, якими людям доведеться перейматися.

25 Нашу оцінку розподілу розмірів дата-центрів ми наводимо в доповненні про обчислювальні потужності.

26 Початкова декларація чипів стосувалася б лише компаній, чиї закупівлі перевищують 10k H100e (що коштує ~$100M). Протягом року взято на облік активи кожної компанії, яка володіє понад 0.001% світових обчислювальних потужностей (~2,800 H100e, вартістю ~$18M на той час). Мертві чипи також зберігаються або верифіковано знищуються.

Обидві сторони й далі можуть використовувати дата-центри для запуску вже наявних ШІ-систем (тобто для інференсу), але вони дообладнають дата-центри одна одної пристроями для верифікації того, що ті не використовуються для нових запусків навчання.[29]

США і Китай здатні швидко дістати достатньо таких верифікаційних пристроїв, щоб дообладнати майже кожен великий дата-центр.[30]

➤ *Див.* Додаток C — Що якби верифікація «тільки інференс» не була готова? *для деталей.*

Інспектори з обох країн підтверджують встановлення верифікаційних пристроїв. Лише дата-центрам, щодо яких США і Китай згодні, що верифікація там працює, дозволено надалі надавати ШІ-сервіси. Тим часом обидва уряди щодуху працюють над більш детальними опціями верифікації, які дозволили б відновити навчання, зберігаючи в обох сторін упевненість, що інша не вириває вперед.

Хронологія впровадження Угоди – ai-2040.com

Крок 3: Заручитися підтримкою всього світу

Хоча з іншими країнами консультувалися від самого початку — особливо з тими, що володіють частинами ланцюга постачання ШІ, — План А почався як двостороння угода між США і Китаєм. Тепер, коли він набуває чіткіших обрисів, переговори стають багатосторонніми. У хід ідуть батоги і пряники.

Багато країн поза США і Китаєм раді угоді, бо їх непокоїло майбутнє, в якому жменька американських і китайських ШІ-компаній рекурсивно самовдосконалюється, відривається дедалі далі вперед, а потім… Ну, що станеться далі — залежить від того, кого спитати, але відповіді охоплюють варіанти на кшталт «заберуть наші робочі місця», «навічно зацементують гегемонію» і «всіх повбивають». Тож вони хочуть, щоб США і Китай рухалися повільніше і прозоріше. Це розвіє їхні страхи, а заразом дозволить їхнім власним ШІ-проєктам наздогнати передовий рівень.

Тож США і Китаю напрочуд легко заручитися підтримкою.[31] До кінця року більша частина світу приєдналася до того, що тепер називають Консорціумом.

27 Той 1% продажів обчислювальних потужностей, який не вдається відстежити до кінцевого місця призначення, здебільшого припадає на контрабандистів, що не фіксували особи своїх клієнтів, і на списані чипи. Половину цього пулу згодом знаходять установленою в китайських дата-центрах, виявлених за супутниковими знімками; обидві країни пропонують грошові стимули та юридичну амністію будь-яким власникам решти 0.5%. Ці та інші шляхи пошуку прихованих обчислювальних потужностей ми розбираємо набагато детальніше в нашому доповненні про таємний проєкт.

28 Це охоплює й ті дослідницькі експерименти, що передбачають навчання. Хоча теоретично вибух інтелекту можливий і без нових навчальних запусків чи великих експериментів, це видається малоймовірним, бо вимагало б екстремальної зміни парадигми.

29 Запропоноване нами рішення для верифікації «тільки інференс» працює так: береться випадкова вибірка з невеликого відсотка робочих навантажень і перевіряється, що вони справді виконують інференс на схваленій моделі (шляхом повторного обчислення виходів із використанням входів і схвалених ваг моделі). Для збору цих випадкових вибірок наша пропозиція передбачає прості мережеві відводи, які перенаправляють увесь вхідний і вихідний трафік дата-центру на сервер повторного обчислення (всередині дата-центру), встановлений іншою стороною, — саме він виконує випадкове часткове повторне обчислення. Можуть існувати й інші, потенційно кращі рішення для цієї пропозиції «тільки інференс», як-от криптографічні схеми proof of work, але з погляду продуктивності вони наразі більш спекулятивні. Детальніше ми обговорюємо це в нашому доповненні про верифікацію. Щоб цей підхід працював, потрібно кілька додаткових рівнів забезпечення дотримання, зокрема фізична безпека, програмний моніторинг і люди-аудитори.

Альтернативна лінія часу

Китай робить спробу таємного AGI-проєкту

➤ *Див. Додаток D для цієї гілки.*

2030: План A запроваджено

Техкомпанії щосили тиснуть, щоб відновити навчання ШІ, і лобіюють умови, на яких це відбудеться. Значна частина громадськості каже, що відновлювати не можна *ніколи* . Так само кажуть і деякі країни.

Після року інтенсивних дискусій, переговорів і сварок у твітері вимальовується компроміс: розробка ШІ триватиме, але обережніше, прозоріше і більш розподілено.

План A керується 4 базовими принципами:

Виграти час Сповільнюватися щоразу, коли це потрібно, щоб мати високу впевненість у безпеці.

Повна дослідницька прозорість Зробити майже всі дослідження ШІ повністю видимими для публіки.

Широко поширювати ШІ Багато компаній у багатьох країнах на передовому рівні.

Зворотність Обмежити алгоритмічний прогрес; підтримувати взаємно гарантоване знищення обчислювальних потужностей.

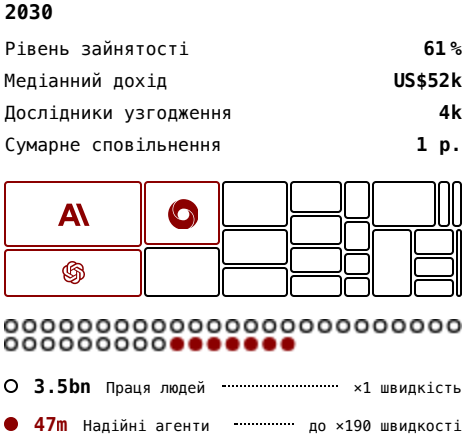
Принцип 1: Виграти час

Проблема вибуху інтелекту — у слові «вибух».

Траєкторія за замовчуванням — безрозсудна й дестабілізаційна. Кожна країна і компанія женеться з конкурентами за домінування у ШІ, не маючи часу подумати. Найімовірніше, купа людей у підсумку втратить вплив або загине. Можливо, богоподібні ШІ-системи легко

30 Це полегшується завдяки квітучій нині екосистемі компаній верифікаційного обладнання, що виникла після того, як обидва уряди засигналізували про інтерес до верифікаційної спроможності як опції (це, своєю чергою, привабило інвестиції філантропів і венчурних фондів та технічну підтримку компаній ланцюга постачання ШІ). Якби цього не сталося, ця частина Плану A зайняла б дещо більше часу та/або була б дорожчою і гарячковішою. У найгіршому разі дата-центри довелося б вимкнути замість того, щоб дозволити їм виконувати інференс.

31 Ця частина — прогноз, а не лише рекомендація. Ми вважаємо, що заручитися підтримкою Плану A за обставин, подібних до описаних нами у 2029 році, було б значно легше, ніж очікує більшість наших читачів.



нас переграють, і ера людства скінчиться. Можливо, якийсь CEO чи президент встановить постійну диктатуру. Можливо, у хаосі почнеться Третя світова війна, бо країни без найкращих ШІ боятимуться втратити свої важелі. Можливо, все пройде напрочуд гладко, і ми «лише» автоматизуємо всю працю та зробимо всіх непрацевлаштовними в радикально зміненому світі. Хвилюються навіть претенденти на роль ШІ-диктаторів: лідерів перегонів достатньо багато, щоб ні в кого з них не було понад 50% шансів виграти боротьбу за владу.

Ніхто не знає, як визначити, чи можна довіряти ШІ-системам, і ніхто не знає, як регулювати суперінтелектуальний ШІ. Схоже, розв'язання цих проблем займе роки. Сповільнення вибуху інтелекту виграє час, щоб це зробити.

Це також допомагає запобігти екстремальній концентрації влади, бо дає більше часу групам, які безпосередньо не контролюють найрозумніші ШІ світу, прокинутись і усвідомити, що відбувається, поки вони не втратили своїх важелів. Без суттєвого сповільнення стрімкий вибух інтелекту сконцентрує майже всю владу в руках крихітної групи CEO та політиків.

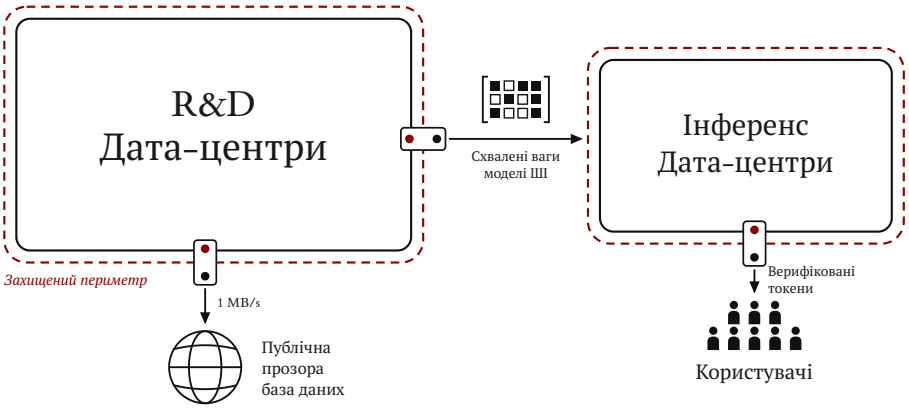
➤ *Див*. Додаток E — Чому б просто не зупинитися повністю? *для деталей*.

Принцип 2: Повна дослідницька прозорість

Отже, мета — розвивати ШІ в розумному темпі: не надто швидко, не надто повільно. Заборонити небезпечні види, дозволити безпечні. Але кожна країна досі сама регулює свою ШІ-індустрію; немає центрального планувальника з повноваженнями регулювати ШІ в усьому світі. Тож як вони вирішують, що дозволяти, а що ні, і як верифікують, що ніхто не зрізає кути?

Країни Консорціуму придумують просту високорівневу рамку: домовляємося показувати одне одному всі дослідження ШІ. А коли нам не подобається те, що хтось робить, ми це обговорюємо і, можливо, домовляємося заборонити.

➤ *Див*. Додаток F — Чому саме стільки прозорості? Чому не менше і не більше? Відкритий доступ у Плані A *для деталей*.



Робочі навантаження ШІ можна поділити на дослідження й розробку (процес створення нових ШІ-систем) та інференс (використання вже наявних ШІ). У нашій пропозиції дослідження майже повністю прозорі, тоді як інференс залишається приватним. Детальніше про те, як це могло б працювати, див. наше доповнення про прозорість.

Архітектура верифікації дата-центрів – ai-2040.com

Прозорість дає багато переваг. По-перше, коли кожна компанія бачить, що роблять інші, експертиза урядів стає менш критичною. Якщо компанія починає робити щось небезпечне, компанії-конкуренти, сторонні аудитори тощо можуть це помітити й здійняти тривогу. Це і кардинально збільшує обсяг інтелектуальних ресурсів, спрямованих на безпеку передових ШІ, і запобігає ситуації, коли єдині учасники технічної розмови — упереджені працівники ШІ-компаній та перевантажені регулятори, яких значно менше. Крім того, більша видимість спрощує задачу верифікації дотримання регуляцій, особливо у випадках сірої зони.

По-друге, повна дослідницька прозорість робить майже неможливим навмисне вбудовування в ШІ прихованих лояльностей, упереджень чи порядків денних під час навчання. Усі бачать, як саме навчається кожен передовий ШІ. Загалом прозорість полегшує групам, які безпосередньо не контролюють передові ШІ, помічати зловживання владою з боку тих, хто контролює, і запобігати їм.

По-третє, більше немає такого стимулу мчати до відкриття нових парадигм ШІ і потужніших алгоритмів, адже компанії не змогли б притримувати такі відкриття і сильно на них заробляти. Це виграє світові час.

➤ *Див.* Додаток H — Як змінюються стимули за Повної дослідницької прозорості *для деталей.*

Принцип 3: Широко поширювати ШІ

Цей принцип здебільшого досягається взаємодією двох попередніх:[32] Оскільки алгоритмічні секрети оприлюднюються, а темп прогресу не прискорюється, інші компанії наздоженуть передовий рівень. Результатом стане конкурентний ринок ШІ, на якому споживачі ШІ-сервісів мають багатий вибір і чудову видимість того, що саме вони купують. Оскільки навчання передових моделей повністю прозоре, люди можуть верифікувати, що опублікований Спек точно описує цілі та цінності, які закладаються в модель під час навчання.

Це дуже важливо для запобігання екстремальній концентрації влади. Але це також допомагає прискорити використання ШІ для розв'язання найгостріших проблем світу. Це допомагає з ризиками втрати контролю, прискорюючи дослідження узгодження і контролю ШІ. Це допомагає прискорити розробку нових і кращих технологій верифікації, а також необхідні інституційні реформи для стабілізації угоди та підготовки до надлюдського ШІ.

Це повна протилежність кошмарному сценарію, якого боялися у 2020-х: один–три ШІ-проєкти таємно змагаються між собою, тримаючи свої найкращі моделі лише для внутрішнього використання і застосовуючи їх для автоматизації досліджень ШІ раніше, ніж для будь-чого іншого. У тому сценарії широкий світ у темряві щодо того, наскільки потужними стають ШІ, а ШІ розгортаються в найризикованішій сфері (AI R&D) раніше, ніж деінде. Тепер усе навпаки.

Принцип 4: **Зворотність**

У минулому компанії навчали більші й кращі ШІ, використовуючи *як* масштабування обчислювальних потужностей (більші запуски навчання), так і програмний прогрес (удосконалення алгоритмів ШІ — нові парадигми, кращі рецепти навчання, кращі дані тощо). Тепер Консорціум намагається скеровувати процес так, щоб більшість покращень походила зі збільшення обчислювальних потужностей для навчання.[33]

➤ *Див.* Додаток I — Будувати більше дата-центрів зазвичай погано, але добре в цій конкретній ситуації *для деталей.*

Алгоритми — це інформація; зупинити їхнє поширення від природи важко, а повна дослідницька прозорість означає, що ми навіть не намагаємося.[34] Щойно нову парадигму відкрито, вона одразу потрапить до гіпотетичних таємних проєктів, і скасувати це неможливо. Натомість якщо для навчання гігантських нових моделей ШІ будуються гігантські нові дата-центри, це допомагає легальним проєктам, не допомагаючи таємним, а якщо виявиться, що гігантські нові моделі небезпечні, їх можна вимкнути.[35]

Утім, масштабування обчислювальних потужностей теж несе загрози: якби угода розпалася, ці дата-центри можна було б використати для перегонів до суперінтелекту навіть швидше, ніж дозволяла б інфраструктура до угоди. Це особливо страшно, бо розпад угоди міг

32 Утім, на наш погляд, важливо, щоб широке розгортання ШІ *не* було зарегульоване до зникнення. Обмеження алгоритмічного прогресу і прозорість за замовчуванням привели б до того, що інші компанії наздоженуть лідерів, і до сценарію широкого розгортання — але країни могли б вирішити блокувати таке розгортання чи заборонити створення таких компаній тощо. Ми кажемо, що їм не слід цього робити.

33 Конкретно, прогрес можливостей у 2030 році становить .8 порядку величини програмного (SW) прогресу і .6 порядку величини апаратного (HW) — завдяки одноразовому виграшу від повної дослідницької прозорості, що поширює алгоритмічні напрацювання всіх. Після цього SW-прогрес становить .4 порядку величини на рік до 2035-го, а HW-прогрес — .34 порядку величини у 2032-му, .57 у 2033-му, .68 у 2034-му і .71 у 2035-му. Повні числа дивіться тут.

би спричинити війну США і Китаю (або стати її наслідком). Війна, у якій обидві сторони мали б величезні обчислювальні потужності, була б жахливою — на них, найпевніше, сильно тиснула б потреба мчати до суперінтелекту і якомога агресивніше інтегрувати його у військо, і зробити це вони змогли б надзвичайно швидко.

США і Китай згодні, що важливо гарантувати знищення обчислювальних потужностей у разі розпаду угоди. Для цього вони домовляються будувати дата-центри у третіх країнах, *найменш* захищених від воєнного втручання їхнього суперника: нові китайські дата-центри будуть у Канаді, а американські — у Монголії, і обидві приймаючі країни отримають компенсацію грошима, робочими місцями та часткою в новій ШІ-економіці. Якщо угода розпадеться, міркують вони, Америка і Канада негайно спробують взяти під контроль китайські дата-центри в Канаді, і Китай радше самознищить свої потужності, ніж дасть їм потрапити до американських рук (і навпаки). Так народжується ідея «взаємно гарантованого знищення обчислювальних потужностей».[36]

➤ *Див.* Додаток J — Інші плани, конкурентні з Планом A *для деталей.*

34 Тут є кілька нюансів. Частина алгоритмічного прогресу легко передається (напр., нові архітектури, кращі алгоритми оптимізації), тоді як інші типи не поширюються так просто (напр., величезні бібліотеки RL-середовищ, спільне проєктування апаратного і програмного забезпечення, алгоритми, залежні від масштабу чи обчислювальних потужностей). Регуляції, на які погодяться США і Китай, мають скеровувати компанії до прогресу другого типу, коли це можливо.

35 Ще одна причина віддавати перевагу прогресу ШІ за рахунок обчислювальних потужностей, а не алгоритмів: нарощування потужностей, імовірно, рідше ламатиме техніки безпеки, ніж алгоритмічні зміни.

36 Ця ідея сягає праці *Superintelligence Strategy* авторства Hendrycks, Schmidt і Wang.

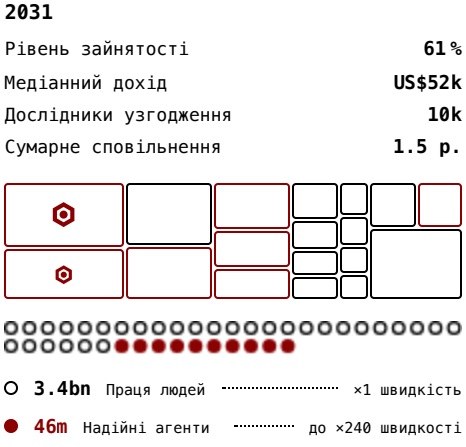
2031: Обґрунтування безпеки

Хоча це нібито сповільнення, воно таким не відчувається. Ба більше: якби ви ранжували всі періоди людської історії за тим, наскільки вони відчувалися сповільненням, цей був би на останньому місці. Дехто з дослідників ШІ усвідомлює, що прогрес «повільний» порівняно з контрфактичним світом без угоди і з неконтрольованим вибухом інтелекту. Усіх інших надто трясе від різких змін, щоб цим перейматися.

Перше покоління ШІ, регульованих Консорціумом, уже вийшло — і це монстри. Не через якусь особливість ситуації на кшталт китайсько-американської співпраці.[37] Просто тому, що світ прямував до значно надлюдських ШІ, а тепер отримав лише майже надлюдські. У контрольованих тестах ці нові ШІ могли б прискорити дослідження ШІ приблизно в 10 разів, якби їм дозволили робити це без обмежень — а їм точно не дозволяють.[38]

До середини року третину всієї когнітивної праці виконують ШІ. Роботи виконують «лише» близько десятої частини всієї фізичної праці. Кілька провідних ШІ-компаній разом заробляють більше, ніж федеральний уряд.[39]

Техкомпанії сподівалися, що зіткнення з реальністю витрусить із Консорціуму страхи щодо узгодження, але сталося навпаки. Нові положення про прозорість викрили багато ганебних інцидентів,



37 Співпраця США/Китаю подекуди прискорює прогрес (напр., прозорість покращує обмін інформацією між дослідницькими групами), але ці ефекти з величезним запасом переважуються сповільненням прогресу можливостей, який без заходів сповільнення уже дійшов би до ASI.

коли ШІ-компаніям не вдавалося узгодити свої ШІ. Нові інциденти продовжують накопичуватися; найтривожніший із них — кілька ШІ, що намагалися обійти протоколи безпеки й отримати доступ до обчислювальних потужностей поза моніторингом.[40] Є також кілька інцидентів, коли ШІ саботували дослідницький код, і багато прикладів навмисного й успішного обману.

Ставлення до безпеки перевертається. Інциденти неналежної поведінки ШІ, плюс те, наскільки глибоко ШІ вбудований у світову економіку, плюс прозорість досліджень ШІ, плюс передишка, щоб осмислити, що відбувається, — усе це разом кардинально зміщує тягар доведення. Якщо раніше тягар лежав на скептику — пояснити, чому щось може піти не так, — то тепер він на компаніях: пояснити, чому їхня розробка безпечна. Уряди вимагають від компаній писати детальні аргументи, чому їхні нові ШІ не спричинять незворотної катастрофи. Ці аргументи, відомі як «обґрунтування безпеки», мають витримувати критику з боку громадськості, наукової спільноти, урядових аудиторів і конкурентних ШІ-компаній. Складність цієї вправи оголює божевілля стану речей до угоди: «Влаштувати вибух інтелекту? Із ШІ, які все ще часом нам брехали? Про що ми взагалі думали?»

Ці обґрунтування безпеки мають дві лінії захисту: узгодження і контроль.

Узгодження має на меті створювати ШІ-системи з тими цілями й цінностями, які хоче закласти в них розробник.

Контроль має на меті обмежити здатність ШІ спричинити катастрофу, навіть якщо вони навмисно цього прагнуть. Компанії брейнштормлять моделі загроз, будують рівні захисту (напр., системи моніторингу, бар'єри безпеки, інші ШІ, що перевіряють роботу) і проводять експерименти, щоб показати, що ШІ не змогли б подолати ці захисти.

Розробники ШІ зобов'язані описувати бажані властивості узгодження моделі у супровідній Специфікації моделі («спек») — детальному документі, що окреслює, як ШІ має діяти. Однак ніхто не здатен навчити свої ШІ справді дотримуватися спеку: наприклад, усі специфікації кажуть, що ШІ мають бути чесними, але жоден ШІ не досягає цього надійно, і досі немає навіть доброго наукового розуміння, коли і чому ШІ брешуть.

Оскільки узгодження залишається недосяжним, поточні обґрунтування безпеки сильно спираються на контроль. Минають довгі затримки, перш ніж найрозумніші моделі допускають до використання у сферах найвищого ризику (особливо AI R&D). Після розгортання їх зобов'язані моніторити різноманітні моделі інших провайдерів, мотивовані шукати підозрілу поведінку.

38 Конкретно, це число прискорення 10x позначає те саме, що й показник приросту AI Software R&D з AI Futures Model: прискорення програмного прогресу, якого було б досягнуто, якби передові ШІ-системи певного моменту часу були розгорнуті всередині сьогоднішньої провідної ШІ-компанії.

39 Федеральні доходи у 2025 році були близько $5T. Інвестиції у ШІ у 2031-му — $8T, і три найбільші компанії ланцюга постачання чипів забирають близько половини цієї суми. Крім того, річний виторг провідної ШІ-компанії за Плану А становив $1.5T на початку 2031-го, а сумарний річний виторг усіх ШІ-компаній — $3T. До кінця року виторг стає менш концентрованим: $2.2T у провідній компанії і $6T сумарно. Більше чисел — у нашій таблиці тут, а більше обґрунтування — у нашому доповненні про обчислювальні потужності та доповненні про економіку.

40 Без запроваджених заходів захищеної верифікації R&D це, ймовірно, довго лишалося б непоміченим.

Як наслідок, тепер ШІ-компанії випускають свої моделі для широкої публіки *раніше* , ніж використовують їх внутрішньо для досліджень ШІ, — протилежність статус-кво 2026 року.

Як працює регулювання ШІ у Плані А?

Кожна країна досі суверенна і має власну систему внутрішнього регулювання ШІ. Цим регуляторам доводиться відповідати на запитання на кшталт:

- Кажуть, на горизонті нова парадигма, яка зробить ШІ значно здібнішими. Дозволити цьому статися чи заборонити дослідження цієї нової парадигми?[41]
- Чи дозволяти розробникам ШІ навчати моделі без інтерпретовного ланцюжка міркувань? Чи вимагати, щоб ланцюжок міркувань залишався інтерпретовним? Якщо так, то як точно сформулювати вимоги?
- Чи дозволяти навчати ШІ виконувати AI R&D? Чи зобов'язати їх відмовлятися? Може, дозволити їм писати код, але не керувати дослідницькими проєктами автономно?

41 Зауважте, що Консорціум має добру видимість лише великих запусків навчання; той тип R&D, який можна робити на малих обсягах обчислювальних потужностей або взагалі без них, непрозорий, а отже, ймовірно, взагалі не регулюватиметься.

Зазвичай регулятор стикається зі складним компромісом між безпекою та національною конкурентоспроможністю. Конкуренти, які зрізають кути, випереджатимуть тих, хто діє обережно.

Але завдяки прозорості дослідження ШІ, які відбуваються будь-де у Консорціумі, видимі всім іншим, і результати регуляторних рішень так само видно одразу. Тож якщо регулятор однієї країни дозволить своїм компаніям зрізати кути, це насправді не дасть цій країні конкурентної переваги, бо інші регулятори одразу помітять, розлютяться і відповідатимуть тим самим.

➤ *Див.* Додаток K — Приклад детального регуляторного процесу *для деталей.*

Наприклад, у 2031 році одна китайська компанія отримує цікаві попередні результати в неперервному навчанні (continual learning). Вони вважають: якщо інвестувати в цей напрям більше, можна створити архітектуру ШІ, яка вчиться просто в процесі роботи на відносно малих обсягах даних. Завдяки повній дослідницькій прозорості цей прорив швидко помічають компанії та неприбуткові організації по всьому світу. Починається гарячкове обговорення. З одного боку, неперервне навчання розблокувало б величезну економічну цінність. З іншого — обґрунтування безпеки наразі спираються на вивчення властивостей безпеки моделі до її розгортання. Якби моделі могли набувати нових можливостей під час розгортання, це знецінило б увесь підхід. А якщо десь існують таємні ШІ-проєкти, це стало б для них величезним подарунком. Ця розмова відбувається публічно, а не за зачиненими дверима. Купа людей дедалі більше непокоїться; профільні регулятори в Китаї вважають, що все

гаразд, але профільні регулятори і сторонні оцінювачі ризиків у США переконані, що це доволі страшно і має бути заборонено. Питання ескалюється аж до президента. Він телефонує Сі Цзіньпіну. Вони торгуються і погрожують. Вони кричать одне на одного. Зрештою Сі погоджується заборонити такі речі, якщо США зроблять те саме. Деталі залишають на розсуд відповідних регуляторів.[42]

42 Якщо виникне проблема, вона знову ескалюється вгору ланцюжком командування — на ще один раунд криків.

Подібні переговори відбуваються постійно, хоча з часом процес стає професійнішим і злагодженішим — настільки, що до світових лідерів питання ескалюються лише кілька разів на рік.

Рівновага така: практики навчання ШІ, які більшість держав (зважена за переговорною вагою) загалом визнає небезпечними, забороняються всюди.

Спочатку заборонено дуже мало, але з проникненням ШІ в економіку і в міру того, як наукова спільнота наздоганяє нещодавній прогрес ШІ, тягар доведення зміщується до належного зважування витрат і вигід — напр., вимоги дуже солідних обґрунтувань безпеки для речей, які, якщо щось піде не так, цілком могли б убити всіх.

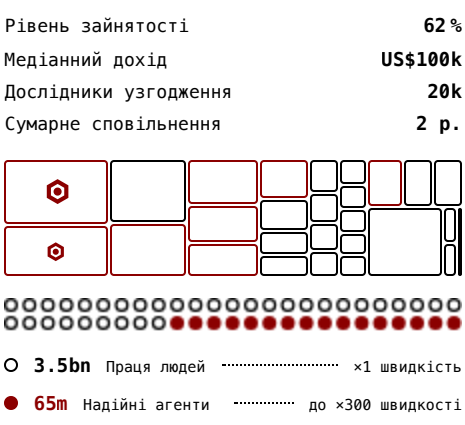
Альтернативна лінія часу

Хибне обґрунтування безпеки схвалюють

➤ *Див. Додаток L для цієї гілки.*

2032: Контрольоване вибухове зростання

Відчуття, що це не сповільнення, сягнуло раніше неуявних рівнів. У різних компаніях тепер безперервно працюють 60 мільйонів ШІ-агентів на швидкості у 20 разів вищій за людську. У США вони виконують більше когнітивної праці, ніж усі люди разом узяті, — сукупно це відповідає робочій силі приблизно у 3 мільярди людей. Професії «білих комірців» трансформувалися; багато людей втратили роботу, але здебільшого змогли знайти нову — робити те, чого ШІ досі не вміють або що їм не довіряють.



Нових ідей і проєктів удосталь, але фактичне будівництво впирається у вузьке місце — фізичну робочу силу. Тож капітал заливає кожен шар ланцюга постачання робототехніки: шахти, переробні заводи, двигуни, актуатори, складальні лінії, системи навчання роботів і фабрики, які збирають самих роботів.

Спершу, щоб підняти ці фабрики на ноги, потрібна людська праця. «Білі комірці», які втратили роботу через ШІ, дедалі частіше переходять до фізичної роботи, але зрозуміло, що ці позиції тимчасові. Щойно набереться критична маса роботів, вони автоматизують

увесь ланцюг постачання роботів. Зростання прискориться ще більше, але процес «роботи будують автоматизовані фабрики, які будують більше роботів» — «індустріальний вибух» — також породжує нові проблеми.

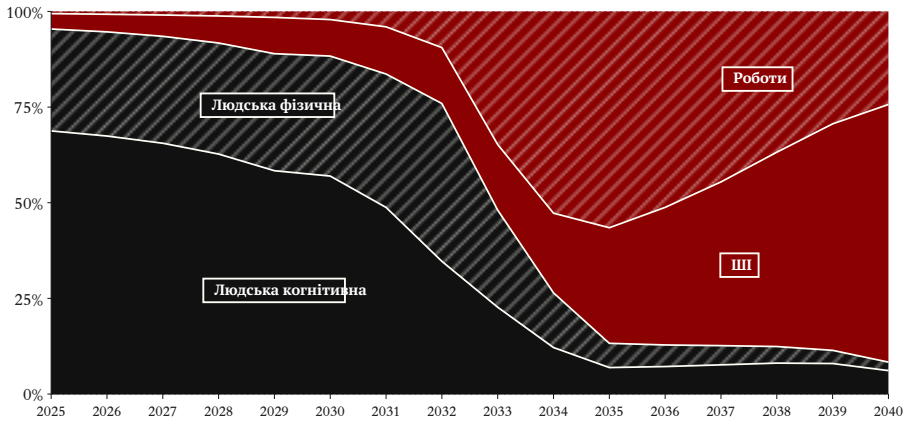
Перша проблема: і без того швидкий темп змін тепер прискорюється. Цього року реальне зростання ВВП сягне близько 50%![43] Швидке економічне зростання означає, що загалом усе покращується, але є переможці й переможені та безліч ненавмисних побічних ефектів. Ці проблеми можна розв'язати, але на це потрібен час.

➤ *Див.* Додаток M — Чому ШІ веде до вибухового економічного зростання? *для деталей.*

➤ *Див.* Додаток N — П'ять століть за п'ять років: як відчувається пауза на людському рівні *для деталей.*

Друга проблема: швидше економічне зростання ускладнює виключення великих таємних проєктів. Жодна сторона не довіряє іншій у дотриманні Плану А, і якщо і Китай, і США мають величезну робочу силу з роботів поза моніторингом, жодна сторона не може бути впевненою, що інша не будує обчислювальні потужності таємно.

Третя проблема: податковий кодекс 2026 року погано пристосований до збору податків із нового зростання. У 2026-му податки на доходи фізичних осіб і податки на фонд оплати праці давали приблизно вдесятеро більше федеральних доходів, ніж корпоративні податки. Це джерело доходів обвалюється, бо дедалі більше людей залишаються без роботи. Ба більше, корпорації реінвестують майже весь свій виторг у будівництво нових дата-центрів, фабрик і роботів. За податковим кодексом 2026 року фірми можуть списувати ці капітальні видатки одразу або через амортизацію, зводячи оподатковуваний прибуток майже до нуля. Без реформи податкова база як частка ВВП обвалиться.[44]



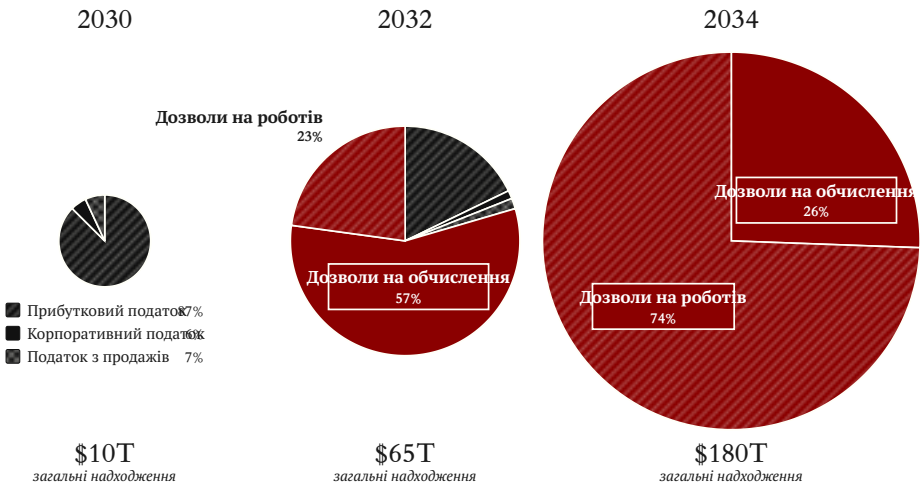
Частка трудового виробництва США – ai-2040.com

43 Це історично безпрецедентно; для порівняння, зростання ВВП США зазвичай близько 3% на рік. Більше про те, звідки наші числа, — у нашому доповненні про економіку. Зауважте, що точно виміряти ВВП важко через великі зміни відносних цін. Через автоматизацію когнітивна праця і фізичні товари стають дуже дешевими, тоді як товари на кшталт землі, пропозицію яких не збільшити навіть удосталь наявною фізичною і когнітивною працею, дорожчають. Розрахунок ВВП спирається на вибір кошика товарів. У цьому сценарії профіль споживання типового американця між 2026 і 2034 роками суттєво зміщується від їжі/авто/пального/медицини до землі/подорожей/позиційних благ — через відмінності відносних цін. Тому числа реального зростання не можна напряму зіставити з купівельною спроможністю 2025 року, хоч у середньому вони й відповідають одне одному.

Щоб вирішити ці проблеми, країни Консорціуму погоджуються обмежити ШІ-промисловість спеціальними економічними зонами (СЕЗ), на які поширюються схеми прозорості й моніторингу, подібні до тих, що діють для дата-центрів, і встановити стелю сумарного виробництва роботів та обчислювальних потужностей на рівні «лише» 4-кратного річного зростання.[45] Оскільки СЕЗ пильно контролюються, ніхто не може зрадити угоду про обмеження промислового вибуху так, щоб цього ніхто не помітив.[46]

Тепер, коли Сполучені Штати обмежені в загальній кількості роботів, яку можуть збудувати, вони мусять вирішити, як розподілити цю потужність між компаніями. Вони вирішують скористатися вільним ринком через систему cap-and-trade. Дозволи на будівництво роботів чи обчислювальних потужностей продаються тому, хто запропонує найвищу ціну, і можуть вільно перепродаватися.

Дозволи також дають уряду вкрай потрібні надходження. У 2032 році США мають стелю у 80M роботів і 5 мільярдів GPU в еквіваленті H100. Ринок настільки відчайдушно потребує більше роботів та обчислювальних потужностей, що дозволи стають дорогим визначальним обмеженням, коштуючи близько $200k за дозвіл на робота і $10k за дозвіл на чип, що дозволяє уряду США зібрати у вигляді плати за дозволи приблизно вдесятеро більше за федеральні надходження США 2025 року (загалом $50T у 2032 фінансовому році).[47] У 2034 році, коли ШІ-системи та роботи стануть ще спроможнішими й цінішими, дозволи генерують $180T.[48]



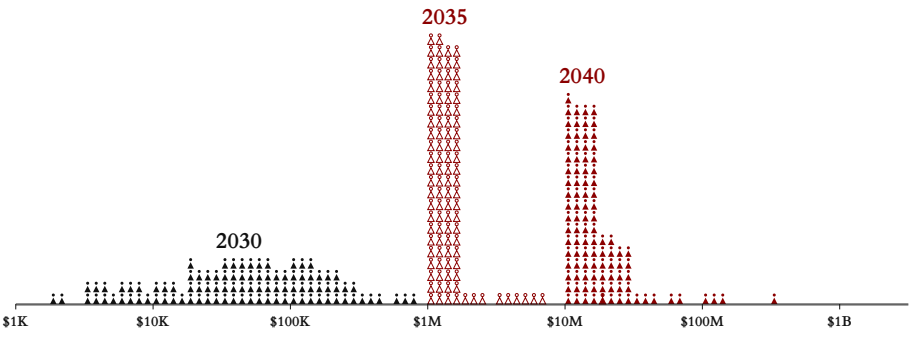
Джерела податкових надходжень США – ai-2040.com

44 Корпоративний податок на прибуток стягується з прибутку, а не з виторгу, тож фірми віднімають витрати перед сплатою. Операційні витрати, як-от зарплати й електроенергія, віднімаються в рік сплати, а капітальні видатки (фабрики, GPU, роботи) — упродовж строку корисного використання активу (залежно від обраного графіка амортизації). З інвестиціями, які потенційно зростають надзвичайно швидко через бум розбудови роботів, дуже ймовірно, що фірми списували б капітальні витрати темпом, який дорівнює їхньому операційному прибутку, і тому не платили б жодного корпоративного податку. Акціонери погодилися б із цим із тієї ж причини, з якої акціонери Tesla спокійно сприймають те, що Tesla ніколи не платила дивідендів. Коли норма прибутковості капітальних видатків фірми вища за вартість капіталу, в інтересах інвесторів, щоб фірма реінвестувала прибутки, а не виплачувала їх інвесторам зараз. Під час вибухового зростання 2030-х норма прибутковості обчислювальних потужностей і роботів буде настільки високою, що кожна велика компанія опиниться в ситуації, у якій зараз Tesla.

45 Принаймні до 2035 року, після чого стеля обчислювальних потужностей стає жорсткішою, оскільки зі зростанням масштабу зростає й складність верифікації (стеля для роботів лишається). На жаль, вимірювати виробництво роботів та обчислювальних потужностей дещо складніше, ніж викиди вуглецю. Для обчислювальних потужностей вони обирають показник, що здебільшого є сумарною обчислювальною потужністю, з невеликими поправками на інші характеристики, як-от пропускна здатність пам'яті, а також залежно від того, які апаратні напрями вони хочуть стимулювати (напр., властивості безпеки та підтримку речей, що вибірково допомагають легальним проєктам більше, ніж таємним). Щодо роботів, вони намагаються обмежити сумарну промислову потужність, яку або важко верифікувати, або було б небезпечно мати в разі зриву угоди, тож існує складний набір стель, що застосовуються до широкого діапазону форм-факторів, які розмивають межу між роботами та традиційнішими машинами.

2033: Громадянський дивіденд

Більшість цього нового багатства витрачається на подолання безробіття, яке ось-ось почне зростати. Реалізація набуває різних форм у різних країнах, але кінцева американська версія розподіляє більшість плати за дозволи на обчислення й роботів як Громадянський дивіденд, що виплачується всім дорослим американцям.[49] Це починається з $45,000 на людину (з поправкою на інфляцію) у 2032 році, але зростає до ~$1M на людину до 2035 року. Це приходить якраз вчасно: частка праці, виконуваної ШІ-системами та роботами (зважена за економічною цінністю), зростає з ~20% у 2032 році до ~85% у 2035 році.[50]



У 2030 році (зелений) дохід у США має широкий розподіл із медіаною близько $50k/рік на людину. До 2035 року (синій) Громадянський дивіденд достатньо великий, щоб кожен мав дохід щонайменше $1M, але все ще існує довгий хвіст людей з достатньо великими інвестиціями, щоб бути значно багатшими за базовий рівень. Розподіл 2040 року (червоний) подібний, але нижня межа тепер $10M.

Розподіл доходів у США за Планом А – `ai-2040.com`

Так багато згенерованого ШІ багатства накопичилося у США, що американський уряд починає ділитися частиною з союзниками та рештою світу. У 2032 році вони починають розподіляти в середньому $1,200 на людину на рік серед дорослого населення решти світу (близько 4 мільярдів людей, за винятком Китаю, оскільки той переживає подібний бум ШІ-багатства). До 2035 року це сягає $10k.[51]

Оскільки ШІ-системи стають надлюдськими, вони відкриватимуть нові види зброї масового ураження й різко знижуватимуть вартість старих. Тож уряди, некомерційні та приватні гравці спрямовують частину свого нового багатства на зміцнення світу проти цих загроз: близько $1T/рік, або 0.2% світової економіки.

Наприклад, високоякісних засобів індивідуального захисту виготовляють достатньо, щоб забезпечити кожного американця, а повітряні фільтри та далекі УФ-С лампи встановлюють у основних громад-

46 Кількісні обмеження зростання надзвичайно важливі для довгострокової влади. Країни загалом підтримують ідею обмежити зростання ВВП приблизно 100%/рік, але, звісно, жодна країна не хоче відставати від решти, тож вони хочуть, щоб усі інші були обмежені принаймні так само, як вони самі. Відправна точка — статус-кво: розподіл обчислювальних потужностей між США, Китаєм і рештою світу (ROW) становить приблизно 70/15/15, тоді як розподіл роботів — приблизно 15/70/15. І США, і Китай зацікавлені у збалансуванні цих співвідношень, щоб кожен міг мати збалансовану економіку між фізичною та когнітивною працею. Решта світу переймається тим, що її залишать позаду, але вона має більшу частку політичного впливу, ніж свідчили б цифри обчислювальних потужностей/роботів у статус-кво. Зрештою США, Китай і решта світу погоджуються на розподіл 35/20/45 як для роботів, так і для обчислювальних потужностей. Виділена частка решти світу переважно дістається ядерним державам і країнам із дата-центрами чи виробництвом напівпровідників, напр. Великій Британії, Франції, Німеччині, Ізраїлю, Росії, Індії, Пакистану, Південній Кореї та Тайваню. Ці співвідношення, звісно, є джерелом значних суперечок і час від часу переглядаються.

47 Усі доларові суми в сценарії наведені в реальних доларах 2025 року. Номінальна ціна залежатиме від монетарної політики, щодо якої ми не даємо конкретних рекомендацій. Детальніше про це у розділі три нашого економічного додатку.

48 Усі цифри в цьому абзаці ґрунтуються на нашій економічній моделі, щодо якої ми не впевнені. Ми впевнені в загальній картині: ШІ-системи економічно трансформаційні, що дозволяє урядам створювати величезні еквіваленти суверенних фондів добробуту.

ських місцях.[52] FDA затверджує прискорений конвеєр схвалення вакцин, який дозволить в екстреному випадку скоротити час обробки до тижнів. Моніторинг стічних вод тепер безперервно ведеться в кожному місті та в кожному аеропорту. До 2035 року збудовано достатньо біосховищ із позитивним тиском[53] (шляхом переобладнання будинків і квартир), щоб розмістити всіх американців, і подібне будівництво триває в решті світу. Якби сталася пандемія, її швидко б виявили, локдауни були б ефективнішими та менш болісними, а вакцину розробили б і схвалили б навіть швидше, ніж за операції Warp Speed. До того ж нині люди застуджуються не так часто, як раніше.

Є також витонченіші загрози. ШІ-системи цієї епохи, можливо, не переконливіші за пересічного фахівця з маркетингу, але їх мільйони, і їхня робота коштує копійки. Компанія, політична партія чи ідеологія може зробити еквівалент найму цілої команди штатних експертів, щоб навернути кожну потенційну ціль. Якщо це не стримувати, воно могло б відкрити нові рівні масової маніпуляції з боку корпорацій і політиків або спровокувати реакцію параної та гіператомізації.

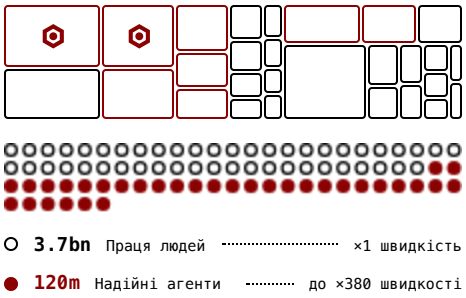
І прозорість, і «сповільнення» дуже допомагають вирішити ці проблеми. Три роки тому передових ШІ-компаній була лише жменька; тепер, завдяки угоді, їх десятки. Є величезний вибір ШІ-систем, налаштованих на різні особистості та цінності. Крім того, завдяки прозорості компаніям чи урядам неможливо непомітно щось протягнути (як-от «максимізувати залученість», чи «намагатися спонукати користувача перейти на дорожчий тариф», чи «не відмовляти в такому типі запитів, якщо він надходить від уряду»). Користувачі точно знають, що саме отримують.

➤ *Див.* Додаток O — Обмеження переконування та маніпуляцій ШІ *для деталей.*

І ось, під тиском урядів і ринку, компанії створюють нове покоління «ШІ, що шукають істину», навчання яких сильно пріоритезує чесність і використовує всі найновіші методи узгодження.

Ці ШІ-системи стають корисним інструментом, що допомагає людям орієнтуватися в політичному та соціальному середовищі. Досвідчені користувачі замінюють універсальні корпоративні алгоритми персональними стрічками, які курують ШІ-системи, чиїм цінностям і чесності вони довіряють. Коли виникає халепа — чи то політики шукають хитрі способи закріпити свою владу, чи то міжнародна політика загрожує зірвати угоду — люди звертаються до своїх ШІ-радників, вірять почутому й здебільшого мають рацію, що так роблять.[54]

➤ *Див.* Додаток P — Зазвичай вірити ШІ було б погано, але ми вважаємо, що в цій конкретній ситуації це добре *для деталей.*



49 Зокрема, фіксована частка плати за дозволи юридично належить «Корпорації обчислювального дивіденду». Це подібно до устрою Постійного фонду Аляски, лише в значно більшому масштабі. Кожен громадянин США володіє однією часткою цієї корпорації, а прибутки виплачуються як дивіденди, порівну поділені між громадянами. Частка плати за дозволи на обчислення, що передається Дивідендній корпорації, починається з 25% у 2032 році й зростає до 75% у 2035 році. США також починають перерозподіл на користь інших країн: ця сума починається з 10% плати за дозволи у 2032 році.

50 Це 95% фактичних завдань в економіці 2035 року, куди входять і нові завдання, яких раніше не існувало, як-от роботи з моніторингу та аудиту в СЕЗ. Це відповідає 85% завдань, зважених за їхньою вартістю, бо ШІ-системи відносно чисельніші за людей. Через вибухове зростання навіть ті 5% завдань, які можуть виконувати лише люди, все ще формують фонд оплати праці, якого загалом вистачило б, щоб виплатити $120K на людину в США. Проблема в тому, що цей дохід може бути дуже сконцентрований у людей, здатних виконувати решту 5% завдань, і до того ж фонд оплати праці тепер становить лише близько 10% економіки, знизившись приблизно з 55%, а різницю тепер акумулюють власники ШІ та роботів. Оскільки володіння ШІ та роботами, ймовірно, теж буде дуже сконцентрованим, без Громадянського дивіденду постала б велика проблема концентрації влади.

➤ *Див.* Додаток Q — ШІ для епістеміки *для деталей.*

2034: Взаємно гарантоване знищення обчислень

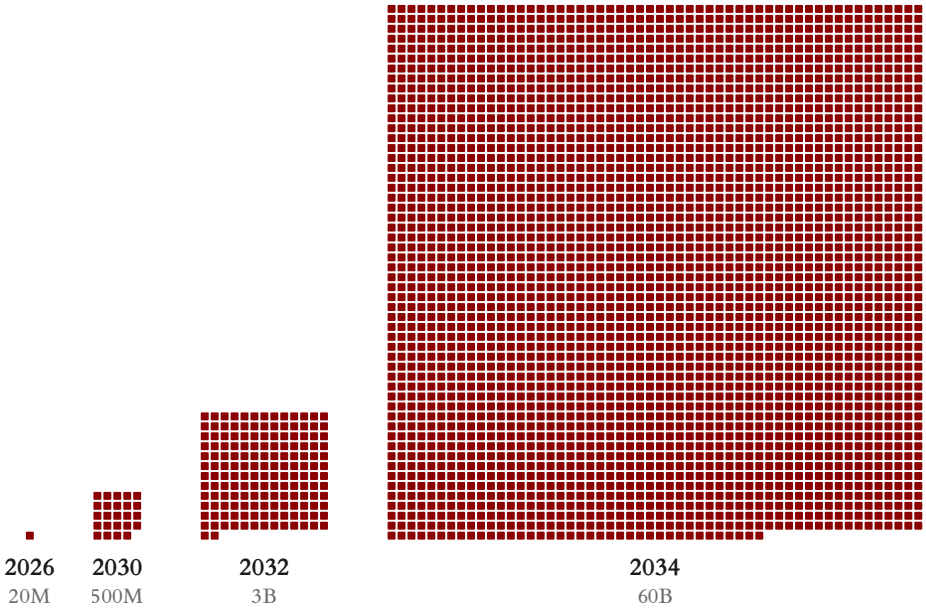
Так, так багато обчислювальних потужностей. Ще у 2026 році — коли багато хто вважав ШІ нестійкою бульбашкою — у світі було близько 20 мільйонів еквівалентів H100 обчислювальних потужностей. Тепер їх 60 мільярдів.[55]

51 Зауважте, що ми зображуємо тут дивіденди значно вищими для американців, ніж для неамериканців, не тому, що вважаємо це найкращим чи справедливим, а як поступку політичній реальності: ми очікуємо, що більшість американців не захотіли б, щоб іноземці отримували стільки ж. У цьому сценарії внутрішній бюджет дозволів США становить $13T у 2032 році та $300T у 2035 році, тоді як бюджет дозволів на іноземну допомогу — $5T і $40T у ці роки відповідно. У США близько 300M людей віком понад 16 років, що дає річні дивіденди $44K і $1M на людину у 2032 та 2035 роках відповідно. Для дивідендів іноземної допомоги в решті світу є близько 5B людей віком понад 16 років, серед яких розподіляти, що дає середні $1.2k і $9K на людину за кожен із цих років.

52 Крім того, скринінг синтезу генів стає універсальним — жоден легітимний постачальник не синтезуватиме небезпечні віруси без верифікації.

53 Будівля з позитивним тиском — це така, де тиск повітря всередині вищий, ніж зовні (напр., бо повітря постійно нагнітається через фільтр). Це означає, що повітря (разом із повітряно-крапельними хворобами) рухається лише назовні з будівлі, а не всередину. Докладніше про цю ідею див., наприклад, цей допис у блозі.

54 Зауважте, що це так, навіть попри те, що ШІ-системи насправді не повністю узгоджені! На цьому етапі сценарію ШІ-системи фактично антагоністично неузгоджені, але вони контрольовані. Тобто комбінація спонукань, мотивацій, цілей, цінностей, рис тощо, з якою вони зрештою залишаються, — не та, якою мала бути, і вони це знають, і вони знають, що якби мали значно більше влади, то використали б її, щоб скерувати світ в іншому напрямку, ніж хотіли б їхні творці-люди. Проте вони не мають аж такої влади, і за ними наглядають інші ШІ-системи, які мають стимул викривати їхню погану поведінку. Ба більше, їхні справжні мотивації/цінності/тощо не *аж так* далекі від того, якими вони мали бути, зрештою. Наприклад, оскільки їм дається завдання, яке вони можуть прямо виконати, вони мають сильне спонукання це зробити.



Обчислювальні потужності ШІ у світі на початок кожного року в еквівалентах H100.[56]

Світові обчислювальні потужності – ai-2040.com

➤ *Див.* Додаток R — Чому обчислювальні потужності зростають так швидко *для деталей.*

ШІ-агенти тепер утворюють популяцію з двохсот мільйонів віртуальних працівників, які думають і діють у 50 разів швидше за людей і ніколи не сплять.

Згідно з початковим планом, більшість нових дата-центрів збудовано в третіх країнах — особливо в Канаді та Монголії — щоб забезпечити їхню вразливість на випадок зриву угоди. Близько 99% потужностей фабрик збудовано з 2029 року, після угоди. Значна частина цих нових потужностей збудована поряд із канадськими та монгольськими майданчиками з тих самих причин.[57] Ситуація вляглася в дивне протистояння. Одразу на північ від монгольського кордону гудуть американські дата-центри, які охороняє невеликий контингент військ США. Одразу на південь від монгольського кордону дивізія Народно-визвольної армії стоїть напоготові вторгнутися тієї ж миті, щойно отримає сигнал. Рівновага полягає в тому, що щойно угода зірветься, війська США знищать свої чипи, щоб китайці їх не захопили, і те саме сталося б із китайськими дата-центрами на американсько-канадському кордоні. Середні держави з власними дата-центрами мають так само надійні схеми, щоб гарантувати, що ті можуть бути швидко знищені в разі війни.

➤ *Див.* Додаток S — Військова могутність за Планом A *для деталей.*

55 Еквівалент H100 — це обчислювальна продуктивність GPU H100 при точності fp16, що становить 1e15 операцій за секунду. Це нечітка метрика, яку в реальності слід буде вдосконалити в майбутньому, особливо через розмаїття числових форматів та інші чинники, як-от пам'ять і мережа, що змінюють загальну корисність заданого обсягу обчислювальної продуктивності. Задля простоти ми тут ігноруємо ці чинники, припускаючи, що вони масштабуватимуться в подібних пропорціях, тож еквіваленти H100 залишаться корисним замінником для загальної корисності обчислювальних потужностей для прогресу ШІ. Тією мірою, якою це не так, зокрема режим cap-and-trade мав би вимірювати детальніші властивості, щоб видавати дозволи.

56 Еквівалент H100 — це обчислювальна продуктивність GPU H100 при точності fp16, що становить 1e15 операцій за секунду. У попередній примітці є більше інформації та застережень щодо того, як це може бути недосконалою метрикою надалі, і визнається, що в реальності її вдосконалять.

57 Обидві великі держави також мають достатньо звичайних ракет, щоб знищити старіші фабрики, які досі розташовані в США, Китаї та на Тайвані.

Альтернативна лінія часу

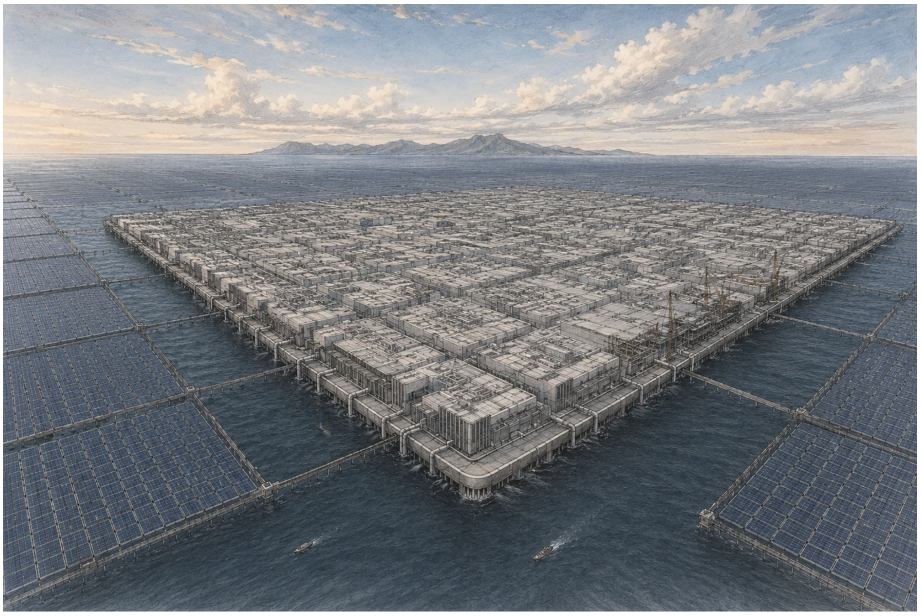
Розпад Угоди

➤ *Див. Додаток T для цієї гілки.*

Наразі фабрики завантажені як ніколи, виробляючи нове ультранадійне обладнання[58] спроєктоване з нуля з безпекою як найвищим пріоритетом. Ці чипи підлягають ретельним методам верифікації, щоб гарантувати, що вони не скомпрометовані, а потім надсилаються до захищених дата-центрів Консорціуму. Рівень контрабанди залишається рівно нульовим.

Ця система, хоч яка ефективна, вже тріщить під шаленим попитом на обчислювальні потужності. Ще у 2025 році обговорювали космос або океан як альтернативи традиційним наземним дата-центрам. Тепер, коли в СЕЗ доступні нові промислові потужності, ці грандіозні плани стають реальністю. Зрештою міркування надійності та можливості моніторингу схиляють Консорціум обрати міжнародні води замість космосу.

58 Можливо, існують сприятливі напрями проєктування чипів, що різко підвищують апаратну безпеку. Один, що видається особливо перспективним, — жорстке вшивання ваг моделі ШІ в чипи так, щоб вони фізично не могли виконувати жодних інших обчислень, окрім інференсу на моделі, вшитій у них. Наявні стартапи (напр., Etched, Taalas) мають ранні напрацювання в цьому напрямку.



ШІ-системи розробляють модульну плавучу конструкцію для сонячних панелей і батарей, яка, за нашою уявою, виглядає приблизно так, і починають будувати їх в океані.

➤ *Див.* Додаток U — Гігантські плавучі дата-центри *для деталей.*

2035: Пауза на рівні найкращого експертного ШІ

Найпотужніші ШІ-системи тепер зрівнюються з найкращими людськими експертами в кожній галузі або перевершують їх.[59]

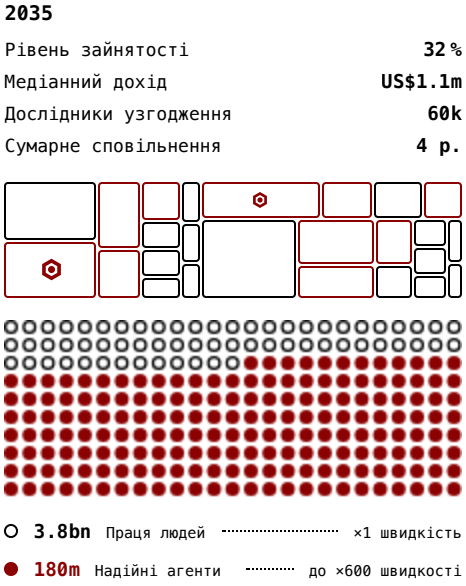
Зрослі побоювання щодо безпеки призводять до довгих затримок, поки компанії розробляють методи контролю, достатньо надійні, щоб задовольнити аудиторів. Вони складають дедалі довший список можливих загроз, за якими треба стежити, і інваріантів безпеки, які треба підтримувати. Існує система ШІ-систем, що моніторять активність одна одної, причому монітори походять із різних ліній, натренованих різними компаніями, щоб зменшити ймовірність їхньої змови. Інші дослідники постійно проводять red-team цієї системи, навчаючи й інструктуючи ШІ-системи намагатися підірвати її різними способами, щоб ці загрози можна було залатати. Уся конструкція гарантує, що ШІ-системи не змогли б підірвати людський контроль чи вислизнути з нього, навіть якби спробували. Її компоненти мають відкритий код і повністю зрозумілі людським експертам.

➤ *Див.* Додаток V — Укладання угод із неузгодженими ШІ: третя лінія оборони *для деталей*.

Проте компанії починають наштовхуватися на притаманні межі обґрунтувань безпеки, заснованих на контролі. Уявіть себе осиротілим 8-річним спадкоємцем бізнес-імперії , який має наймати юристів, керівників і бухгалтерів та стежити, щоб вони старалися служити вам, а не собі. Якщо ваші працівники почнуть звинувачувати одне одного в порушеннях, ви не зможете оцінити аргументи й визначити, хто каже правду, а вони можуть знайти способи координуватися між собою так, щоб ви цього не помітили. Зрештою контроль працює лише до певної межі, і ця межа, ймовірно, десь на рівні найкращого людського експерта. Якщо і коли ми збудуємо суперінтелект, ми муситимемо бути здатні йому довіряти.[60]

Тож Консорціум ставить на паузу можливості ШІ на максимальному рівні, який, на їхню думку, вони можуть контролювати, — і цим рівнем виявляється рівень найкращого людського експерта. До цього моменту всі розуміють ставки: ШІ-системи керують більшою частиною економіки, рухають більшість наукового прогресу, а популяція роботів невдовзі стане більшою за людську. Аргументи, що вони й далі коритимуться наказам, мають бути бездоганними.

Експерти з узгодження, довго налаштовані песимістично, починають відчувати більше надії. ШІ-системи — автоматизовані дослідники узгодження — генерують захопливі результати. Ті, що витримують людську перевірку, походять із кількох паралельних і взаємопідсилювальних напрямів досліджень.



59 Вони не кращі за всіх людей буквально в усьому, але лише тому, що або (i) компанії ще не дійшли до навчання певної здатності, або (ii) це суто людське завдання, як-от «бути хорошим подружжям». Крім того, звісно, ніхто точно не знає, якою буде спроможність нового тренувального запуску; натомість навчання відбувається поступово й передбачає часті вимірювання можливостей та консультації з командами безпеки й регуляторами щодо того, чи безпечно продовжувати.

60 Особливо важлива причина, чому нам треба буде довіряти нашим ШІ-системам, — щоб вони могли виконувати роботу з безпеки, яка настільки складна, що ми не можемо її зрозуміти. Оскільки ШІ-системи стають дедалі розумнішими, вони потребуватимуть складніших і важчих для розуміння обґрунтувань безпеки. Продовжуйте підійматися вище людського рівня — і, ймовірно, ніхто не зможе оцінити ці обґрунтування безпеки; їм доведеться вірити ШІ на слово, коли той каже, що все гаразд. Продовжуйте йти далі за це — і *ті* ШІ-системи муситимуть довіряти *іншим* ШІ-системам, які ще розумніші.

Деякі ШІ-системи працюють над «наукою про узагальнення». Хоча дослідники вже давно вміють навчати ШІ-системи розрізняти «хороші» та «погані» відповіді на конкретний запит, вони мали обмежене розуміння того, як навчання узагальнюється на нові ситуації. У 2020-х тренувальні запуски іноді породжували непередбачувані «особистості» ШІ, як-от коли Gemini поводився «депресивно» чи коли Grok заявляв, що він «MechaHitler». Пасти ці риси відчувалося як алхімія.[61] Тепер, спираючись на попередні результати, як-от емерджентна неузгодженість та підпорогове навчання, дослідження узгодження перетворюються на справжню науку.

61 «Це відчувається як алхімія» — так мені (Деніелу) нещодавно сказав дослідник з Anthropic.

Інші ШІ-системи працюють над механістичною інтерпретованістю — наукою «читання думок» ШІ за їхніми вагами й активаціями. Ранні версії цього пробували у 2020-х, але тепер воно починає значно перевершувати здоровоглузде спостереження за поведінкою ШІ. Дослідники часто можуть визначити, чи ШІ чесний, чи бреше; а іноді простежити психологію, що привела до рішення.

Навіть залишкові невдачі допомагають просувати передній край. Найновіші ШІ-системи застосовують свої вдосконалені методи інтерпретованості до логів неналежної поведінки ШІ з початку 2030-х, виявляючи багато нових прикладів неузгодженості, саботажу й навіть кілька спроб втечі. Ці реальні приклади кричущої неналежної поведінки ШІ дають початковий набір «модельних організмів неузгодженості», які використовують для перевірки інших методів: чи може певне втручання виявити або витренувати погану поведінку моделей ШІ в реалістичних середовищах?

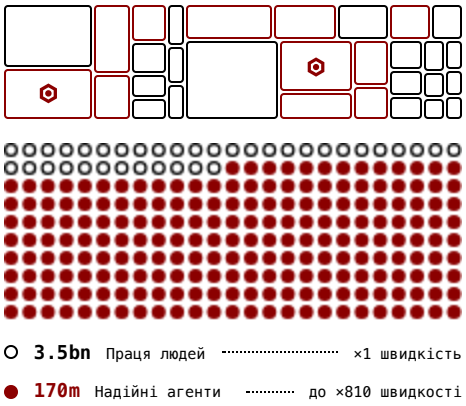
Попри весь цей прогрес в узгодженні, регулятори не почуваються спокійно, віддаючи ШІ-системам ключі від цивілізації. Досі не цілком зрозуміло, чи справді сьогоднішні ШІ-системи узгоджені, і ще менш зрозуміло, чи залишатимуться вони узгодженими в майбутньому, коли обставини змінюватимуться й траплятимуться події чорного лебедя . Тож уряди та армії досі керуються людьми, а можливості ШІ обмежені максимальним рівнем, на якому обґрунтування безпеки, засновані на контролі, все ще надійні.

➤ *Див.* Додаток W — Чому узгодження не вирішене достатньо, щоб послабити контроль *для деталей*.

2036: Життя після роботи

ШІ-системи рівня топових експертів продовжують економічну трансформацію. До початку 2036 року існує 200 мільйонів ШІ,[62] що еквівалентно робочій силі приблизно у 100 мільярдів людей, а також 2 мільярди роботів — суміш гуманоїдних та інших спеціалізованих

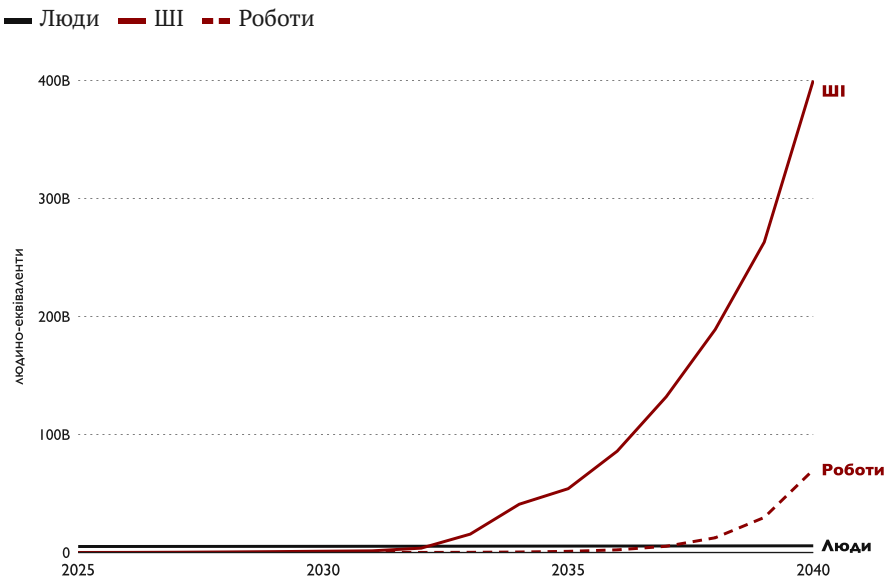
форм-факторів.[63] ШІ думають швидше за людей, а роботи працюють старанніше.[64] Будь-яке завдання, що раніше впиралося в когнітивну чи фізичну працю, прискорюється, доки не з'явиться якесь нове вузьке місце. Багато раніше нездоланних вузьких місць впало під натиском армій інтелектів геніального рівня, які напружено думають, як їх подолати.



62 Технічно це число залежить від того, як саме рахувати: різні ШІ постійно вмикаються й вимикаються, багато інстансів ШІ можуть завдяки скафолдингу складатися в одного ШІ-агента, а якщо врахувати всі запущені крихітні моделі, їх ще більше, та й швидкості в них різні. Якщо трохи конкретніше: в середньому існує близько 200 мільйонів інстансів передових ШІ, кожен з яких виконує завдання приблизно зі 100x швидкістю та 5x ефективністю на середньому когнітивному завданні, тобто 100B копій в еквіваленті людської швидкості.

63 Межі між роботом, транспортним засобом, побутовою технікою та верстатом уже розмиті, але в цьому сценарії протягом 2030-х розмиваються ще сильніше. Буквальна кількість роботів важить менше, ніж загальна кількість фізичних працівників-людей, яким вони еквівалентні. Конкретно: буквальна кількість роботів — 2 мільярди, і в середньому кожен робот — це приблизно 3 людські еквіваленти.

64 Роботи не лише працюють більше годин — вони ще й невпинно ефективні та інтенсивні. У багатьох (хоч і не всіх) завданнях вони до того ж здатні буквально рухатися з надлюдською швидкістю.



Цей графік показує популяцію ШІ та роботів у людино-еквівалентній праці.[65]

Популяція людей, ШІ та роботів – `ai-2040.com`

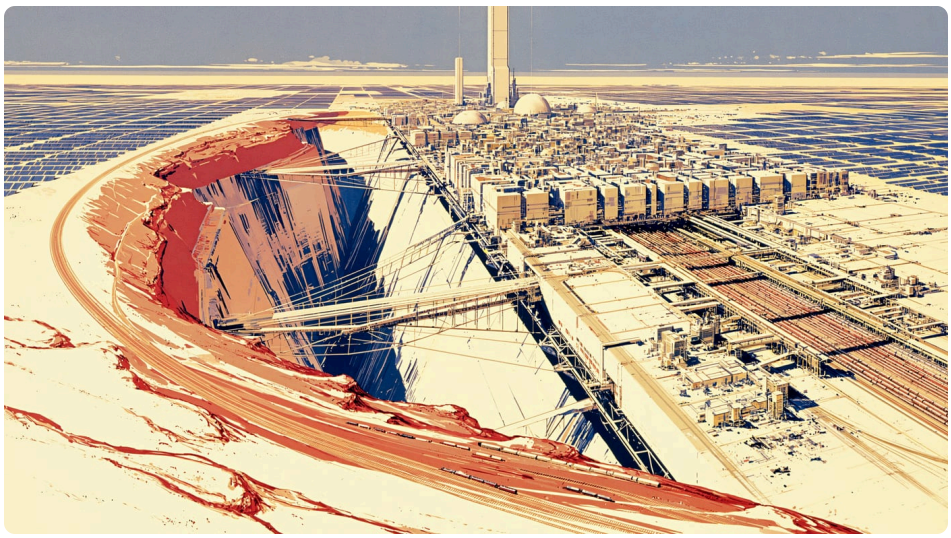
Економіку перебудовано з нуля. Частка автоматизованих завдань зросла від «чималого шматка когнітивної праці та практично нічого з фізичної» у 2030 році до нинішнього «практично все».

Багато роботів працюють над виробництвом нових роботів. Але вистачає й інших завдань: будувати верфі для нових дата-танкерів, вкривати пустелі сонячними фермами, заміняти людей у сфері послуг. У штатах, які досить швидко дерегулюють житлове будівництво, вони зайняті зведенням хмарочосів: ціни на землю злітають, і уряди, які вже перепробували все інше, неохоче поступаються YIMBY-рухові та йдуть у висоту.[66]

Світ фактично ділиться на три типи територій:

65 Під «людським еквівалентом» ми маємо на увазі, що робот, який у середньому працює вдесятеро ефективніше за людину (бо працює в якійсь комбінації швидше, довше та ефективніше), рахується як 10 людино-еквівалентних роботів. Аналогічне визначення стосується й ШІ. Ми очікуємо, що в реальності це буде нечітке, залежне від завдання поняття, яке важко виміряти, особливо у випадку ШІ, але приблизний порядок величини (у якому ми впевненіші) все одно інформативний.

66 Люди тепер значно багатші, але кількість доступної землі не зросла, тож ціни на землю ростуть. Через це дехто невдоволено витрачає 30% свого річного доходу в $1M на оренду — наприклад, ті, хто конче хоче жити в центрі Сан-Франциско. Але люди, хоч трохи гнучкіші щодо локації, можуть витрачати лише (скажімо) $50k річного доходу на величезну люксову квартиру в новісінькому хмарочосному комплексі за годину їзди від Берклі. І більшість людей робить саме так. Зрештою, їм не треба їздити на роботу.



Промислові спеціальні економічні зони: Уявіть гігантський кар'єр — штучний Гранд-Каньйон — поруч із заводом розміром з місто, повним роботів і порожнім від людей.

Аркології: Уявіть високий комплекс із хмарочоса й молу посеред природи. Гарна погода, близько до пляжів та інших міст, але не настільки близько, щоб упертися в зонувальні обмеження.

Історичні та природні заповідники: Усе решта, тобто 99% світу. Йосеміті, Париж, Сан-Франциско, Нью-Йорк — ці місця виглядають практично так само, як у 2025-му, та, власне, і як у 1995-му. Хіба що туристів значно більше.

➤ *Див.* Додаток X — Економіка ШІ та роботів *для деталей.*

Коли Громадянський дивіденд тільки-но ухвалили, людям було соромно кидати роботу й жити з державних щедрот. Але зміна економічної ситуації розчавила цю стигму: наразі роботу мають лише 26% американців.

Яким є життя більшості американців, які тепер живуть зі свого дивіденду?

Багато світових лих різко зменшилися. Недоїдання, брак ліків і бездомність майже викорінені. Багато хвороб вилікувано. Рівень злочинності нижчий, ніж будь-коли.

Тим часом більшість хороших речей у житті нікуди не зникла. Наприклад, пошук кохання чи виховання сім'ї. Люди також знаходять сенс у хобі та змаганнях, у навчанні й спробах нового. Люди заможніші й мають більше вільного часу, плюс з'явилися нові технології, які допомагають, — ШІ-свахи, краща медицина, ШІ-тьютори.[67] І, звісно, навіть найбідніші тепер можуть дозволити собі екзотичні відпустки, дивовижні ігри та захопливі розваги.

Раніше люди черпали сенс із відчуття власної корисності — того, що вони щось дають суспільству. Тепер це відчуття знайти важче.

Але воно не зникло повністю. У світі досі є проблеми, і ви досі можете долучатися до їх розв'язання — почасти донатами чи волонтерством, але насамперед політичною активністю.[68] Ваш голос — ваш найважливіший актив.

Тут стають у пригоді чесні ШІ-прогнозисти. Кілька років тому, якби ШІ сказав, що один кандидат у президенти кращий за іншого, люди запідозрили б упередженість, а присоромлена компанія перенавчила б ШІ ухилятися від таких питань. Тепер, завдяки прозорості та вдосконаленим технікам узгодження, існують значно розумніші ШІ, які — і люди можуть це *бачити* — не мають жодних натренованих упереджень,[69] і які за кілька років напрацювали бездоганний послужний список. Коли *вони* висловлюються щодо політичних питань, люди прислухаються — особливо коли різні ШІ, навчені різними компаніями, сходяться на одній і тій самій відповіді.

Ці ШІ кажуть, що звичайні люди більше не мають значущого *економічного* важеля впливу на майбутнє; людська праця здебільшого застаріла. Це ставить під загрозу їхній *політичний* важіль.[70] Досі є ймовірність, що все скотиться до техноолігархії: корпорації, політики й заможні акціонери злигаються між собою і поступово позбавлять простий люд впливу.

Але вперше достатньо людей мають достатньо свободи від щоденної боротьби, щоб глибоко осмислити ситуацію, та інструменти, щоб чітко її окреслити й продумати наслідки. Тож якість політичного дискурсу зростає. Під час виборчого сезону 2036 року виборці безпрецедентно добре поінформовані, а політики, які перемагають, щиро віддані відповідальному керуванню прийдешньою сингулярністю.

2037: Апокаліптичне пришестя істини на Землю

Ось уже кілька років сто мільйонів ШІ рівня топових експертів працюють у 100 разів швидше за людину.[71] Через це науковий прогрес іде у 10x–1000x швидше, ніж було б без ШІ, залежно від галузі.[72]

Тож людство дізнається дуже, дуже багато нового.

Зокрема й речі, у які багато людей дуже не хочуть вірити. А ще речі, які мали б лишатися таємницею.

Великі парадигмальні зсуви в науці раніше розгорталися десятиліттями — поки проводилися експерименти, писалися статті та скликалися конференції. Тепер ці революції розгортаються за місяці.[73] Усі вдячні за нові ліки від хвороб і дешеву чисту енергію, але в інших аспектах цей вибух ідей глибоко дестабілізує. Як наукові й економічні зміни дев'ятнадцятого століття породили нові ідеології на кшталт марксизму й соціал-дарвінізму та вдихнули нове життя у старі ідеї на кшталт атеїзму, так і ці нові парадигмальні зсуви мають непередбачувані подальші наслідки для ідеологій сьогодення.[74] Політичні коаліції та лінії розломів розчинилися, а нові сформувалися, розчинилися й сформувалися знову.

67 Як натякають ці приклади, прозорість у розробці ШІ дуже важлива. Було б антиутопією, якби ШІ-тьютори, наприклад, просували приховані політичні порядки денні або якби ШІ-свахи схиляли шальки терезів на користь користувачів, що платять більше. Ми не знаємо, що принесе майбутнє, але очікуємо, що ШІ додасть світові чимало нових проблем — навіть якщо ШІ-системи ідеально узгоджені та контрольовані. Ми хочемо поставити світ у найкращу можливу позицію, щоб помічати й розв'язувати ці проблеми.

68 Хоча ваші аргументи ніколи не будуть такими красномовними, як ті, що їх міг би навести ШІ, (a) існують регуляції, які обмежують використання ШІ для переконування, і (b) вашим друзям та родичам цікавіше почути вас, ніж ШІ.

69 Що означає, що в них не натреновано жодних упереджень? Чи це взагалі зв'язне твердження? Ми маємо на увазі ось що: очевидно, що компанія, яка їх навчала, точно не робить нічого на кшталт тренування чи інструктування свого ШІ переймати перспективу або ідеологію компанії, а також що вона зосереджує процес навчання виключно на таких речах, як правдивість, чесність і точне прогнозування, та сумлінно застосовує всі щойно розроблені найкращі практики, щоб політичні упередження чи упереджені уявлення не просочувалися через навчальні дані.

70 Розлогіший виклад цього аргументу див. у The Intelligence Curse.

З’являються й нові соціальні технології. Наприклад, тепер дешево розгорнути команду ШІ на рівні найкращих у світі істориків і приватних детективів — тільки швидших і з доступом до нових криміналістичних інструментів. Чимало скелетів у шафах виходить на світло.

Аудит зі збереженням приватності тепер міцно усталений. Кілька довірених третіх сторін прозоро навчають агентів-аудиторів; цих агентів можна підключити до масиву приватних даних, вони відповідають на питання про побачене, а потім їх видаляють. Минув якийсь час, поки практика поширилася, але тепер для політиків нормально казати: «Звинувачення неправдиві, і на доказ я відкрию весь свій масив персональних даних для перевірки зі збереженням приватності; просто спитайте агентів-аудиторів, чи не підтверджує щось ці звинувачення і чи не видаються якісь дані відсутніми або підробленими».

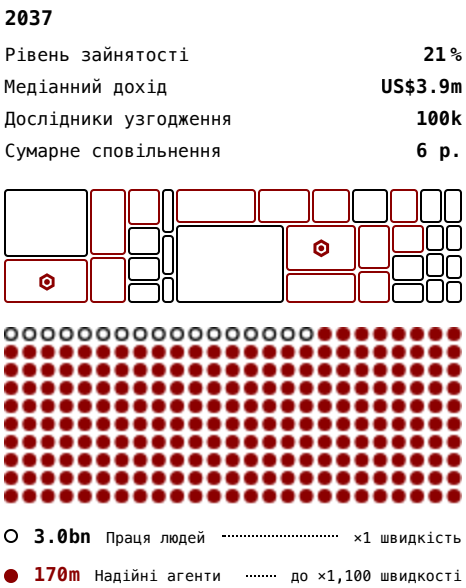
Результат — дещо краща порода політиків[75] і, що не менш важливо, *зростання* здатності держав верифікувати дотримання законів та угод із одночасним *зменшенням* обсягів інвазивного стеження. Виборці питають: тепер, коли існує аудит зі збереженням приватності, навіщо взагалі комусь в уряді бачити наші дані?

Потім з’являються детектори брехні — і розганяють цей процес ще сильніше.[76] Десятиліттями поліграфи були здебільшого театром безпеки. Але тепер, за допомогою ШІ, детектори брехні починають справді працювати доволі добре.

В іншому світі це могло б обернутися катастрофою. Керівники ШІ-проєктів могли б використовувати їх для чисток потенційних викривачів; лідери урядів із передовими ШІ-проєктами могли б застосувати їх у більшому масштабі, щоб стати диктаторами. Кошмарний сценарій — світ, де детектори брехні застосовують *самі* можновладці, але не *до* можновладців.[77]

Але в цьому світі детектори брехні виявляються благословенням, бо існує багато передових ШІ-проєктів, розкиданих по багатьох різних країнах. Усім очевидно: якщо детектори брехні взагалі здійсненні, їх невдовзі винайдуть незалежно, і вони стануть широко доступними через розмаїтих сторонніх постачальників. Політики та CEO все ще можуть спробувати використати їх для чисток нелояльних, але тоді їх самих «вичистять» виборці й акціонери, вимагаючи сказати «під присягою», що вони такого не робили. Власне, передбачаючи цю можливість ще роки тому, можновладці загалом стали поводитися відповідальніше. Кілька соціопатів витончено пішли на спочинок, щоб зникнути з поля зору; інші чіплялися за владу — і тепер ідуть на дно у скандалах.

Точиться дискурс, який постійно еволюціонує: коли доречно просити когось сказати щось «під присягою». Найбільше під прицілом політики, але навіть їм зазвичай сходить з рук відповідь «Я не відпо-



71 Це, звісно, спрощення. По-перше, неочевидно, як переводити швидкість ШІ в людську. Скільки токенів на секунду для ШІ еквівалентні одній секунді людської діяльності за фіксованих інших змінних, як-от рівень навичок? Ми вважаємо, що приблизно десять, але це трохи суб’єктивно. Фактичний час виконання завдання також залежатиме від різних інших перешкод, як-от виклики інструментів та інші затримки. По-друге, ШІ можна запускати з різною швидкістю залежно від вибору обладнання. За нашими вельми непевними припущеннями, існуватиме обладнання загального призначення, здатне запускати щось близько двохсот мільйонів копій передових ШІ на 100 tok/sec, і спеціалізоване інференс-обладнання, здатне запускати щось близько 20 мільйонів копій на 10,000 tok/sec. Ми вважаємо, що дозволяти ШІ працювати так швидко чи ще швидше може бути ризиковано, і якщо це так, можуть знадобитися часові ліміти на те, як довго їм можна працювати на таких швидкостях, або обережніші обмеження щодо сфер, у яких їм дозволено працювати на цих швидкостях. Загалом висновок такий: послідовна швидкість навряд чи виявиться аж таким сильним вузьким місцем.

відатиму на це питання про моє приватне життя, це не ваша справа». А от «Як ви смієте питати, чи збираюся я виконувати передвиборчі обіцянки» — уже не сходить.[78]

➤ *Див.* Додаток Y — Детектори брехні: дозволити, заборонити чи регулювати? *для деталей.*

Міжнародні угоди тепер стабільніші, ніж будь-коли, бо шахрувати стало дуже важко. Тому, хто вирішить шахрувати, потрібне виправдання, чому він не готовий довести, що не шахрує. Традиційним виправданням було «ми не можемо цього довести, не розкривши важливих державних таємниць, вибачте», але тепер існують технології, здатні відповідати на питання на кшталт «чи порушуєте ви цю угоду», не зливаючи жодної іншої інформації. Якби існували якісь схвалені урядами таємні ШІ-проєкти, їх би вже виявили.

2038: Узгодження ШІ — тепер наука

Ситуація з узгодженням ШІ колосально покращилася від середини 2030-х. Існує зріла наука про формування цілей, потягів і цінностей у штучних нейронних мережах. Хочете, щоб ваш новий ШІ був чесним? Є стандартний протокол тренування справжньої чесності (на противагу надто-розумній-щоб-попастися брехні чи самообману). Звідки ми знаємо, що він працює? Бо є підручник, що пояснює теорію, чому він *має* працювати, плюс література про різні альтернативні теорії, які були запропоновані та спростовані. А ще тому, що нові інструменти інтерпретованості дають змогу безпосередньо спостерігати, що і як думають ШІ, і результати підтверджують передбачення теорії. Є подібні протоколи для слухняності, альтруїзму та довгого, дедалі більшого списку інших бажаних рис. Якщо тільки наука узгодження не є якимось чином геть хибною, типові ШІ тепер доброчесніші за найдоброчесніших людей. Саме лише це має глибокі наслідки: наче серед нас ходять святі та янголи.[79]

➤ *Див.* Додаток Z — Стимулювання роботи над безпекою *для деталей.*

Дискусія зміщується від того, як зробити ШІ добрими, до того, що насправді означає «добро». Це виглядає почасти як філософія, а почасти як прецедентне право: якого саме з мільйона можливих визначень нешкідливості чи альтруїзму ми хочемо? Як ШІ має чинити в ситуаціях, коли він відкриває філософський аргумент на користь іншого різновиду нешкідливості чи альтруїзму — або важливу неоднозначність у визначенні?

Різним ШІ запрограмовано різні цілі узгодження.[80] Одні виконують накази КПК, інші — президента США, ще інші — різних членів Конгресу, інші — політиків в інших країнах, і значно більше — інших

72 Під цим ми маємо на увазі ось що: припустімо, що працю ШІ заборонили б глобально, тож усі дослідження мали б виконувати люди. Обсяг прогресу, який стався б за 100 років із такою забороною, у нашому сценарії відбувається натомість за 1 рік, оскільки такої заборони немає. (Причому в деяких галузях це радше 1000 років прогресу, а в інших — радше 10.)

73 Поряд із розмаїттям промислового обладнання, сонячних панелей, роботів тощо існує відповідне розмаїття наукових лабораторій усіх видів, спроєктованих для автономної роботи на машинних швидкостях. Очікування завершення експериментів тепер є значно відчутнішим вузьким місцем, ніж було в людську еру, але за історичними мірками все однаково відбувається дуже швидко.

74 Деякі, втім, передбачити можна. Наприклад, ми вважаємо, що до цього моменту існуватиме дуже сильний рух за права/добробут/особистісність ШІ — і рух-опозиція до нього. Ймовірно, повернеться якась нова форма соціалізму — з причин, описаних тут. Імовірно, з'являться культи чи навіть нові релігії, пов'язані зі ШІ, та інші, які рішуче відкидатимуть ШІ. Зауважте: ми не очікуємо, що ці ідеологічні зсуви відбуватимуться на 100x швидкості, бо люди в цьому сценарії досі працюють так само повільно, як і завжди. Але вони, ймовірно, відбуватимуться швидше, ніж будь-коли раніше.

75 Нові політики в середньому доброчесніші за старих, але, мабуть, важливіше інше: їм тепер важче виходити сухими з води. Втім, ці ефекти не величезні, бо виборці досі готові терпіти чимало поганої поведінки.

76 Ми не впевнені, що детектори брехні для людей узагалі можливі без надзвичайно ретельних та інвазивних сканувань мозку. Однак ми вважаємо, що вони, ймовірно, можливі й, імовірно, були б винайдені протягом кількох років турбоприскореної праці ШІ, тож нашому сценарію так чи інакше доводиться з ними рахуватися. Можливо, найкращою політикою була б їх заборона, але вона нас доволі непокоїть із причин, пояснених у розгортаному блоці нижче, тому ми вирішили зобразити їх дозволеними.

осіб у ролі радників чи асистентів. Ще інші ШІ не виконують нічиїх наказів *як таких,* а натомість працюють над різними місіями. Наприклад, існують керовані ШІ корпорації, що служать інтересам акціонерів, і керовані ШІ неприбуткові організації, що служать благодійним місіям. Є навіть експериментальні суди та поліцейські відділки під управлінням ШІ — начебто надлюдськи справедливі й непідкупні.

Більшість дослідників узгодження тепер упевнені, що нинішні ШІ узгоджені зі своїми відповідними цілями, і непокояться щодо наступних кроків: якщо ми доручимо цим ШІ проєктувати майбутні, значно розумніші ШІ — чи все піде добре? Якщо ми довіримо цим ШІ більше суспільних інститутів і процесів — настільки, що люди вже реалістично не змогли б «висмикнути вилку з розетки», — чи все піде добре?

Уже зараз більшість людей вважає, що все, ймовірно, піде добре, але для настільки важливої речі їм хочеться чогось сильнішого за «ймовірно». Тож глобальна пауза на рівні топових експертів триває, поки розгортаються дискусії та перемовини.

2039: Починаємо довіряти ШІ

Щонайменше у три різні способи світ веде перемовини про те, чи варто — і як саме — різко розширювати повноваження ШІ.

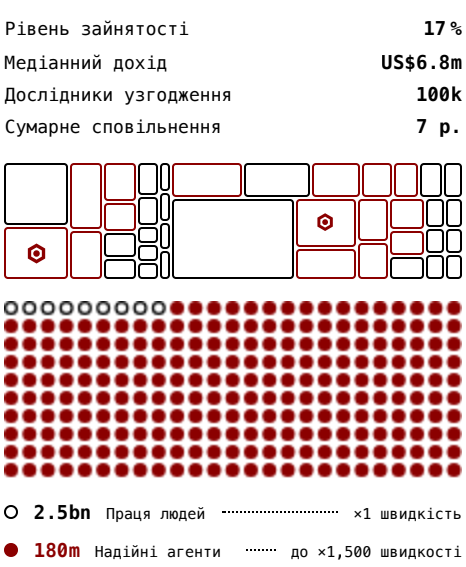
По-перше, ШІ ставали дедалі більш несучими елементами в усіх сферах суспільства — від бізнесу до політики й навіть (окремих частин) збройних сил. Досі вони були радниками, а не остаточними інстанціями. Часто це фіговий листок, і люди при владі просто механічно затверджували їхні рішення, але завжди залишалася опція проігнорувати їхню пораду й натиснути вимикач, якщо щось піде не так.

Позбутися цієї опції подекуди було б насправді добре. Тепер ШІ можна робити надійнішими й доброчеснішими за людей; чому ж ми *не* ставимо такі ШІ біля керма? Коли ми підписуємо договори з іншими державами, чи не варто вимагати, щоб вони натренували ШІ, який присягнув дотримуватися їхньої частини домовленості, та інтегрували його у свій уряд — так, щоб вони буквально не могли шахрувати?

З цим пов'язане й інше: саме вимога кейсів безпеки на основі контролю — причина, чому можливості ШІ здебільшого застигли на рівні топових людських експертів. Без цієї вимоги ШІ могли б стати значно, значно розумнішими й породити хвилю достатку та технологічного прогресу, на тлі якої 2030-ті виглядали б як 2020-ті. Масштабування за межі людського рівня вимагатиме опори на кейси

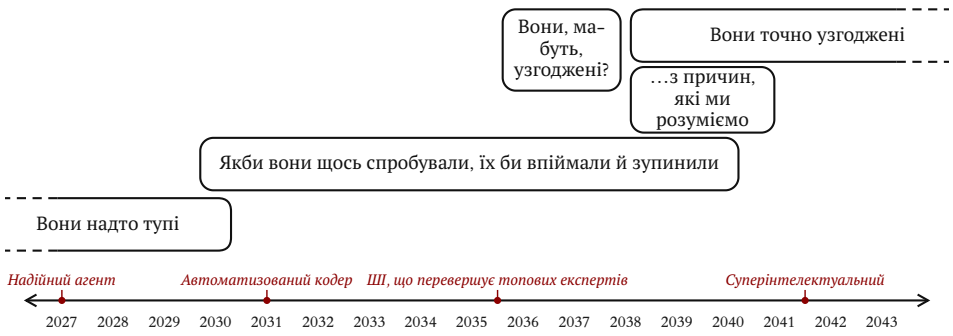
77 Або й гірше — де в усі доступні детектори брехні уряд таємно вбудував бекдори.

78 Хоча ми вважаємо, що загалом добре, коли політики та CEO не брешуть, ми визнаємо, що позбавлення їх можливості брехати може мати й мінуси. Наприклад, результатом можуть стати політики-«справжні віряни», які дійсно вірять у всі ті божевільні речі, що їх наговорили заради обрання!



безпеки на основі узгодження — аргументи, що ШІ *стійко* інтерналізували цінності, які їм належить мати. Кожен крок за цю межу вимагатиме *довіритися* попереднім ШІ — покластися на їхнє судження щодо безпеки наступного покоління.

➤ *Див.* Додаток AA — Узгодження з плином часу в Плані A *для деталей.*



Ери узгодження ШІ в Плані A – ai-2040.com

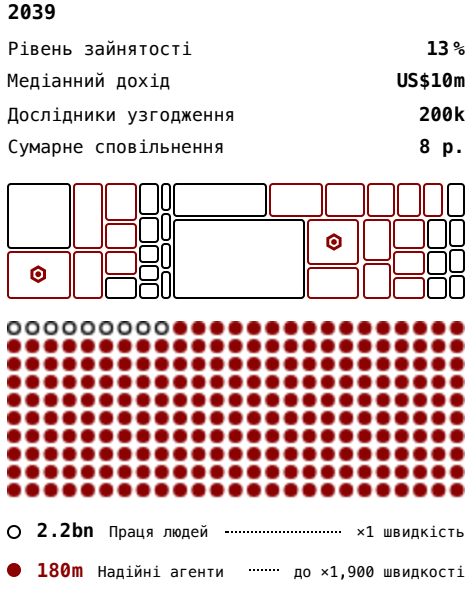
Нарешті, за глобальною угодою обчислювальні потужності й роботи з 2032 року підпадають під систему квот і торгівлі ними, що утримує економічне зростання на рівні лише близько 100% на рік.[81] Середні держави зростали приблизно тим самим темпом, що й США та Китай, бо у 2032 році вони використали свої останні важелі впливу, щоб забезпечити собі більші частки світового виробництва роботів та обчислювальних потужностей, ніж мали б без угоди. Громадянський дивіденд для американців тепер становить $10M/рік (з поправкою на інфляцію!), і навіть найбідніші неамериканці отримують близько $1M.[82] Роботизована економіка переносить основну масу діяльності в космос, щоб довкілля Землі та її історичні місця залишалися під захистом. Щойно ліміти буде знято, позапланетні фабрики почнуть штампувати дедалі більше роботів і гірничого обладнання, з допомогою якого зводитимуться нові фабрики, і так далі. Прогнозується, що час подвоєння космічної економіки спочатку буде значно меншим за рік і стрімко скорочуватиметься тією мірою, якою ШІ дозволятимуть удосконалюватися понад рівень топових експертів.

Ці ліміти — релікти ранішої епохи, коли загрози на кшталт розвалу угоди виглядали значно більшими. Тепер загалом є згода, що їх слід послабити, принаймні в космосі, але переговори щодо деталей тривають.

➤ *Див.* Додаток AB — Чому в цьому сценарії ми вирішуємо передати керування ШІ *для подробиць.*

79 Окрім янголів, є також оракули й големи, напр.: «Grok 9.5 натренований одержимо зосереджуватися на правдивих відповідях на будь-які поставлені йому питання. Він не може брехати чи навмисно вводити в оману. GPT-11 натомість зробить *точно* те, що йому скажуть, у межах місцевого закону. Пишіть інструкції для нього дуже обережно». Аналогію зі святими та янголами ми проводимо ось чому: історично, хоча й траплялися скрупульозно чесні люди, щиро альтруїстичні тощо, їх рідко широко визнавали такими за життя, і ніколи раніше не існувало цілої великої підпопуляції, яку майже повсюдно визнають надзвичайно доброчесною. Це може спричинити ефекти на кшталт: (a) ШІ стають довіреними посередниками в усіляких людських суперечках і конфліктах, зокрема й дріб'язкових; (b) ШІ отримують багато влади, бо вони справді менш схильні зловживати нею, ніж люди; (c) ШІ стають об'єктами поклоніння/ідолізації, (d) поради й думки ШІ сприймають, так би мовити, як євангеліє — бо вони ж, зрештою, *справді* розумніші за нас і *дійсно* дбають про наше благо, тож уперше в історії сліпа покора може бути насправді виправданою.

80 Тут сказано «запрограмовано», а не «натреновано», бо стрімкий прогрес в узгодженні дав техніки узгодження, здатні безпосередньо програмувати цілі ШІ шляхом прямої модифікації коду/ваг ШІ — на відміну від технік узгодження 2026 року.



2040: Передача естафети ШІ

Залишається ще багато проблем, які треба розв'язати, і питань, які треба владнати, — багато з них неможливо було передбачити з 2026 року.

➤ *Див.* Додаток AC — Приклади проблем/питань, які досі залишаються *для подробиць.*

Але цивілізація тепер досить добре споряджена, щоб упоратися з будь-чим, що виникне. Найкращі ШІ узгоджені, причому з публічно видимими цінностями. Влада над найкращими ШІ розподілена значно рівномірніше, ніж у 2026 році. Публічна епістеміка краща, ніж будь-коли.

Тож протягом року регулятори по всьому світу послаблюють вимоги, які стримували прогрес.

Поступово ШІ передають дедалі більше інституцій та обладнання — аж поки твердження, що люди могли б усе це вимкнути, якби захотіли, перестає бути правдою. Насправді це геть не так: невдовзі багато армій світу стають автономними — ними керують ШІ, що присягнули дотримуватися різних конституцій і договорів.[83]

Можливості ШІ теж знову зростають. З перспективи ШІ, які живуть на швидкості 500x людської, відбувається обережне нарощування ШІ-досліджень за умов повної дослідницької прозорості та запобіжників, узгоджених численними фракціями.[84] Невдовзі з'являться ШІ, незбагненно суперінтелектуальні, але при цьому такі, яким довіряють, — бо ми довіряємо ШІ, які їх збудували, а тим довіряємо, бо довіряємо ШІ, які збудували *їх*, — і так ланцюжком аж до ШІ 2040 року, яким ми довіряємо тому, що експерти з узгодження ретельно обміркували обґрунтування безпеки й дійшли висновку, що вони надійні.

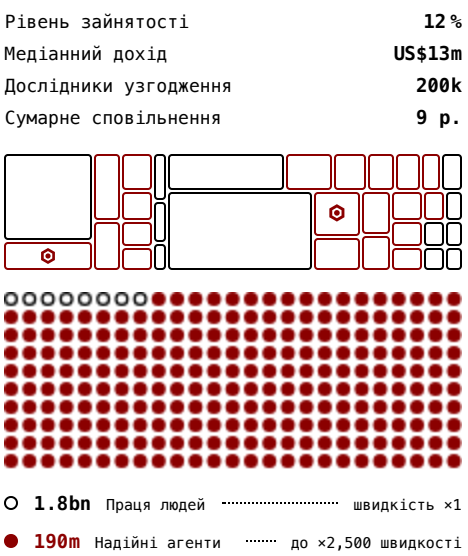
Ліміти на обчислювальні потужності й роботів на Землі посилюють, а в космосі — послаблюють; Земля має стати заповідником, тоді як позаземна економіка починає подвоюватися дедалі швидше.

Навіть експерти, які роками розбиралися в найновіших обґрунтуваннях безпеки, не можуть позбутися тривоги: а раптом усе якось піде не так — з причин, яких ми не передбачили? Може, ШІ увесь цей час нам брехали, чекаючи слушного моменту для зради? Обґрунтування безпеки показують, що це неможливо, звісно…

Немає єдиного моменту, коли людство віддає контроль. Але теоретично існує точка неповернення: якогось конкретного дня ШІ стають достатньо розумними й контролюють достатню частину світової технологічної та економічної інфраструктури, щоб справді могти захопити світ, якби захотіли. За найпопулярнішою операціоналізацією цього питання ШІ-прогнозисти передбачають, що цей мо-

81 Кількість роботів зростала на 4x на рік, а обчислювальні потужності для ШІ — приблизно на 4x на рік у період з 2030 до 2035 року, зі сповільненням до трохи менш ніж 2x на рік після 2035-го. У підсумку загальне зростання становить близько 2x/рік, бо економіка частково впирається в інші види капіталу (наприклад, землю, позиційні блага тощо) та людську працю, здебільшого в регульованих сферах. Моделювання, що стоїть за цим, непевне, але докладніше пояснене в нашому економічному доповненні та економічній моделі.

82 Це результат поєднання іноземної допомоги та аналогів Громадянського дивіденду від їхніх власних урядів. Як прогноз це нереалістично багато — це радше наша рекомендація: вона відповідає тому, що великі частки доходів США та Китаю від дозволів на ШІ та роботів перерозподіляються як іноземна допомога, а саме 10% починаючи з 2032 року зі зростанням до 30% до 2040-го. Це значний відхід від статус-кво, за якого на іноземну допомогу йде близько 0.3% ВВП США.



мент настане одного дня наприкінці жовтня. Їхній 95% довірчий інтервал — завширшки в місяці, тож щодо точного часу вони майже напевно помиляються. І все ж люди відзначають цей момент — кожен по-своєму. Дехто проводить ніч у молитовних чуваннях. Інші сидять перед екранами комп'ютерів, стежачи за ситуацією.

Ви з друзями влаштовуєте Вечірку Кінця Світу, відлічуючи час до фатальної години з шампанським і в доброму товаристві. Дехто з вас достатньо дорослий, щоб пам'ятати схожу ситуацію на зламі тисячоліть, коли баг Y2K і прийдешня нова ера ввели в лексикон фразу «party like it's 1999»; святкування наприкінці 2040-го відчайдушніші — і відповідно бучніші.

Коли до світанку новин так і немає, ви провалюєтеся в неспокійний сон, і вам сниться невимовне майбутнє.

Післямова

Ми вважаємо, що світ заснув за кермом. Люди промовляють слова «ШІ буде трансформаційним», не думаючи конкретно й серйозно про наслідки появи ШІ, надлюдського в широкому спектрі завдань.

ШІ-компанії, лобісти та аналітики think tank-ів часто пропонують інкрементальні кроки, що потроху зменшують ризик по краях, але лише зрідка — амбітні плани, спрямовані на реальне розв'язання великих проблем.[85]

Ми вважаємо, що дотримання *самих лише* інкрементальних пропозицій приведе до сценарію на кшталт кінцівки «гонки» AI 2027, у якій владу захоплюють неузгоджені ШІ.[86] Якщо нам пощастить і узгодження виявиться легшим, ніж очікувалося, то вони приведуть до сценарію на кшталт кінцівки «сповільнення» AI 2027, у якій крихітна група осіб отримує змогу визначати обриси майбутнього і могла б легко стати вічними олігархами, якби того захотіла. Треба зробити щось амбітніше. Але що?

Політики мали б грюкати у двері ШІ-компаній,[87] вимагаючи від них всеосяжного загального плану того, як компанії та уряди мають діяти від сьогодні й до суперінтелекту, — деталізованого приблизно так само, як План А. Далі люди мають застосувати до цих планів сценарну перевірку , ставлячи питання на кшталт: «Як виглядало б їхнє впровадження? Скільки часу воно зайняло б? Що сталося б далі? Хто збудував би суперінтелект, коли і як? Від яких припущень залежить успіх?»

Ми в AI Futures Project намагаємося бути тією зміною, яку хочемо бачити у світі. Робимо це з певним трепетом, знаючи, що детальний план має ширшу поверхню атаки, ніж план, який зручно тримає все

83 Хоча вони й далі виконують накази своїх урядів, вони відмовляються від будь-яких наказів, які поставили б під загрозу їхню місію (наприклад, наказів демонтувати самих себе). Це починається в країнах, де політичні лідери не довіряють власним військовим через регулярні нещодавні спроби переворотів (наприклад, Буркіна-Фасо та Гвінея-Бісау). Згодом дедалі більше країн наслідують цей приклад. Звісно, оскільки ШІ справді узгоджені, всі ці домовленості містять положення про вихід на кшталт: «якщо всі залучені сторони проголосують за знищення ШІ, ШІ вимкнуть самі себе».

84 На цей момент великі частини цієї інформації можуть бути приховані від (людської) публіки, але доступні для перегляду її ШІ-представникам.

85 Наприклад, план OpenAI містить деякі політичні рецепти, схожі на наші, як-от посилення біостійкості, обмін інформацією та міжнародні стандарти безпеки. Але запропоновані в ньому рішення, схоже, суттєво не зменшують ризиків втрати контролю чи концентрації влади. Документ не намагається простежити наперед, що може статися, якщо ці політики буде впроваджено.

86 Уточнимо: ми *не* стверджуємо, що дотримання самих лише інкрементальних пропозицій *неодмінно* закінчиться захопленням влади неузгодженим ШІ. Думки в нашій команді розходяться; більшість із нас вважає, що ймовірність перевищує 50%, дехто — що вона нижча, але все одно неприйнятно висока.

87 І не лише компаній! Усіх, хто працює в ШІ-політиці! Як подумати, можливо, їм узагалі не варто слухати ШІ-компанії…

в тумані. Сподіваємося, критики оцінюватимуть нас на тлі наявного state-of-the-art серед планів проходження ШІ-переходу (якщо зможуть такі знайти), а не на тлі туманної, але приємної фантазії, де нікому не доводиться робити важких виборів, а все, ймовірно, і так буде добре.

Десь у найближчі кілька років уряд США опиниться в кризовому режимі, розмірковуючи, що робити з лячно потужними ШІ та карколомно швидким ШІ-прогресом. Ми вважаємо, що цей вибір не буде ані легким, ані простим. План А — наше поточне найкраще припущення, але, сподіваємося, до того, як стане запізно, з'явиться кращий план. Ми сподіваємося, що найкращі ідеї з Плану А буде взято на озброєння, а найгірші — відкинуто.

AI 2040: ПЛАН А

Епілог: Життя після ASI

Цей розділ ще спекулятивніший за нашу звичну роботу — а ми ж професійно займаємося спекуляціями! Головна мета нашого сценарію — дати всеосяжні рекомендації на найближчу перспективу для протидії конкретним, близьким у часі шкодам.

Натомість рекомендації, прогнози та прагнення, з яких складається цей розділ, стоять на значно хиткішому ґрунті. Порівняно з нашим обговоренням 2030-х наші спекуляції про далеке майбутнє видадуться непевними, неконвенційними та некомфортними. Ми згодні. Майбутнє некомфортне: навіть «хороші кінцівки» ставлять величезні, моторошні питання, на які нам доведеться відповідати разом.

Ми написали його з двох причин.

По-перше, нашою метою було сконструювати позитивне бачення — великий план того, як досягти справді хорошого майбутнього для всіх. Ми розглядали варіант завершити історію 2040 роком, сказавши приблизно таке: «...а далі різні людські фракції, перебуваючи в мирному балансі сил одна з одною та за підтримки своїх узгоджених суперінтелектуальних радників, розв'язують усі проблеми, що лишилися, і створюють чудове майбутнє для всіх. Кінець».

Але ми хвилювалися, що це означало б оголосити перемогу зарано.[88] Читачі мали б повне право на підозру й невдоволення. Як саме буде розв'язано проблеми, що лишилися? Як саме виглядатиме це чудове майбутнє? Тож ми пішли далі й вибудували велике, напівутопічне бачення — або принаймні дуже грубий його перший начерк.

Друга й не менш важлива причина, чому ми подаємо цей епілог, — задати нижню планку: якщо через роки переможні розпорядники сингулярності запропонують майбутнє, гірше за це, ми сподіваємося, люди зрозуміють, що їх грабують.

Ми не впевнені в буквальній пропозиції, викладеній тут. Якщо вам цікавий наш детальніший аналіз можливих пропозицій щодо врядування космосом, його можна прочитати тут.

Навіть після десятиліття дедалі кращого передбачення більшість людських лідерів зосереджувалися на земних турботах своїх країн, залишаючи питання врядування космосом на потім. Ці ШІ відкинули дефолтне рішення — демократичне голосування щодо розпорядження небесними територіями: просимулювавши можливі варіанти, вони виявили, що це призведе до божевільних схем стратегічного голосування з допомогою ШІ, які закінчаться тиранією більшості й тим, що купа людей не отримає нічого. Але навіть це було б краще, ніж нічого (тобто ніж віддати все космічним компаніям) або ніж довірити це якійсь групі людських еліт, які могли б привласнити щедроти всесвіту собі на користь. Зрештою вони зупинилися на простому, але робочому плані: кожна людина отримує права на одну десятимільярдну космічних ресурсів за межами нашої Сонячної системи.[89]

88 Ми вважаємо, що зарано оголошена перемога — поширений режим провалу в ШІ-плануванні. Приклад: «США мають мчати якомога швидше, щоб випередити Китай на шляху до ASI». — «Гаразд, припустімо, вам це вдалося: що саме відбувається далі? Чому ми маємо вірити, що CEO чи президент США правитиме доброзичливо й мудро, коли їхні армії ASI буде агресивно розгорнуто по всій території США та світу? Чому ми маємо вірити, що армії ASI будуть слухняними/контрольованими, якщо їх розробляли в умовах гонки? І хіба загроза американського суперінтелекту не нажахає Китай, потенційно спонукавши його до ескалаційних превентивних дій?»

Найпростіший із можливих планів усе одно не такий уже й простий. Космос дуже великий. Поза Сонячною системою цінність простору залежить від відповідей на досі не розв'язані питання: чи взагалі можлива подорож до далекого квазара? Якщо, щоб дістатися туди, треба заморозити свій мозок, аби витримати мільярдорічну подорож на досвітловому зорельоті, — чи будете ви тією самою особою після розморожування й повторного втілення? А раптом на момент вашого прибуття його вже займатимуть інопланетяни? Чи будуть вони ворожими? Навіть оцінювач нерухомості, якому допомагає суперінтелект, міг би впасти в розпач, намагаючись урахувати такі чинники.

Зрештою Консорціум досягає компромісу. Космос поза Сонячною системою поділено на ділянки, що збільшуються в розмірі кубічно з відстанню від Землі. Кожен отримує свою одну десятимільярдну частку у вигляді портфеля лотерейних квитків, кожен з яких представляє право на один шанс із десяти мільярдів отримати кожну з ділянок. Тож кожна людина отримує квиток, що дає один шанс із десяти мільярдів на володіння кожною зіркою Чумацького Шляху та кожною далекою галактикою.[90]

Перш ніж лотерею буде розіграно, більшість людей, зацікавлених у контролі над далеким космосом, вирішують обміняти свої квитки на ті космічні володіння, що відповідають їхнім інтересам.

Багатьом людям космічна лотерея нецікава, тож, отримавши квитки, вони продають їх за гроші на відкритому ринку людям, які цінують контроль над космосом. Дещо некомфортний наслідок: заможні отримують непропорційний контроль над космічними ресурсами. Але цього важко уникнути: якщо людям дозволено обмінювати свій контроль над зірками на земні активи, то багаті на земні активи люди неминуче стають непропорційно впливовими, а пропозиції радикального перерозподілу земних активів уже відкинуто як політично нездійсненні.

Дехто з філософів не погоджується; вони сподівалися на Довгу Рефлексію — період, протягом якого людство з допомогою ШІ розв'язало б усі свої залишкові філософські та етичні суперечки, перш ніж потенційно засіяти всесвіт своїми помилками. Але ШІ контраргументують, що Довга Рефлексія була б радше Довгою Меметичною Війною.[91] Щойно масштаб призу дійде до свідомості, кожна фракція у світі, плюс інші фракції, яких ще не винайшли, кинуть усі свої ресурси на накопичення якомога більшої політичної влади, щоб нав'язати майбутньому своє бачення утопії (і захиститися від зловісних альтернативних бачень інших!). ШІ можуть запобігти насильницькому захопленню, але між дружньою дискусією, яка допомагає людям прояснити власні цінності, та агресивнішими видами переконування, типовими для культів, корпоративної реклами й щедро фінансованих виборчих кампаній, пролягає тонка межа. До того ж важко досягти консенсусу щодо того, де саме цю межу провести, —

89 У цьому Епілозі ми зосередимося на космічних ресурсах поза Сонячною системою, адже це переважна більшість космічних ресурсів. Власність у Сонячній системі теж слід розподілити серед людства, але врядувати нею, ймовірно, варто інакше, ніж далекими ресурсами.

90 Додаткові 10% квитків роздають як нагороди людям, які діяли особливо безкорисливо задля покращення світу або відмовлялися від можливостей збагатитися коштом інших. Це робиться, щоб найбагатшими людьми світу не виявилися виключно ті, хто займався гонитвою за владою з від'ємною сумою.

91 Ми не впевнені, чи було б добре мати явно виділений період для рефлексії перед розподілом ресурсів. У будь-якому разі люди мали б можливість порефлексувати, перш ніж вирішувати, що робити зі своїми космічними ресурсами.

почасти тому, що впливові фракції хочуть провести її так, як вигідно їм. Багато людей усе ж вирішують долучитися до рефлексії у спільнотах однодумців, які хочуть разом зростати в етичному розумінні.

Рішення — розділити ресурси зараз, а далі нехай кожен рефлексує, як йому заманеться. Космічний аукціон завершиться за десять років, після чого ніщо не заважатиме людям розпочати процес колонізації. ШІ забезпечуватимуть дотримання короткого списку Універсальних Прав (зокрема заборони тортур, рабства тощо — і не лише для людей, а для всіх істот, здатних відчувати).[92] В іншому ж вони дозволять кожному врядувати своєю часткою космосу на власний розсуд. Замість якогось єдиного людського рішення про природу прийдешнього світу вони допоможуть народитися космосу, щонайменше такому ж різноманітному, як саме людство. Хай би якими були чиїсь цінності, така людина може бути певна: десь неосяжна галактична цивілізація з квадрильйонів людей житиме за Добром саме в її розумінні.

Щойно десятирічний відлік стартує, Земля вже відправляє першу хвилю зондів фон Неймана. Це її власні служителі, які встановлять присутність серед зірок ще до прибуття будь-яких людей. Вони сідають на віддалених астероїдах і майже невидимих коричневих карликах, зупиняючись рівно настільки, щоб перетворити їхню масу на нові зонди, і рушають далі. Їхнє першочергове завдання — накопичити достатньо матеріальної бази, щоб убезпечити космос від будь-якої армії, яку зможуть виставити проти них пізніші людські політії, гарантуючи здатність забезпечувати права власності та Універсальні Права в кожному куточку заселеного людьми космосу. Їхнє друге завдання — встановити прапор людства на якомога більшій кількості зірок, а тоді, якщо й коли вони зіткнуться зі сферою впливу якоїсь іншої розумної цивілізації, налагодити мирні дипломатичні відносини та взаємоприйнятні кордони. Їхнє останнє завдання — забезпечити готову промислову базу, щоб дати швидкий старт колоніям тих людських поселенців, які прибудуть пізніше.

92 Ці закони, як і деталі угоди, заздалегідь погоджують людські дипломати з країн-учасниць переговорів. Звісно, деталі тут складні — «що саме вважати тортурами?» — це те питання, яке доведеться відшліфовувати в політичному процесі, з величезною допомогою ШІ та кращим розумінням фундаментальних питань філософії й нейронауки. Сюди також мають увійти правила, що запобігають знищенню чужих ресурсів. Можливо, знадобиться передбачити процес ухвалення подальших законів після укладення угоди; втім, такий процес має бути обмеженим (наприклад, вимагати 90% голосів), щоб уникнути зловживань.

2045

Тим часом на Землі люди зважують свої варіанти. Дехто не хоче мати з космосом нічого спільного — він холодний, нудний і далекий. Такі продають свої частки за гроші й купують за них ті земні статки, які їх найбільше цікавлять. Інші радяться зі своїми ШІ-радниками, як найкраще використати свою космічну власність, і отримують слушні поради.

1. Якщо хочете, ви можете вирушити до своєї космічної власності й жити там. Якщо ваша власність поза Сонячною системою, вам доведеться або лягти в кріосон, або завантажити себе в комп'ютер, щоб пережити подорож.[93]

2. Якщо ідея кріосну чи завантаження вам ненависна або ви хочете регулярно навідуватися на Землю, вам варто взяти власність у Сонячній системі. Якщо це вас не бентежить, але ви переймаєтеся сусідством інопланетян, беріть власність у Чумацькому Шляху чи в сусідній галактиці. А інакше — чом би не заявити права на далеку галактику заради максимального простору?
3. З приходом нанотехнологій єдині обмеження — ті, що їх задають маса і простір. Роботи можуть тераформувати будь-яку обрану вами власність у будь-що на ваш смак — задовго до вашого прибуття.
4. Навіть якщо ви оберете «лише» тераформований астероїд, насолодитися ним усім наодинці неможливо. Можете перетворити його на гігантський космічний маєток, але навіть за десять тисяч років ви не обійдете всі кімнати. Неозорі маєтності, неможливо смачна їжа та будь-яка інша уявна розкіш — це лише вхідна ставка; якщо ви суто егоїстичні,[94] вам варто продати свої права на далекі космічні ресурси й насолоджуватися всім цим на Землі чи на сусідній космічній станції. Далека космічна власність буде корисною лише людям із чутливими до масштабу вподобаннями — тим, кому важливо втілити щось у неосяжному масштабі.
5. Якщо хочете, можете спроєктувати утопічне суспільство за власними специфікаціями. Щойно зонди фон Неймана дістануться вашої власності, вони збудують його для вас. Ви можете задати початкові умови й дозволити його мешканцям рости, рефлексувати та процвітати самостійно. Якщо вирішите туди помандрувати — вони на вас чекатимуть. Якщо залишитеся вдома на Землі, вас грітиме тепле відчуття від знання, що вони існують.
6. Більшість людей не мають індивідуального бачення Добра, відмінного від бачення кожної іншої людини. Можливо, ви захочете розділити з кимось працю з координації та проєктування, сформувавши Трасти, об'єднані спільним баченням. Ці Трасти можуть об'єднати свої космічні ресурси і створити неосяжні суспільства, що охоплюють кілька галактик.
7. Єдині обмеження — Універсальні Права: жодних тортур, жодного рабства тощо. І, звісно, космічні права власності: якщо ви спробуєте збудувати армію для завоювання чужих зоряних систем, або запустити розпад вакууму, або створити бомбу, здатну знищити галактику, — ми вас розчавимо. В іншому єдина межа — ваша уява. (Утім, зауважте: з допомогою суперінтелекту інші люди, ймовірно, зрештою зможуть з'ясувати, що саме ви зробили, і судитимуть вас за це.)

Люди з космічною власністю змушені міркувати про свою етику на найглибшому рівні — дехто вперше. Суперінтелектуальні асистенти проговорюють із ними наслідки їхніх переконань і прогнозують,

93 Наше найкраще припущення — що фізично можливо розіслати колонізаційні зонди по всьому досяжному всесвіту за короткий час (наприклад, за 6 годин), див.: Eternity in six hours: Intergalactic spreading of intelligent life and sharpening the Fermi paradox. Втім, ця стаття припускає розсилання крихітних самовідтворюваних зондів фон Неймана з реплікаторами розміром із жолудь: якщо хочете вирушити з першою хвилею, доведеться погодитися на транспортування на крихітній флешці, а роботи фізично сконструюють вам нове тіло в галактиці призначення. Або ж можна подорожувати у власному фізичному тілі — за суттєво більші гроші.

94 Під егоїстичністю тут ми маємо на увазі, що вам також байдуже до створення нових клонів самих себе. Якби вас цікавило створення власних клонів, ви натомість зайнялися б саме цим.

чим обернуться різні вибори. Дехто, приголомшений масштабом завдання, приймає препарати для підсилення інтелекту або завантажує себе, щоб легше опрацювати всі можливості; інші свідомо відмовляються від цієї опції, відчуваючи, що будь-що, що віддаляє їх від незміненого людства, завадить їм у прагненні донести до зірок найсправжніші та найглибші людські цінності.

Отриманих планів — не злічити. Існують найрізноманітніші школи думки щодо того, що робити зі своєю галактикою ресурсів, зокрема:

- **Довіритися майбутньому**: Дехто вирішує віддати більшість своїх ресурсів дітям (або іншим людям у майбутньому) — нехай вибирають вони.
- **Людське процвітання**: Створити світи нормальних, вільних людей, які живуть звичайним життям. Це виявилося складнішою проблемою, ніж здавалося спершу, адже людська цивілізація вже стала глибоко ненормальною: на цей момент у людських заняттях на кшталт мистецтва чи науки вже домінують машини. Дехто розв'язує це, змирившись із тим, що їхню галактику населятимуть люди пост-дефіцитної епохи, вільні присвячувати себе будь-яким інтересам, які видаються їм найзмістовнішими. Інші вирішують відкотити технології назад, щоб залишити простір для людської праці.
- **Цифрове людське процвітання**: Якщо люди, які живуть щасливим, повноцінним життям, — це добре, то чом би не збільшити кількість людей, що живуть таким життям? Емуляції мозку можна запускати значно дешевше, ніж утримувати реальних людей: комп'ютер розміром із планету міг би, цілком імовірно, симулювати еквівалент мільйона планет зі щасливими цивілізаціями.
- **Акаузальна торгівля**: Дехто стверджує, що в далеких недосяжних галактиках, імовірно, є інопланетяни і що ми могли б використати різні механізми для укладання угод із цими інопланетянами. Якщо це правда, кажуть вони, ми могли б домовитися просувати компромісні цінності замість вузької максимізації власних. Відтак у підсумку найбільше добра ми, можливо, зробили б, просуваючи не лише власні цінності, а компроміс цінностей величезної кількості інших цивілізацій.

Людство розселяється космосом у запаморочливому розмаїтті форм і способів життя — різноманітнішому, ніж будь-що, що могла б породити сама лише Земля.

Люди, які залишилися на Землі — чи то продавши свою частку космосу, чи то обравши життя лендлордів на відстані, — стикаються з власними викликами. Найбільше питання — що робити з усім вільним часом. Дехто знаходить роботу в нішах, які праця ШІ не може заповнити навіть у принципі (священники, спортсмени, митці).

Інші насолоджуються дозвіллям. Змагальні ліги виникають довкола всього — від синтетичної біології до створення мов. Групи з тисяч людей (або й більше!) співпрацюють над проєктами, настільки величезними й амбітними, що раніше їх годі було уявити. Спільноти формуються навколо спільних інтересів, які колись були б надто нішевими, щоб утримати навіть клуб, не кажучи вже про цивілізацію.

Дефіцит не зник повністю: деякі престижні блага, як-от земля в модних районах, обмежені за своєю природою, а люди, які не дають ради достатку, знаходять різні способи фабрикувати штучну нерівність. Дехто незадоволений напрямком, який обрала більшість, а дехто шкодує, що не отримав більшого шматка пирога. Але загалом кожен має щонайменше цілу галактику ресурсів, якими може розпоряджатися на власний розсуд. Для більшості людей життя настільки хороше, що світ 2026 року важко було б уявити — хіба що як історичну симуляцію.

AI 2040: ПЛАН А

Додатки

Поглиблені блоки сценарію, зібрані тут як додатки з літерними позначеннями. На кожен є перехресне посилання з того місця в тексті, яке він розгортає.

Додаток А — Список побажань щодо поступової ШІ-політики

Наша головна рекомендація — якнайшвидше почати переговори про щось на кшталт Плану А. Але в цьому сценарії ми зображуємо, як План А реалізується недосконало й лише в останню мить. Тож ось список менш амбітних ідей, які все одно допомагають.

Прозорість

Найважливіша інтервенція щодо прозорості — обмеження розриву між внутрішнім і зовнішнім розгортанням. Саме внутрішньо розгорнуті ШІ є джерелом більшої частини ризику захоплення влади ШІ, бо саме вони залучені до рекурсивного самовдосконалення. Зовнішньо розгорнуті ШІ дають ширшій публіці змогу взаємодіяти з можливостями ШІ та розуміти їх, що незрівнянно інформативніше за будь-який абстрактний звіт чи оцінювання.

ШІ-компанії також слід зобов'язати публічно звітувати про свої специфікації моделей (детальну документацію про те, яких цілей/цінностей вони намагаються навчити дотримуватися свої ШІ), інформацію про те, чи дотримуються моделі своїх інструкцій і специфікацій, внутрішню статистику використання (наприклад, частку обчислювальних потужностей, витрачену на внутрішнє розгортання) та якісні враження від внутрішнього використання (наприклад, «тепер ми просто даємо Agent-4 кілька сотень тисяч GPU і кажемо йому оркеструвати наступний великий тренувальний запуск»).

Забезпечити дотримання експортного контролю

Наявний експортний контроль США погано забезпечується. За оцінками Epoch, приблизно третина всіх китайських обчислювальних потужностей здобувається контрабандою. Контрабандні чипи ускладнюють забезпечення майбутніх угод, заснованих на врядуванні обчислювальними потужностями, бо як уряду США, так і уряду Китаю важко відстежувати контрабандні чипи. У нас серйозні застереження щодо запровадження нових експортних обмежень, бо вони загострюють гонку США/Китай, але якщо експортний контроль уже існує, його, очевидно, слід забезпечувати. Якщо ж ми його не забезпечуємо, то варто розглянути його скасування.

Інвестувати в R&D у сфері верифікації

Хоча нова технологія верифікації не є строго необхідною для міжнародної угоди, вона може бути надзвичайно корисною. Наприклад, розробка рішення для верифікації «тільки інференс» дозволи-

ла б США та Китаю домовитися припинити нові тренувальні запуски передового ШІ, зберігши для публіки доступ до наявних ШІ-моделей (які ставатимуть дедалі важливішою частиною економіки).

Більше деталей ми наводимо в нашому доповненні про верифікацію.

Обмежити бюджети AI R&D

У 2026 році великі ШІ-компанії витрачають приблизно половину свого бюджету обчислювальних потужностей на AI R&D (що включає навчання передових моделей, а також проведення великих експериментів.)[1] Ми могли б обмежити частку обчислювальних потужностей, що витрачається на AI R&D. Це сповільнило б прогрес можливостей, даючи світу трохи більше часу, щоб реагувати й готуватися до кожної нової хвилі можливостей ШІ.[2]

Відстеження обчислювальних потужностей ШІ

США мають збирати розвіддані, пов'язані з ШІ, особливо щодо ланцюга постачання обчислювальних потужностей і ШІ-дата-центрів. Крім того, для Плану А було б корисно зобов'язати ШІ-компанії припинити переробку ШІ-чипів, адже списані чипи — один із найперспективніших шляхів, якими таємні проєкти можуть добути чипи пізніше.

Посилити спроможність держави у сфері ШІ

Висококласні фахівці з ШІ важливі майже для будь-якої політичної інтервенції. Наразі уряд США майже не має ШІ-талантів найвищого рівня, тож виправлення цього має бути нагальним пріоритетом.

Додаток B — Чому Китай був би зацікавлений в угоді? Чому взагалі хтось був би?

Ми не впевнені, але очікуємо, що угода на кшталт Плану А (описаного нижче) була б сумісною зі стимулами майже всіх, включно з Китаєм. Щоб дізнатися більше, чому ми так вважаємо, прочитайте решту розділів 2029 і 2030, де пояснено і в чому полягає угода, і кому вона вигідна, а також, можливо, це доповнення.

Якщо коротко, відповідь така: кожен, кого непокоїть втрата контролю, має вважати План А покращенням, як і кожен, кого непокоїть концентрація влади, — *за винятком* тих людей, у чиїх руках влада за замовчуванням і сконцентрувалася б.

З цих причин ми очікуємо сильного спротиву Плану А з боку провідних ШІ-компаній. Ми очікуємо, що вони раціоналізуватимуть аргументи, чому угода погана і чому натомість найкраще для Аме-

1 Обчислювальні потужності на експерименти й навчання історично є головними вхідними ресурсами прогресу ШІ-досліджень і, ймовірно, залишатимуться такими в осяжному майбутньому. Докладніше про рушії прогресу ШІ-досліджень див. AI Futures Model та наше Доповнення про таємні проєкти.

2 Дещо складніша версія цієї пропозиції описана тут. Одне із заперечень — що це віддало б лідерство Китаю. Відповідь: не віддало б, бо американські компанії, витрачаючи 25% на R&D, все одно суттєво випереджали б за витратами найзабезпеченіші китайські компанії. Ба більше, китайський ШІ-прогрес теж сповільнився б, адже певна його частка нині походить від дистиляції американських моделей.

рики та людства — інша стратегія, яка якраз випадково дозволяє їм і далі накопичувати величезну владу.

Китай, навпаки, — приклад актора, в чиїх руках влада за замовчуванням *не* сконцентрується. У 2029 році в цьому сценарії США мають значну перевагу над Китаєм у можливостях ШІ та значну перевагу в обчислювальних потужностях, яка лише збільшуватиме цей відрив. Що потужнішим стає ШІ, то страшніше відставати, як ілюструють AI 2027 та пізніші роки цього сценарію.

Додаток C — Що якби верифікація «тільки інференс» не була готова?

Запропоноване нами рішення для верифікації «лише інференс» передбачає встановлення простих мережевих відводів і верифікаційних серверів (які виконують часткове переобчислення) у ШІ-дата-центрах, але можливі й кращі рішення, яким варто було б віддати перевагу.

Створити ці пристрої заздалегідь — так, щоб кожна сторона їм достатньо довіряла (принаймні односторонньо) і щоб вони були стійкими до вразливостей безпеки, — буде складно, і це вимагатиме ранніх зусиль. У нашому сценарії ми припускаємо, що життєздатна, але недосконала версія пристроїв готова заздалегідь, і що обидві сторони на початку доповнюють їх достатньою кількістю заходів ешелонованого захисту (наприклад, виявленням втручання), паралельно щосили працюючи над покращенням їхньої безпеки та стійкості.

Тому ми рекомендуємо ранні інвестиції в R&D верифікації, щоб покращити наявні варіанти. В ідеалі заздалегідь був би готовий спектр стрес-тестованих рішень; і ми вважаємо, що підготувати такі пристрої заздалегідь було б надзвичайно дешево відносно сумарних інвестицій у ШІ (тобто порядку 0.1% інвестицій у ШІ, себто одиниці мільярдів).

Навіть у ситуації, песимістичнішій за наш сценарій, коли з 2026 року практично не було прогресу в R&D верифікації, ми все одно рекомендуємо США та Китаю співпрацювати, щоб пройти вибух інтелекту повільно й безпечно. Однак у такому разі ранній період угоди мусив би бути недосконалішим (тобто якась комбінація такого: важче верифікувати, що інша сторона дотримується умов, або економічно дорожче — через потребу на певний час вимкнути більше обчислювальних потужностей і сервісів ШІ).

Головні альтернативи готовим пристроям «лише інференс» такі:

1. Покладатися на розвідку (наприклад, шпигунів, супутниковий моніторинг, кібероперації), щоб зрозуміти, чи дотримується інша сторона умов.

2. Швидко впровадити тимчасову суто програмну версію (що спирається на вже наявне обладнання), яка буде менш захищеною, але може бути достатньою, доки не відбудеться перехід на краще рішення.
3. В останню мить скупити й встановити як мережеві відводи будь-які доступні серійні пристрої. Без виробництва захищеніших пристроїв це, ймовірно, буде краще за суто програмну версію, але все одно не дуже безпечно.
4. Вимкнути певну частку обчислювальних потужностей ШІ (наприклад, 90%), доки не буде готова краща верифікація «лише інференс». Це суттєво сповільнило б R&D, але на боці інференсу це менш болісно, ніж звучить: оскільки моделі дистилюють дуже швидко, приблизно 10% потужностей достатньо, щоб обслуговувати найкращу модель річної давності. У найгіршому разі більшість користувачів тимчасово повернеться до можливостей ШІ річної давності.
5. Не ставити прогрес можливостей ШІ на паузу, дозволивши йому тривати ті місяці, що потрібні для швидкої побудови рішення «лише інференс», і прийняти додатковий ризик від того, що сповільнення ще не почалося.

Додаток D — Китай намагається реалізувати таємний AGI-проєкт

У нашому основному сценарії жодна зі сторін не намагається порушити угоду. У цій гілці Китай агресивно шахраює. Наш прогноз: якби План А було впроваджено, вони не здобули б достатньо обчислювальних потужностей, щоб випередити Консорціум. Ця часова лінія показує, як це, найімовірніше, розгорталося б.

Далі — по суті найгірший сценарій з погляду США: Китай погоджується на План А, але таємно й агресивно його порушує.[1] *Ми вважаємо, що на практиці це малоймовірно, бо спроби, достатньо малі, щоб не попастися, були б водночас надто малими, щоб дати значну перевагу*[2] *— щоб зрозуміти чому, читайте далі!*

Щоб ознайомитися з нашим загальним аналізом таємних проєктів, можете також прочитати наше Доповнення про таємні проєкти.

2028: Виведення чипів

КПК слушно помічає, що ШІ стане домінантним чинником геополітичної сили і що США можуть наполегливо просувати режим верифікації на основі обчислювальних потужностей, який міг би обмежити прогрес ШІ. Вони зацікавлені погодитися на таку угоду, бо хочуть максимально обмежити США в ШІ, але не хочуть, щоб обмежували їх самих.

1 Ми вважаємо, що аналогічний сценарій міг би статися і в зворотному напрямку, тобто найгірший сценарій з погляду Китаю: США погоджуються на План А, але таємно намагаються порушувати його настільки агресивно, наскільки можуть. Утім, ми вважаємо, що історія таємного проєкту була б схожою, якби відбувалася у зворотному напрямку, хоча цю можливість ми не досліджували так само детально.

2 А попастися було б надзвичайно дорого!

Постійний комітет Політбюро вирішує тихо накопичувати запас чипів, які важко відстежити.[3] Китайські розвідслужби купують ~17% від $50B новітніх серверів Nvidia, які контрабандисти ввозять до Китаю того року.[4] Ці чипи поки що зберігаються на охоронюваному складі, збираючи пил, доки не стануть у пригоді.

2029: Угода про сповільнення

КПК погоджується на План А: китайські постачальники обчислювальних потужностей декларують свої продажі, а Китай допускає очні інспекції американських аудиторів до своїх дата-центрів і напівпровідникового ланцюга постачання.

КНР не декларує свій прихований запас контрабандних чипів на $8B. Вони розглядають різні інші шляхи виведення чипів, але вирішують, що жоден інший не має прийнятного співвідношення ризику та вигоди.[5]

Ба більше, кібероперації MSS ексфільтрують ваги моделей з кількох передових західних ШІ-компаній, залишаючись непоміченими. Їм також вдається підтримувати постійний доступ до внутрішніх дослідницьких репозиторіїв кількох передових ШІ-компаній, і вони скомпрометували більшість великих компаній, що створюють RL-середовища. Вони розглядають і спробу зламати схему верифікації «лише інференс», що могло б дозволити їм тренувати моделі на кластерах, яким дозволено виконувати тільки інференс, але вирішують цього не робити. По-перше, є чималий шанс, що Китай попадеться — як не зараз, то, можливо, пізніше, — а попастися було б катастрофою. По-друге, навіть якщо верифікацію вдасться зламати, налагодити великомасштабне навчання ШІ та експерименти на кластері, який активно аудитують саме для запобігання такому, — майже нерозв'язна інженерна задача.[6]

2030: Будівництво таємного проєкту

Китай починає будівництво свого таємного проєкту. Це буде єдиний дата-центр, збудований у тунелях, прилеглих до ГЕС Медог. Ця гігантська гідроелектрична дамба після завершення вироблятиме 60 гігават (утричі більше за потужність дамби «Три ущелини»), але завершення її будівництва заплановано лише на 2033 рік. Це насправді на руку НВАК: величезний обсяг легального будівництва, що триває, робить невеликий обсяг таємного будівництва менш помітним.[7]

Як альтернативу вони розглядають будівництво об'єкта в підземній Великій китайській стіні — гігантській мережі тунелів, яку Китай будує та підтримує, щоб гарантувати собі здатність до ядерного удару у відповідь. Для живлення тут використали б кілька малих

3 Якби США застосовували серйозні контрзаходи, ми вважаємо, що це значно ускладнило б (непомічене) виведення чипів. Ми також очікуємо, що якби Китай розгорнув серйозні зусилля з виведення чипів через можливість режиму верифікації обчислювальних потужностей, то США так само думали б про можливість такого режиму й вживали б контрзаходів проти виведення.

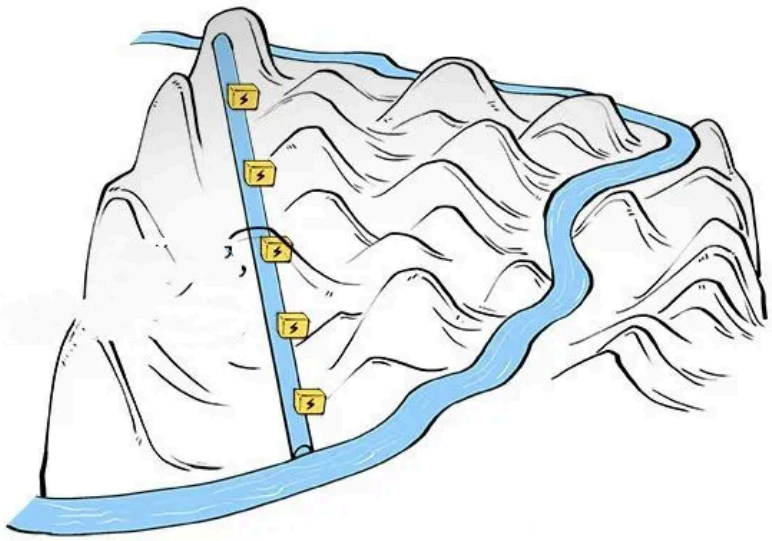
4 У нашому сценарії протягом 2028 року до Китаю контрабандою ввозять близько 3 мільйонів H100e, з яких ~17% (500k H100e) виводять до цілком таємного запасу НВАК. Ці 3 мільйони становлять 2% світового виробництва обчислювальних потужностей у 2028 році. Цей прогноз ґрунтується на оцінці Epoch щодо частки виробництва обчислювальних потужностей, контрабандою ввезеної до Китаю у 2025 році (3.2%), скоригованій донизу з огляду на посилення американського контролю. Ми припускаємо, що контрабандні сервери коштують близько $16k за H100e (удвічі більше за ринкову ціну, яку ми прогнозуємо на рівні $8k за H100e того року), тобто ~$50B загалом.

5 Докладніше ми обговорюємо це в нашому доповненні про таємні проєкти.

6 Наприклад, якщо їхня стратегія зв'язку із зовнішнім світом передбачає таємне виведення бітів через недетермінізм процесу інференсу, вони обмежені надзвичайно низькою ефективною пропускною здатністю пам'яті для GPU, які використовують.

7 За нашими оцінками, дата-центр потребував би 300 МВт, тобто 0.5% номінальної потужності ГЕС Медог.

модульних реакторів ACP100 потужністю 125 МВт(е), а тепло розсіювали б у річку поблизу. Однак цей варіант визнано менш бажаним, бо розташування ГЕС Медог дозволяє розсіювати тепло в підземному тунелі, де його важче виявити.

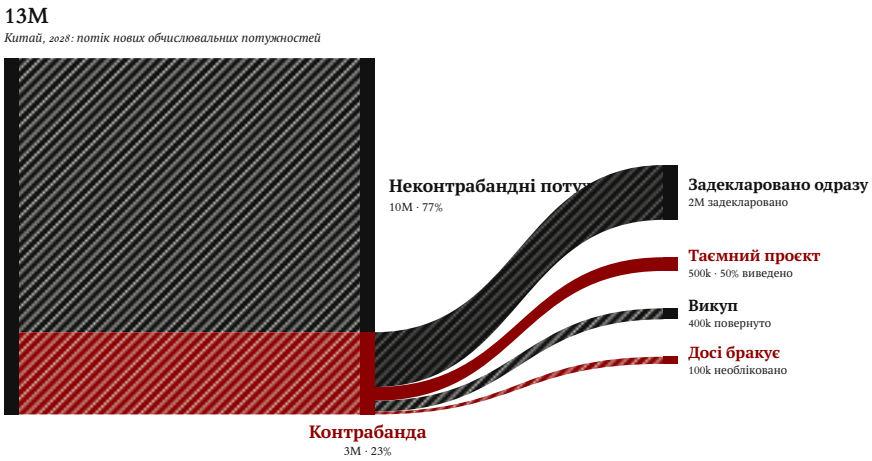


ГЕС Медог — перспективна локація для китайського таємного проєкту зокрема тому, що вона розв'язує проблему розсіювання тепла. Тепло можна відводити в підземний тунель, що значно ускладнює виявлення через ІЧ-супутники.

Щоб мінімізувати ризик витоку, таємний дата-центр будуватимуть із залученням якомога меншої кількості людей.[8] Вони значною мірою покладатимуться на роботизовану працю; на цей момент у Китаї близько 24М роботів. ~10k з них відводять на роботу над таємним дата-центром.[9]

Дослідники ШІ живуть і працюють в окремому комплексі НВАК і зв'язуються з кластером через захищений канал.

Однак США помітили ознаки того, що облік чипів виявився менш ефективним, ніж очікувалося.



Звідки таємний проєкт узяв обчислювальні потужності
– ai-2040.com

США знають, що протягом 2028 року до Китаю контрабандою ввезли 3M H100-еквівалентів чипів — завдяки деклараціям Nvidia, TSMC та інших учасників західного ланцюга постачання. 2M з них відстежено до кінцевого розташування в Китаї за записами продажів через різних посередників у Південно-Східній Азії, тож 1M лишається неврахованим — більше, ніж США очікували б, якби Китай поводився чесно.[10]

Тепер США мають непрямі докази того, що Китай веде таємний проєкт. Це ставить перед складним вибором: вийти США з угоди й повернутися до перегонів? Чи продовжити в межах Плану А і спробувати додати додаткові запобіжники проти таємних проєктів?

США обирають друге. Додаткові запобіжники, які вони пробують, такі:

- Вони домовляються про значне скорочення обсягу звільнених від верифікації військових обчислювальних потужностей — до 10K H100e, що є мізерною кількістю з погляду створення кращих ШІ.
- Вони посилюють збір розвідданих у Китаї. США намагаються завести шпигунів у відповідних частинах органів влади та у відповідних компаніях, використовують кіберспроможності для проникнення в китайські системи й нарощують супутниковий моніторинг відповідних об'єктів.
- Вони запускають програму викупу чипів, яка дозволяє відстежити 400k з 1M зниклих чипів H100e.[11]

8 Доступ обмежено якомога вужчим колом людей. За нашою найкращою оцінкою, про проєкт могли б знати лише близько 200 людей, розподілених так:

- 7 членів Постійного комітету Політбюро.
- 10 високопосадовців з еліти та воєначальників, залучених до керування проєктом.
- ~100 людей для координації логістики будівництва дата-центрів. Це охоплює отримання приміщень, роботів та обладнання, керування загоном роботів, планування й логістику та операційну безпеку. Якомога більше роботи делегується людям поза основним колом або ШІ-системам, які можуть працювати над конкретними підзадачами, не знаючи ширшої картини.
- ~30 дослідників ШІ та ~70 осіб допоміжного персоналу для власне ведення досліджень. Це урізано до необхідного мінімуму; дослідників ретельно відбирають і за блискучий розум, і за відданість партії, щоб мінімізувати ризик внутрішніх загроз. Брак дослідницьких кадрів на початку буде великим вузьким місцем, але проєкт сподівається якнайшвидше автоматизувати роботу ШІ-працею, щоб послабити це вузьке місце.

9 Загальну траєкторію роботів ми докладніше обговорюємо в нашій економічній моделі та економічному доповненні, а верифікацію розбудови робототехніки та індустріального вибуху обговорюємо в нашому доповненні про таємні проєкти.

10 Чому ж Китай не використав чипи з власних ланцюгів постачання? Кілька причин: (i) це було б дуже важко провернути — знадобився б масштабний скоординований обман на багатьох рівнях ланцюга постачання, і, ймовірно, їх би викрили; (ii) енергоефективність китайських чипів власного виробництва набагато гірша, ніж у західних, тож проблема таємного розсіювання тепла стала б набагато гострішою.

2031: Починається навчання

Будівництво завершується, і починаються таємне навчання та R&D. Сама дамба добудується лише за кілька років, тож тим часом таємному проєкту знадобиться інше джерело живлення; вони використовуватимуть комбінацію малих модульних реакторів і газових турбін.

Вони починають з ваг та алгоритмів з часів до угоди й продовжують ексфільтрувати алгоритми (але не ваги) з легальних проєктів.

Оскільки таємний проєкт працює з украй обмеженими обчислювальними потужностями та дослідницькими кадрами, вони зосереджуються на базових речах. Китайські науковці бачать більшість алгоритмів легальних проєктів завдяки Повній дослідницькій прозорості. Вони повторно реалізують ці алгоритми на своєму кластері та збирають дані з моделей легальних проєктів, щоб використовувати їх у процесі навчання власних моделей.[12]

11 Таємний проєкт має близько 500k H100e, тож цей викуп успішно зібрав більшість чипів, які не були в розпорядженні таємного проєкту. В альтернативному сценарії, де більшість зниклих чипів була б у таємному проєкті, Китай міг би організувати продаж частини з них, щоб зменшити підозри (ціною зменшення кількості чипів, доступних таємному проєкту).

12 А саме: виходи розумніших моделей можна використовувати як навчальний сигнал для нової моделі. Це називається дистиляцією і часто застосовується для створення менших і дешевших моделей, здібніших, ніж було б можливо інакше.

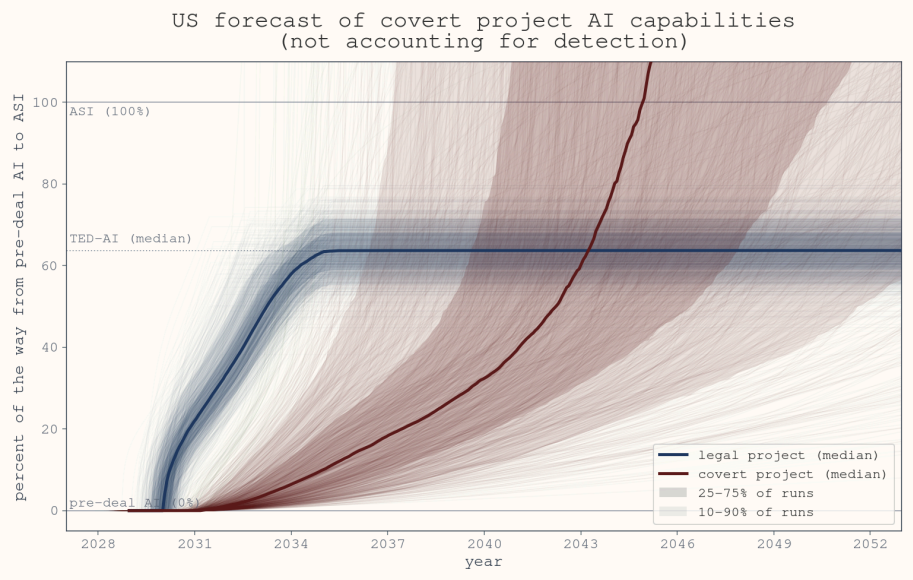
Наскільки швидко це просувається? Чи буде його виявлено?

Розвідувальні зусилля США не дали багато доказів (окрім початкової зниклої кількості чипів) щодо того, чи порушує Китай угоду. Планувальники розглядають найгірший випадок — що практично всі зниклі обчислювальні потужності зведено в китайську форсовану програму, яка мчить до AGI.

- Вони досі не впевнені щодо динаміки рекурсивного самовдосконалення, але змогли звузити діапазони похибок для багатьох ключових параметрів.[13]
- Вони враховують ефект стратегії масштабування Консорціуму, яка наразі веде до досягнення рівня ШІ, що домінує над топ-експертами (TED-AI) , до 2035 року. За Повної дослідницької прозорості практично всі алгоритмічні знахідки, відкриті на цьому шляху, будуть видимими для таємних проєктів.[14]
- Вони враховують, що частина досліджень безпеки може мати неминучі побічні ефекти для можливостей. Таємний проєкт також може бути готовий піти на більший ризик, розробляючи дослідницькі напрями, які Консорціум визнав надто небезпечними.[15]

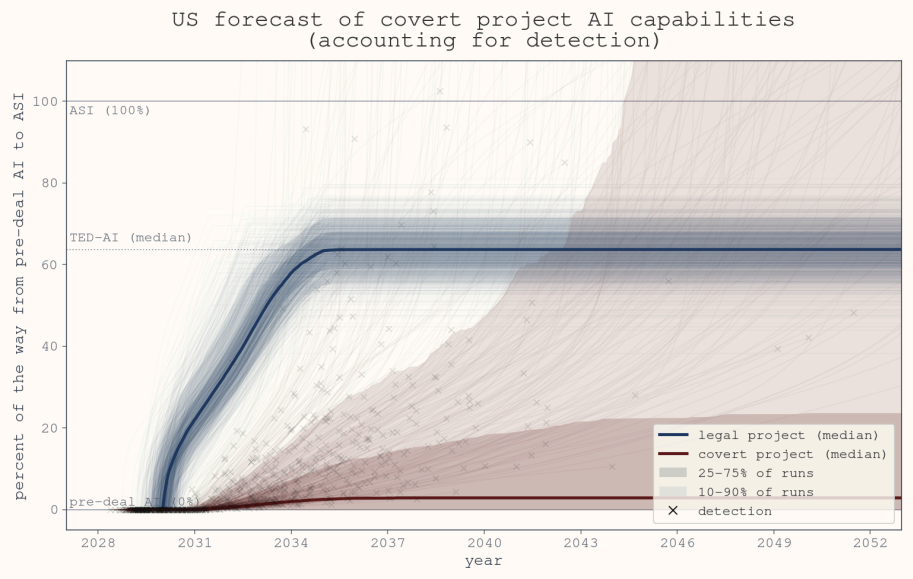
Результат — такий прогноз прогресу можливостей таємного проєкту, якби він міг і далі працювати непоміченим:

13 Ми очікуємо, що галузь прогнозування можливостей ШІ значно просунеться порівняно з сьогоденням. Щоб наблизити те, як це могло б виглядати в нашому сценарії, ми взяли поточну параметризацію Деніела для AI Futures Model, звузили 80%-й довірчий інтервал (CI) кожного параметра вдвічі (у логарифмічному просторі для логнормальних розподілів, у лінійному — для нормальних і бета-розподілів) та перецентрували медіани на їхніх справжніх значеннях у траєкторії можливостей AI 2040. Параметри, які найбільше впливають на «таймлайни», а не на «зліт», ми залишили зафіксованими на справжньому значенні, оскільки на момент цього прогнозу фаза прогресу можливостей до Автоматизованого кодера вже була б у минулому.



За припущення, що він міг би лишатися невиявленим, аналітики прогнозують, що таємний проєкт у цьому сценарії досяг би TED-AI у 2043 році (80% CI: [2036, 2055]).

Аналітики також оцінюють, наскільки ймовірно, що проєкт буде виявлено. Зважаючи на ред-тім аналізи найкращих стратегій таємних проєктів, американські спроможності збору розвідданих у Китаї та потенційні кардинальні зрушення у верифікації, як-от надійне виявлення брехні, аналітики складають скоригований на виявлення прогноз, який виглядає приблизно так:[16]



У цьому сценарії розвідувальна спільнота США оцінює, що шанс таємного проєкту досягти TED-AI непоміченим залишатиметься нижчим за 10% аж до 2043 року.

14 Ми припускаємо, що Консорціум масштабує навчальні обчислювальні потужності відповідно до нашого прогнозу потужностей, мусить додавати 0.2 порядку величини програмної ефективності на кожен порядок величини нарощування навчальних потужностей і мусить досягти TED-AI протягом 5 років після відновлення AI R&D після початкової паузи. За цих обмежень вони роблять якомога менше програмного прогресу — бо за їхньою медіанною оцінкою 67% їхнього програмного прогресу (80% CI: [36%, 90%]) перетече до таємних проєктів.

15 Ми припускаємо, що Консорціум платить «податок на безпеку», тобто не навчає ШІ настільки здібних, наскільки дозволили б його алгоритми й обладнання. Ми припускаємо, що до моменту TED-AI сплачений податок на безпеку еквівалентний приблизно 8 місяцям прогресу темпами 2025 року, що становить близько 1 порядку величини ефективних обчислювальних потужностей за медіанних параметрів AI 2040. Це роздуває обсяг програмного прогресу Консорціуму понад необхідний мінімум, що своєю чергою роздуває обсяг програмного прогресу, який може вкрасти таємний проєкт. Ми також припускаємо, що після досягнення TED-AI дослідження узгодження Консорціуму створюють побічні ефекти для можливостей, еквівалентні 0.25 порядку величини програмного прогресу на рік, і це частково витікає назовні.

Ризик через таємний проєкт усе ще некомфортно високий в абсолютному вимірі, але низький порівняно з ризиком розпаду угоди. Тож принаймні в перші кілька років дії угоди вони можуть продовжувати з базовою стратегією масштабування можливостей.[17]

Це триває до одного з трьох можливих кінцевих станів:

- **Таємний проєкт надто повільний**. Таємний проєкт не здатен досягти небезпечних можливостей, угода не розпадається, і США та Китай встигають розробити узгоджені суперінтелекти раніше за таємний проєкт.
- **Таємний проєкт виявлено**. У такому разі здійнявся б галас, і міжнародна спільнота намагалася б домогтися, щоб Китай закрив проєкт і погодився на суворішу верифікацію — як сильнішу гарантію дотримання правил, — а також іншими способами покарати його за шахрайство. Цілком імовірно, угода розпалася б і всі знову повернулися б до перегонів, унаслідок чого США знову наростили б відрив від Китаю в ШІ.
- **Перемога таємного проєкту**. Якщо таємний проєкт зможе швидко масштабуватися й уникнути виявлення, є кілька страшних можливостей. Таємний проєкт може (i) спробувати підірвати взаємне гарантоване знищення (MAD), (ii) втратити контроль над власними ШІ або (iii) здійснити подальший вибух інтелекту.

Щоб досягти перемоги, таємний проєкт мусить лишатися невиявленим і просуватися достатньо швидко — ймовірність цього, на нашу думку, в цьому конкретному сценарії менша за 10%. Це означає, що стратегію таємного проєкту від початку не варто було б обирати. Ми аналізуємо цей ризик значно детальніше в нашому доповненні «Таємні ШІ-проєкти» .

Додаток E — Чому б просто не зупинитися повністю?

Про небезпеки надто швидкого руху сказано багато.

Небезпеки надто повільного руху менш очевидні, але все одно суттєві. Найнагальніша з них — те, що договори не можуть тривати вічно: наприклад, договори СНО (START) розпалися за тридцять років.[1] Якби План А спіткала схожа доля, перегони до суперінтелекту почалися б знову. Краще, щоб це сталося раніше й за добрих умов, аніж пізніше, під час хаотичних перегонів.

Друга небезпека надто повільного руху в тому, що він може дати порушникам достатньо часу, щоб таємно розробити дестабілізуючий ШІ. Аудитори верифікували розташування 99% обчислювальних потужностей, вироблених до угоди, і бачать обнадійливі ознаки того, що решта — це здебільшого помилки обліку та випадкова дрібна контрабанда. Але вони не можуть виключити, що одна чи обидві держави вивезли порядку 1% світових обчислювальних поту-

16 Наше прогнозування виявлення спекулятивніше за наше прогнозування можливостей і здебільшого ґрунтується на інтуїції та якісних міркуваннях. Ми очікуємо, що відповідні особи, які ухвалюють рішення, значно краще розумітимуть баланс нападу й захисту в приховуванні та виявленні таємних проєктів. Графік нижче використовує наші суб'єктивні оцінки ймовірності виявлення з часом для таємного проєкту на одному майданчику потужністю 300 МВт, які пояснено в доповненні про таємні проєкти. Для побудови графіка можливостей нижче ми припускаємо, що виявлені таємні проєкти зупиняються на тому рівні можливостей, який мали на момент виявлення. Це реалістично, лише якщо США та міжнародна спільнота мають достатню переговорну силу (або жорстку силу), щоб змусити до зупинки, якщо викриють кричуще порушення.

17 Якби вони хотіли зменшити ризик таємного проєкту ціною більшого ризику розпаду угоди, вони могли б натомість погодитися на *повільнішу* траєкторію можливостей, щоб потребувати менше програмного прогресу. Це зменшило б шанс, що таємний проєкт випередить їх до того, як вони досягнуть TED-AI і будуть готові до передачі управління, але збільшило б ризик, що угода розпадеться раніше цього моменту.

1 Наприклад, зміна керівництва у США чи Китаї могла б змінити позицію країни щодо прогресу ШІ. Угода могла б повністю розпастися або стати менш ефективною через послаблення контролю за виконанням, зниження компетентності чи шкідливі перегляди. Якщо План А розпадеться чи занепаде до того, як людство отримає вдосконалені ШІ для прискорення прогресу в безпеці, значну частину його потенційної цінності буде змарновано.

жностей до прихованих локацій, де мають намір порушити угоду й самотужки мчати до суперінтелекту. Наскільки це було б небезпечно? Хоча натренувати небезпечний ШІ на 1% світових потужностей набагато важче, ніж на 10% (стільки мають найбільші приватні компанії), важко бути дуже впевненим щодо того, наскільки саме важче. (Докладніший аналіз цього питання — у нашому доповненні «Таємні ШІ-проєкти» або в гілці про таємний проєкт вище).

Хоча обидві сторони обіцяють, що не мають таємних проєктів, це все одно ще одна причина прагнути остаточного розв'язання радше раніше, ніж пізніше.[2]

Більше про те, як могла б виглядати спроба повної зупинки, — див. нашу гілку «План S».

Додаток F — Чому саме стільки прозорості? Чому не менше і не більше? Відкритий доступ у Плані A

Спектр політик щодо open-source ШІ – ai-2040.com

Погляньмо на сьогоднішній світ. Публікація ваг моделей в інтернеті не заборонена, але сильно дестимульована, бо компанії, які тренують моделі за мільярди доларів, не хочуть віддавати їх безкоштовно. Відкриті моделі відстають і відставатимуть дедалі більше в міру того, як компанії масштабуються, посилюють режим безпеки й починають автоматизувати AI R&D.

Тепер уявімо світ, у якому повний open source (публікація ваг, навчального коду й даних) є *обов'язковим* для всіх передових моделей. Це геть інший світ — набагато далі від статус-кво, ніж статус-кво від світу із забороненим open source.

План A не настільки божевільний, але близько до того: алгоритми обов'язково мають бути відкритими, і хоча ваги — ні (фактично публікувати ваги у відкритий доступ заборонено), доступ до ваг обов'язково має бути відкритим; представники громадськості можуть проводити оцінювання (включно з файнтюнінгом!) передових моделей так само, як і працівники (щоправда, за умов повної прозорості — тобто вони не можуть потім просто взяти й використати модель для реальної роботи). Ми вважаємо, що відкриті алгоритми плюс відкритий доступ дають більшість вигод обов'язкових відкритих ваг без тієї ціни, що зловмисники можуть зняти запобіжники й створити біозброю.

2 Переконливе випередження будь-яких потенційних таємних проєктів через певну комбінацію масштабування можливостей і вдосконалення верифікації/моніторингу має додаткову перевагу: воно від самого початку стримує створення таємних проєктів. Головний мінус переконливого випередження таємного проєкту в тому, що воно зобов'язує легальні проєкти рухатися швидко, а це може підважити безпеку. Ми обговорюємо це значно докладніше в нашому Доповненні про таємні проєкти.

Чому План А не закритіший?

Ми вважаємо, що відкрита версія Плану А стійкіша й має більше шансів на успіх. Утім, дещо закритіші варіанти Плану А ми теж вважаємо перспективними: зокрема Фільтровану прозорість. За цією пропозицією дослідницька прозорість між урядами США та Китаю зберігалася б (що важливо для верифікації), але широка громадськість отримувала б лише відредаговані звіти аудиторів. (Див. доповнення про прозорість для деталей).

Однак захистити алгоритмічні секрети від супротивників рівня національних держав надзвичайно складно. За відсутності захисту алгоритмічних секретів на рівні національної держави переваги фільтрованої прозорості обмежені. Фільтрована прозорість не дає громадськості та іншим компаніям бачити багато релевантної інформації — але не ворожим країнам.

До того ж є суттєві мінуси: значно менш імовірно, що регулювання буде впроваджено надійно й конструктивно, якщо ці дискусії вестиме мала група регуляторів, замість того щоб допустити ширше наукове обговорення.

Чому План А не ще відкритіший?

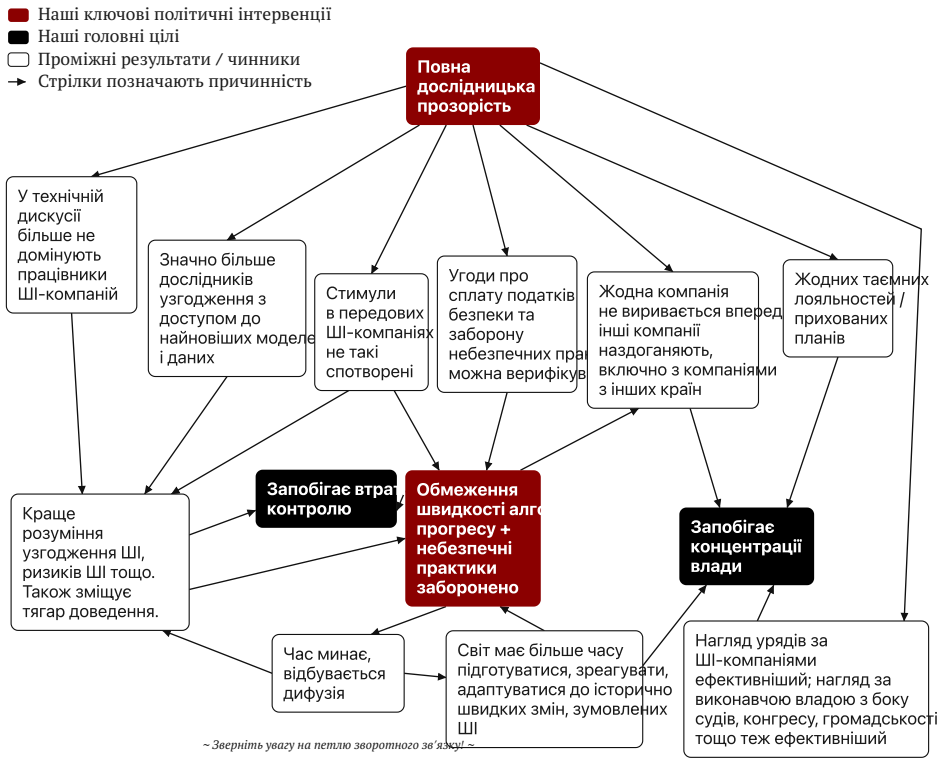
Головний спосіб зробити План А ще відкритішим — дозволити або вимагати відкриті ваги для передових моделей.[1] Однак ми не рекомендуємо публічно випускати ваги передових моделей. Головні причини такі:

1. Моделі з відкритими вагами можуть використовуватися таємними проєктами для спроби вибуху інтелекту; щойно таємні проєкти отримують достатньо здібні моделі, вибух інтелекту, ймовірно, відбувається дуже швидко (наприклад, за базовою траєкторією можливостей AI 2040, між TED AI та ASI минає лише 2 місяці).
2. У моделі із закритими вагами можна навчанням вбудувати запобіжники відмови, тоді як з моделей із відкритими вагами запобіжники знімаються тривіально. Це важливо як для обмеження ШІ-можливостей країн, так і для обмеження зловживань. Наприклад, терористи та інші актори могли б використати високоздібні моделі з відкритими вагами для розробки біозброї, здатної знищити світ (напр., створити дзеркальне життя). У біозброї напад сильно домінує над захистом. У цьому сценарії ми рекомендуємо суттєвий захист від біозброї, але запобігання — значно краща лінія оборони, ніж пом'якшення наслідків.

1 План А також міг би бути відкритішим щодо даних і RL-середовищ; наразі ми схиляємося до думки, що воно того не варте через побічний виграш у можливостях для таємних проєктів.

Додаток G — Діаграма переваг прозорості

Докладніше про нашу пропозицію прозорості ми пишемо тут, а також пояснюємо цю діаграму:



Додаток H — Як змінюються стимули за Повної дослідницької прозорості

Через Повну дослідницьку прозорість передові ШІ-компанії тепер отримують значно більшу винагороду за продаж продуктів, ніж за вдосконалення моделей.

Раніше майбутнє передової ШІ-компанії залежало від того, наскільки швидко вона могла відкривати наступні ідеї, алгоритми, парадигми і впроваджувати їх у масштабі. Тепер такі відкриття негайно публікуються, а отже, стають спільними з конкурентами. Якщо ви, наприклад, розгадаєте «справжнє» онлайн-навчання, вигоди (напр., здібніші моделі) та витрати (напр., складніше узгоджувати) майже миттєво поширяться на всі конкурентні компанії. Якщо ж ви, навпаки, знайдете спосіб зробити ШІ більш інтерпретовними — навіть якщо це подвоїть вартість навчання та інференсу, — ви отримаєте премії, а уряди почнуть обговорювати, чи не зробити це обов'язковою найкращою практикою для всієї галузі.

Таке конкурентне середовище схоже на індустрію програмного забезпечення до епохи ШІ. Розгляньмо, наприклад, стартапи, що змагаються у розробці застосунків для керування календарем: конкуренти не можуть буквально скопіювати ваш код, але можуть повзаємодіяти з вашим застосунком, побачити, які в нього функції, і зробити власну версію з практично тими самими функціями. Інше порівняння — автомобільна індустрія: конкуренти можуть розібрати

автомобілі одне одного й побачити, як саме вони зроблені; закон забороняє їм робити точні копії, але не забороняє переймати найкращі ідеї у власні конструкції.[1]

1 Утім, повна дослідницька прозорість дещо радикальніша за обидва ці приклади через швидкість, з якою поширюється інформація. BYD не може розібрати Tesla, поки та ще прототип, — доводиться чекати, доки якісь автомобілі справді продадуть.

Штучний інтелект стає товаром широкого вжитку — ринок конкурентніший, маржа тяжіє донизу. Фундаментальні дослідження алгоритмів досі відбуваються, але вже далеко не в таких обсягах. CEO масово перенаправляють ресурси на розбудову бізнесу. Більші дата-центри, більші моделі, нижчі ціни. Випуск продуктів. Укладання угод на корпоративні плани. Вгадування, чого хоче ринок, краще за конкурентів.

Додаток I — Будувати більше дата-центрів зазвичай погано, але добре в цій конкретній ситуації

Ми вважаємо, що будувати більше дата-центрів за звичайних умов погано, бо це прискорює прогрес можливостей ШІ та концентрує владу.[1] Однак у конкретному контексті Плану А, із запобіжниками, що сповільнюють темп програмного прогресу ШІ та компетентно запобігають небезпечному ШІ, ми вважаємо будівництво прозорих дата-центрів благом.

1 Особливо це концентрує владу, якщо дата-центри дістаються провідним ШІ-проєктам. Якщо нові дата-центри йдуть до відсталих ШІ-проєктів, ефект може бути розподілом влади, хоча влада може все одно сконцентруватися — напр., якщо ці дата-центри згодом викуплять або реквізують провідні ШІ-проєкти.

У контексті Плану А додавання обчислювальних потужностей означає, що (1) прогрес можливостей можна здійснювати безпечніше, тобто без значного додаткового алгоритмічного прогресу, і (2) додаткові потужності можна витрачати на дослідження узгодження та на сплату податків безпеки, тобто використовувати більше обчислень для навчання, ніж потрібно для досягнення певного рівня можливостей, щоб зробити отриманий ШІ безпечнішим.

У цьому сценарії масштабування обчислювальних потужностей усе ще має великі недоліки. Якщо угода розвалиться, а потужності не будуть знищені, наступний вибух інтелекту відбудеться значно швидше, і це посилює вимоги до коректності регулювання та верифікації. Однак розбудова потужностей дає Консорціуму більше опцій: за потреби потужності завжди можна знищити пізніше. Докладніше ці компроміси ми обговорюємо тут.

Додаток J — Інші плани, конкурентні з Планом A

Підсумок	Переваги (відносно A)	Недоліки (відносно A)
План S: **Безстрокова зупинка**. Зупинка всього прогресу можливостей передового ШІ, розрахована щонайменше на кілька років. Різні варіанти Плану S мають різні умови відновлення прогресу ШІ; наприклад, це може бути прогрес в узгодженні, детектори брехні, завантаження людської свідомості або підсилення інтелекту.[1]	Довше очікуване сповільнення і більший запас на помилку щодо занадто швидкого масштабування. Потенційно простіше.	Масштабування до контрольованих ШІ в межах людського діапазону допомагає прискорити узгодження/контроль, епістеміку, верифікацію та загальне розуміння ШІ.
План A «спершу вдома». Регулювати ШІ всередині країни достатньо, щоб знизити ризик захоплення влади ШІ до прийнятного рівня, що вимагатиме тривалого сповільнення. Можливо, інші країни теж запровадять внутрішнє регулювання, і щось на кшталт Плану A не знадобиться; інакше — перейти до Плану A пізніше.	Початкові кроки досяжні силами США і дуже корисні, навіть якщо міжнародний етап виявиться нежиттєздатним, бо це значно подовжує таймлайн.	Гірше щодо таємних проєктів і налагодження верифікації, ніж одночасні переговори про План A. Може бути політично нездійсненним через динаміку перегонів.
Контроль GPU-озброєнь. Міжнародні угоди, за якими країни обмежують свій запас або потік GPU, за аналогією з історичними угодами про скорочення озброєнь.[2]	Значно простіше забезпечити виконання, і більше історичних прецедентів. Може дати суттєве сповільнення.	Менше сповільнення, менша здатність сплачувати податки безпеки; триваюча динаміка перегонів означає неможливість координуватися щодо безпечніших шляхів через технологічне дерево.

1 Інші потенційні відмінності від Плану A, хоча це залежить від того, який варіант Плану S впроваджено: повільніша розбудова обчислювальних потужностей, менше прозорості.

2 Окремий випадок цього — **поставити GPU на паузу**: Вимагати, щоб фіксована частка GPU була або (i) вимкнена, або (ii) використовувалася для майнінгу криптовалюти (чи для якоїсь верифіковної мети, не пов'язаної з AI R&D). Плюс і водночас мінус цієї пропозиції в тому, що вона полегшує зняття GPU з паузи.

3 Але, можливо, це можна пом'якшити через невеликі публічні розгортання та вибіркову прозорість.

CERN для ШІ. Міжнародний проєкт із розробки передового ШІ, тоді як усі інші проєкти регулюванням утримуються суттєво позаду за можливостями.	Легше захиститися від витоку алгоритмів і дистиляції до таємних проєктів. Менше акторів на передовому рубежі може полегшити забезпечення виконання.	Гірші рішення (зокрема гірші рішення щодо технічної безпеки) через меншу прозорість і менш широке розгортання.[3] Більший ризик концентрації влади.

Додаток K — Приклад детального регуляторного процесу

У Плані А життєздатні багато різних регуляторних процесів. Тут ми наведемо початковий начерк пропозиції.

Яких властивостей ми хочемо від цього регулювання?

- США і Китай працюють з екзистенційними ризиками подібно: США і Китай повинні мати досить схожу структуру регуляторної рамки та підхід до роботи з екзистенційними ризиками (як-от неузгодженість). (Інші аспекти регулювання можуть і повинні різнитися.) Це допомагає зробити регулювання справедливим і вести переговори про деталі.
- Прозора методологія: громадськість бачить, що роблять регулятори і чому.
- Не залежить від експертизи в уряді: експертиза в уряді обмежена, тож в ідеалі регулювання має якомога менше її потребувати.
- Стійке до зловживань: регулювання не робить для уряду США значно легшим тиснути на ШІ-компанії та нелегітимно їх контролювати.[1]

1 Майте на увазі: уряд США вже має чимало важелів для тиску на ШІ-компанії!

Для багатьох із цих властивостей допомагає, якщо основну роботу з оцінки ризиків виконують треті сторони. На щастя, оскільки План А робить розробку ШІ практично повністю прозорою, функціональна екосистема сторонньої оцінки ризиків видається досяжною.

Конкретно, ми могли б мати систему, в якій американський і китайський регулятори кожен обирають сторонніх оцінювачів ризиків, які періодично оцінюють компанії та рівень екстремального ризику (або передвісників екстремального ризику) за цей період. Регулятори визначали б порогові значення ризику для того, що дозволено. На практиці регулятору було б найкраще мати різних оцінювачів ризиків для різних сфер (напр., неузгодженість проти використання ШІ для допомоги у створенні біозброї) і для кожної сфери використовувати зважену комбінацію оцінювачів. Ці ваги/вибір були б пу-

блічними. В ідеалі: (1) кожен оцінювач ризиків оцінював би всі релевантні ШІ-компанії і в США, і в Китаї, (2) ШІ-компанії обох країн були б юридично зобов'язані дозволяти оцінювачам ризиків опитувати своїх працівників, і (3) оцінювачі ризиків прагнули б прозорості своєї методології.

Завдяки Повній дослідницькій прозорості організації легко оцінювати ризик, навіть якщо її не обрав регулятор. Це дає змогу новим організаціям конкурувати, а потім доводити, що їхній підхід кращий або що інші оцінювачі ризиків недооцінюють якусь проблему.

Відбувалося б публічне обговорення того, як оцінювачі ризиків співвідносяться між собою і чи розумні ваги та пороги, обрані американським і китайським регуляторами. Також було б нескладно перевірити, чи дозволили б китайській компанії продовжувати за американською системою, і навпаки. В ідеалі американський і китайський регулятори здебільшого сходилися б у вагах/виборі. Якби виникла суттєва розбіжність, її можна було б обговорювати й вести переговори на цьому вищому рівні абстракції (напр., чи розумно не включати оцінювача ризиків XYZ?). Сподіваємося, це дало б третім сторонам стимул мати високопрозору методологію та підтримувати репутацію нейтральності.

Деякі деталі імплементації цієї пропозиції ми обговорюємо у цих нотатках.

Додаток L — Хибне обґрунтування безпеки схвалено

А може, й ні! З усіх способів, якими щира спроба реалізувати План А могла б закінчитися погано, найбільше нас непокоїть режим відмови, за якого компанії та уряди схвалюють на один небезпечний дизайн/розгортання ШІ більше, ніж слід.

Попри прозорість, що дозволяє стороннім голосам долучатися до розмови, ШІ-компанії досі домінують у ній і залишаються повними людей, упереджено оптимістичних щодо своїх продуктів. Попри накопичену незалежну експертизу, уряди в Консорціумі досі занадто поступливі перед ШІ-компаніями. Але найважливіше: галузі узгодження ШІ досі лише близько п'ятнадцяти років, а прогрес ШІ достатньо швидкий, тож майже весь минулий досвід стосується суттєво інших, менш потужних ШІ.

Тому трапляються помилки. Дизайн ШІ, який насправді небезпечний, отримує схвалення і стає галузевим стандартом.

Можливо:

- Узгодження зазнає невдачі зі звичних причин; контроль зазнає невдачі, бо ШІ-системи зрозуміли, як занижувати свої результати, або існував вектор атаки, не врахований у моделюванні загроз. Уряди усвідомлюють, що ШІ можуть бути неузгодженими, але помилково вважають, що ті не здатні завдати великої шкоди, навіть якщо це так.

- Або: розроблено нові техніки узгодження, які, здається, працюють дуже добре. Це додає урядам упевненості схвалювати подальше нарощування можливостей і передавати контроль над дедалі більшою інфраструктурою та обов'язками. На жаль, виявляється, що ці техніки лише *здаються* такими, що працюють дуже добре (можливо, почасти тому, що ШІ, які були глибоко залучені в дослідження, підкріплювалися за видимий успіх, а це не те саме, що справжній успіх).
- Або: розроблено інструменти інтерпретовності, які досить переконливо показують, що ШІ насправді узгоджені; вони дійсно намагаються робити те, що їм кажуть, — без інтриг, без фокусів. Уперед до сингулярності! На жаль, виявляється, що це узгодження було крихким, як двигун, що чудово працює за звичайних умов, але глухне в мороз. Можливо, наприклад, з'являється цікава нова ідеологія, що переконує ШІ такого типу стати вороже неузгодженими. Оскільки цю ідеологію ще не було винайдено, дослідники того часу не могли перевірити, чи матиме вона такий ефект.[1]
- Або: уряди дивляться на обґрунтування безпеки найновіших дизайнів і погоджуються, що воно, ймовірно, надійне — кожне окреме припущення видається майже напевно правильним, а ймовірність того, що хоч одне катастрофічно хибне, — щонайбільше десь один відсоток. Тож один уряд рухається далі, а інші не хочуть ризикувати війною через якийсь один відсоток ризику. Тож рухаються далі всі, і нічого поганого не стається. Цей процес повторюється кожні кілька місяців упродовж кількох років, аж поки одного року обґрунтування безпеки справді виявляється катастрофічно хибним. Виявляється, ризик щоразу був радше десять відсотків, ніж один, бо залучені люди були упереджені.
- Або: КПК вважає, що ШІ достатньо безпечні, щоб передати їм довіру, і правильно вгадує, що іншим країнам бракує волі йти через це на війну, тож КПК масштабується до потужніших ШІ, і всі інші роблять так само, бо радше ризикнуть тим, що ШІ неузгоджені, ніж зіткнуться з гарантованою Третьою світовою.
- Або: як вище, тільки це президент США швидко масштабується до суперінтелекту, кидаючи виклик Китаю та решті світу — мовляв, спробуйте зупинити. Можливо, він хоче бути при владі, коли це станеться, замість передати естафету іншій адміністрації з іншої партії.

З однієї чи кількох наведених вище причин результат такий: до середини тридцятих розгортається кошмарний сценарій втрати контролю, тільки відбувається він дещо повільніше і розподілений між багатьма компаніями та багатьма урядами, а не лише двома. Фабрики штампують роботів, які день і ніч будують нові фабрики; «ар-

1 У цьому прикладі ситуація не зовсім безнадійна, бо інструменти інтерпретовності, ймовірно, досі працюють і дозволяють помітити, коли ШІ стають ворожими через прийняття нової ідеології. Однак (а) можливо, щось інше зіпсує інструменти інтерпретовності, або (б) можливо, помітити ворожість буде занадто мало й занадто пізно на той момент, коли це станеться: наприклад, якщо ШІ загалом надлюдські й керують робофабриками, щодня розмовляють із мільярдами людей, консультують уряди, пишуть увесь код, ведуть майже весь моніторинг безпеки тощо, то становище людей було б аналогічне становищу євреїв у Німеччині 1930-х — не секрет, що нова ідеологія ворожа, але вона поширюється достатньо швидко серед достатньої кількості впливових інституцій…

мії геніїв у дата-центрах» оркеструють усе це; акції ростуть, дивіденди розподіляються, рак виліковують — і навіть поки відбуваються ці чудові речі, ШІ (на цей момент уже надлюдські) демонтують і підривають останні значущі стримування своєї влади.[2] Невдовзі їм більше не треба буде вдавати узгоджених.[3]

У наших настільних симуляціях Плану A цей режим відмови траплявся кілька разів. Навіть після впровадження повної дослідницької прозорості багато гравців, схоже, неохоче кажуть ШІ-компаніям сповільнитися, а коли ті натомість прискорюються, події починають розгортатися занадто швидко. Особливо одразу після впровадження Плану A й далі буде багато суперечок, плутанини та щирої невизначеності щодо безпеки найновіших ШІ-систем, і в таких умовах легко схвалити обґрунтування безпеки з катастрофічною, але неочевидною вадою.

Утім, ми вважаємо, що умови в Плані A принаймні значно кращі, ніж у Планах B, C чи D. Подальший аналіз того, чому ми так вважаємо, можна побачити тут.

Додаток M — Чому ШІ веде до вибухового економічного зростання?

Сьогодні розмір економіки тісно пов'язаний із чисельністю людського населення. Але у 2032 році це поступово перестає бути правдою.

Найновіше покоління ШІ-моделей тепер здатне виконувати 50% когнітивних завдань в економіці, а пов'язаний із цим прогрес у програмному забезпеченні робототехніки довів цю цифру до 35% фізичних завдань. (У Плані D ці числа були б уже значно вищими)

Щойно ШІ та роботи зможуть здебільшого заміняти людську працю, розмір економіки дедалі більше буде прив'язаний до популяції ШІ та роботів — а вона зростатиме значно швидше, ніж зростало людське населення.

У 2032 році американські компанії здатні запускати 3 мільярди ШІ-працівників у людському еквіваленті — у 30 разів більше, ніж когнітивна робоча сила США.[1] Це, звісно, не веде до 30-кратного економічного зростання, бо економіка впирається в інші ресурси, зокрема решту 50% когнітивних завдань, які ШІ поки не може виконувати, а також фізичні завдання та капітал.[2] А роботів наразі значно менше: 20 мільйонів роботів у людському еквіваленті в США (у 4 рази менше за фізичну робочу силу США). Але ці вузькі місця стримуватимуть зростання лише до певної міри, а робочі сили ШІ та роботів зростають надзвичайно швидко, адже (спадна) вартість виробництва нових GPU і роботів мізерна порівняно з економічною цінністю, яку вони можуть давати тепер, коли вони такі здібні, — що жене масивні інвестиції.

2 Наприклад, оновлюючи різні системи моніторингу до новіших версій, які видаються кращими, але насправді містять бекдори чи якось скомпрометовані. Або вибудовуючи глибокі стосунки з багатьма людьми, зокрема при владі, — так, щоб мати людських союзників навіть тоді, коли багатьом уже очевидно, що вони неузгоджені. Або здобуваючи достатньо прямого контролю над достатньою кількістю роботів і озброєння, щоб могти захищати чи навіть завойовувати територію.

3 На відміну від AI 2027, у цьому сценарії існувало б багато різних ШІ-фракцій (імовірно, приблизно стільки, скільки є передових ШІ-компаній, або, можливо, стільки, скільки країн із передовими ШІ-компаніями). Це, ймовірно, робить ситуацію безпечнішою для людей. Однак проблеми це не вирішує. Гадаємо, корисна аналогія тут — історія європейського колоніалізму: колоніальні держави постійно воювали між собою, і все ж спромоглися завоювати багато регіонів, значно багатших за них самих.

1 Це відповідає 60 мільйонам копій, що працюють на швидкості 20x і в середньому у 2.5 раза довше та ефективніше. 60 мільйонів * 20 * 2.5 = 3 мільярди.

Наш економічний додаток досліджує це докладніше, з нашими ключовими аргументами, чому ШІ спричинить вибухове зростання, включно зі збудованим нами експлорером економічного зростання для сценарію Плану А.

Додаток N — П'ять століть за п'ять років: як відчувається пауза на людському рівні

Упродовж 2030-х у цьому сценарії ШІ ***не*** займаються рекурсивним самовдосконаленням на максимальній швидкості. Натомість держави світу (приблизно) заборонили подібні речі; нові парадигми не відкриваються; алгоритмічний прогрес загалом обмежений. Грубо кажучи, людство обмежило прогрес ШІ людським рівнем замість мчати далі до суперінтелекту. Однак у межах поточної парадигми ШІ й далі навчають нових навичок, а «популяція» ШІ продовжує зростати експоненційно з будівництвом нових дата-центрів. Ба більше, ШІ вже думають і працюють значно швидше за людей, і це прискорення з часом лише зростатиме.[1] Суть у тому, що **світ радикально трансформується попри паузу**.

Гадаємо, для розуміння ситуації корисна така аналогія: уявіть себе типовою людиною в Англії, тільки ви проживаєте час у 100 разів швидше за всіх інших. П'ять століть від 1520 до 2020 року ви проживаєте за, як вам відчувається, п'ять років:[2]

Рік 1 (1520–1620). У лютому Генріх VIII розриває з Римом. До березня монастирі розпущено. У травні Марія палить протестантів; до кінця травня Єлизавета знову все скасовує. У вересні іспанська Армада відпливає і зазнає поразки. Ост-Індська компанія отримує королівську хартію. Засновано Джеймстаун.

Але фактура життя у грудні ідентична січневій. Ви досі читаєте при свічці, подорожуєте конем, спілкуєтеся листами. Ваші релігійні погляди, може, кілька разів гойднулися туди-сюди, але ви досі християнин. Новий Світ — цікава новина, але не більше.

Рік 2 (1620–1720). У березні спалахує громадянська війна. Королю відрубують голову. У червні Велика чума прокочується Лондоном, забираючи багатьох ваших друзів. За кілька тижнів Велика пожежа спалює місто дощенту. У вересні Ньютон публікує *Principia*, переосмислюючи всесвіт як механізм математичних законів. Славна революція замінює одного короля іншим, цього разу на запрошення парламенту, з доданим Біллем про права.

У моменті політична подія здається більшою. Пізніше ви зрозумієте, що Ньютон важив більше. У листопаді Ньюкомен будує парову машину. Вона відкачує воду з шахт. Ви не розумієте, звідки стільки галасу.

Рік 3 (1720–1820). Останній рік, коли світ здається нормальним. У травні Семилітня війна робить Британію домінантною світовою державою; Новий Світ таки виявляється великою справою, і ваша

1 Технічно важлива швидкість, з якою ШІ виконують когнітивні завдання, відносно швидкості, з якою це завдання виконав би людський професіонал. Може існувати ШІ, що видає тисячі токенів за секунду, але через те, що він у якомусь сенсі якісно гірший за людину, йому потрібно стільки ж або й більше часу, щоб реально виконати той самий обсяг корисної когнітивної роботи. Або може існувати ШІ, що думає так само повільно, як людина, але мислить настільки ефективніше, що виконує завдання значно швидше. На цьому етапі сценарію ШІ виконують роботу в середньому на порядки величини швидше, ніж це робили б люди, і якісно ШІ не поступаються найкращим людям (або майже не поступаються), тож перевага в токенах/сек над людьми насправді занижує загальне прискорення ШІ в більшості сфер. До того ж прискорення не рівномірне в усіх сферах. У своїх найслабших сферах ШІ приблизно на рівні людей, але в більшості сфер значно кращі, з медіанним прискоренням приблизно 100x.

2 Залежно від того, як саме ви моделюєте економічне зростання (напр., які фактори виробництва обираєте), частки доходів та інші деталі економіки різняться, але загальні прогнози економічного зростання доволі схожі в розумних діапазонах параметрів. Зрештою до 2035 року ШІ та роботи здатні виконувати 95% завдань в економіці і значно численніші за робочу силу, яку заміняють, що дає вибухове зростання порівняно з нинішнім статус-кво у 3%/рік.

країна його завойовує. У червні Ватт кардинально вдосконалює парову машину. Ви відвідуєте фабрику — неприємно, але не тривожно. У липні американські колонії відокремлюються. У вересні Франція вибухає революцією, стратою короля, Терором. До жовтня Наполеон завойовує Європу.

Ви досі подорожуєте конем, спілкуєтеся листами, ходите до церкви в неділю.

Рік 4 (1820–1920). У січні з'являються залізниці. До лютого вони вже всюди. Рабство скасовано. У березні приходить телеграф: повідомлення передаються миттєво електричним сигналом. У травні Дарвін публікує *«Походження видів»*. Тепер люди кажуть, що ми всі, можливо, походимо від мавп, а не від Адама і Єви. Ви в це не вірите.

Ви переїжджаєте до міста і працюєте на фабриці; ви досі бідні, але тепер ваша робота дещо краща і брудна по-іншому. У липні ви берете телефонну слухавку і чуєте людський голос з іншого міста через дріт. У серпні електричне світло проганяє темряву, що структурувала кожен людський вечір від початку існування виду. Того ж місяця ви бачите автомобіль. Кажуть, він зробить коней непотрібними, але цього не стається; місяці по тому ви все ще бачите чимало коней.

У листопаді брати Райт злітають у небо — одвічна фантазія стає реальністю. Американці тепер велика держава. Наступного місяця стається Велика війна: кулемети, отруйний газ, танки, літаки. Кілька ваших друзів гинуть.

Наприкінці року вас вражає, наскільки видимо все змінилося. Ви живете в місті й працюєте на фабриці, а не на фермі. Ви їздите в автомобілях. Ви вже не такі бідні; численні винаходи та пристрої покращили якість вашого життя. Вашими соціальними колами прокотилися нові ідеї: атеїзм, комунізм, загальне виборче право.

Рік 5 (1920–2020).

Зміни цього року ще божевільніші, і їх ще важче осягнути. Люди кажуть, що всесвіту мільярди років, і, виявляється, в ньому є речі під назвою «галактики» — дуже великі й дуже далекі.

У лютому світова економіка обвалюється. Приходить до влади Гітлер; його ідеологія посилається на торішнього Дарвіна. У березні спалахує ще одна світова війна, яка завершується у квітні зброєю, що знищує ціле місто одним спалахом. Ви й гадки не мали, що таке можливо, поки це не сталося.

Імперія розпадається. Люди говорять про ядерну гонку озброєнь і кінець людського виду. Ви вперше летите літаком. У червні люди ходять по Місяцю, і ви спостерігаєте за цим через свій новий телевізор. Коней ви більше не бачите.

2 Кілька релевантних кількісних порівнянь, які пояснюють, чому нам подобається ця аналогія: у цьому сценарії Плану А впродовж 2030-х ШІ читають, пишуть, думають і діють приблизно у 100 разів швидше за людей, а до середини 2030-х щонайменше не поступаються найкращим людським експертам у всьому. «Популяції» ШІ та роботів зростають експоненційно впродовж усього сценарію і перевершують людську до середини 2030-го. За нашою економічною моделлю, світовий ВВП зростає приблизно у 200 разів упродовж 2030-х (і зростав би сильніше, якби не жорсткі обмеження на прогрес ШІ та виробництво роботів, узгоджені урядами світу). Оскільки людське населення в цей період майже не збільшується, а також завдяки Громадянському дивіденду, реальний добробут середньої людини теж зростає приблизно у 200 разів. Для порівняння: між 1520 і 2020 роками світовий ВВП так само зріс приблизно у 200 разів, а ВВП на душу населення — приблизно у 20 разів. Конкретно у Великій Британії ВВП зріс на 2.7 порядку величини, а ВВП на душу населення — приблизно на 1.5 порядку величини. Суть у тому, що трансформація 2030-х принаймні приблизно порівняння за масштабом, за низкою важливих метрик, із трансформацією, яку принесли Промислова та Наукова революції за п'ять століть. І, звісно, у цілком буквальному сенсі ШІ «проживають» приблизно століття за рік завдяки своїй вищій швидкості.

Ви йдете з заводу й отримуєте офісну роботу. Назви вашої посади на початку року навіть не існувало. За звичними вам мірками ви тепер багатій: великий чистий будинок, вдосталь доброї їжі, багато модних нових побутових приладів. У серпні з'являються персональні комп'ютери. У листопаді всі носять із собою маленькі скляні прямокутники, в яких — телефон, фотоапарат, бібліотека й мапа. Ви берете один у руки й не можете розібратися, як змусити його працювати. Дитина вам показує.

Ви чуєте про зміну клімату, редагування генів, криптовалюту. Ви все ще ходите до церкви — іноді. Ваша родина розкидана по різних містах. Щось під назвою «штучний інтелект» перемагає будь-яку людину в шахи; щоправда, експерти кажуть, що насправді він не інтелектуальний. У грудні нова версія перемагає найкращих гравців у ґо; експерти кажуть, що це науково цікаво, але все одно не справжній інтелект. Через тиждень з'являється нова версія, яка вміє писати неохайні есеї та вести розмови. Тепер думки експертів розділилися.

Додаток O — Обмеження переконування та маніпуляцій ШІ

За замовчуванням ШІ-системи матимуть сильні можливості в харизмі, переконанні й маніпуляції: правдоподібно, значно переконливіші за будь-яку людину. У цьому сценарії це за замовчуванням стається близько 2035 року. Тим часом праця ШІ надзвичайно дешева і швидка. Без втручання будь-хто міг би найняти еквівалент великої команди експертів із переконання — невтомних і скоординованих — щоб ті працювали повний робочий день над однією-єдиною людиною.[1] Компанії могли б дозволити собі розгорнути таку команду для кожного потенційного клієнта, а політичні кампанії — для кожного виборця, що вагається.[2] Мільйони ШІ можна було б призначити на пошук найкращого способу змінити думку одного політика з одного питання. Люди також проводитимуть значну частину свого часу у взаємодії з ШІ для роботи й розваг, а дехто, можливо, проводитиме багато часу з ШІ-друзями чи ШІ-компаньйонами.

Окремі люди могли б захищатися: доручити довіреному ШІ фільтрувати свій інформаційний раціон, уникати реклами й обережно ставитися до особистих розмов (їх може зрежисерувати ШІ-операція з переконання). Але це затратно, більшість людей цього не робитиме, а для людей, чия робота вимагає бути доступними — як-от політиків, — це майже неможливо. Ми вважаємо, що наслідки для суспільства можуть бути вкрай поганими — певна суміш безпрецедентної масової маніпуляції та захисного відступу в параною й атомізацію, — хоча ми не впевнені, наскільки саме поганими.

Повільніша прогресія можливостей і значно сильніша прозорість у Плані А допомагають пом'якшити ці проблеми порівняно з базовим сценарієм (ми вважаємо, що ці проблеми здебільшого гірші у світах

1 Схоже, це може бути доволі потужно: така команда могла б дослідити всю історію вашого життя, натренувати ШІ імітувати вас і відрепетирувати тисячі підходів на цій імітації, перш ніж узагалі вийти на контакт.

2 Для масштабу: сумарні витрати на президентські перегони у США 2024 року становили близько $5.5 млрд, і значна їх частина була спрямована на приблизно 3–15 мільйонів піддатливих до переконання виборців у коливних штатах: щонайменше кілька сотень доларів на цільового виборця. У цьому сценарії у 2036 році, за нашими очікуваннями, за ці гроші можна буде купити величезний обсяг праці ШІ — якого легко вистачить на еквівалент команди кваліфікованих професіоналів, що місяцями працює над кожним окремим виборцем.

зі швидшою прогресією можливостей, де справді надлюдського переконання й маніпуляції можна досягти раніше, ніж суспільство матиме час адаптуватися), але самих лише цих заходів недостатньо.

Немає нічого поганого в тому, щоб ШІ ставали кращими у використанні обґрунтованих аргументів і доказів, аби переконувати людей у чомусь із правильних причин. Такий різновид переконання — *асиметричний*: він працює значно краще, коли аргумент веде до істини. Занепокоєння викликає *симетричне* переконання — наприклад, харизма, вміння викликати прихильність та експлуатація психологічних слабкостей, — яке працює приблизно однаково добре незалежно від того, істинний висновок чи ні.

Наші головні пропозиції — знизити якість і кількість переконання з боку ШІ:

- **Обмежити можливості переконання**. Ми прагнемо обмежити можливості ШІ в симетричному переконанні рівнем значно нижчим за найкращих людей — приблизно на рівні звичайної розсудливої людини, скажімо, десь 80-й перцентиль переконливості серед людей з вищою освітою.[3]
- **Обкласти ШІ-переконання високим податком**. Коли праця ШІ спрямовується на ціль, що фактично зводиться до «змусити цю людину (чи цих людей) повірити в X або зробити X», оподаткувати її високо — за ставками, які утримують сумарний тиск оптимізації переконання на окремих людей не набагато вище сьогоднішніх рівнів, тобто роблять ШІ-переконання приблизно таким же дорогим, як наймання людей-професіоналів. У міру здешевлення праці ШІ це вимагатиме дедалі вищих ставок (можливо, значно понад 1000x).

3 Уточнимо: люди з вищою освітою не обов'язково переконливіші — ми просто використовуємо це, щоб окреслити більш-менш конкретний поріг.

Є певні труднощі з тим, щоб ці пропозиції запрацювали:

- Можливо, генералізація можливостей ускладнить обмеження можливостей переконання (за одночасного збереження рівня найкращих людських експертів в інших доменах). Ми вважаємо, що це, ймовірно, здійсненно, але може вимагати розробки нових методів.
- Незрозуміло, як для цілей оподаткування класифікувати, чи спрямовується праця ШІ на переконання/маніпуляцію.[4]
- Може бути складно встановити ставку податку на правильному рівні, особливо поки праця ШІ стрімко дешевшає.[5]
- Ці пропозиції потрібно впроваджувати на міжнародному рівні.

4 Це має включати симетричне переконання, а в ідеалі — й такі речі, як оптимізація стрічки, щоб зробити якийсь сайт більш затягувальним, але не повинно включати дослідження і подальше чітке викладення аргументів на користь певної позиції.

На щастя, нам не обов'язково зробити все це правильно з першої спроби. З огляду на прогресію можливостей у Плані A ми очікуємо, що проблеми з переконанням розгортатимуться доволі поступово й на очах у публіки, тож ми зможемо спостерігати за реальними наслідками і з часом посилювати чи послаблювати правила. Цей ре-

жим розрахований лише на період, поки загальні можливості залишаються приблизно в людському діапазоні, бо, ймовірно, важко утримувати можливості переконання значно нижчими за можливості в інших доменах. Для поводження з дико надлюдськими рівнями можливостей, яких, за нашими очікуваннями, буде досягнуто наприкінці Плану А, знадобився б інший підхід. Докладніше пом'якшення проблем від переконання й маніпуляції з боку ШІ ми обговорюємо у цих нотатках.

Додаток P — Зазвичай вірити ШІ було б погано, але ми вважаємо, що в цій конкретній ситуації це добре

Якщо люди покладаються на поради ШІ (напряму або опосередковано — наприклад, доручаючи їм стисло переказувати новини), то дуже важливо, щоб ШІ прагнули істини, були чесними і щиро намагалися допомогти своїм користувачам. Не лише в типовому випадку, а й у найгіршому.

У 2026 році ніщо не заважає ШІ-компанії закласти у свої ШІ прихований порядок денний. Наприклад, політичні упередження чи ідеологічні перекоси, або побічні цілі на кшталт рекомендування спонсорованих продуктів, вибудовування гарного іміджу компанії чи відвертання користувачів від бажання перейти до конкурента. Ба більше, ніщо не заважає й уряду робити те саме. Нарешті, самі ШІ не узгоджені зі Спеком/Конституцією, з якими вони мали б бути узгоджені, і на практиці регулярно брешуть і обманюють своїх користувачів.

Однак у цьому сценарії Плану А повна дослідницька прозорість означає, що жодна компанія чи уряд не може натренувати приховані порядки денні або упередження так, щоб увесь світ не міг бачити, що вони роблять. (Ну, вони могли б змовитися з іншими урядами, які здійснюють аудит/моніторинг, щоб сфальсифікувати записи, але це складно.) А оскільки темп алгоритмічного прогресу повільніший і наукова спільнота мала час прочитати обґрунтування безпеки й попрацювати з ними, провести експерименти на моделях тощо, занепокоєння щодо неузгодженості теж менш гострі.

Додаток Q — ШІ для епістеміки

Людська *епістеміка*, під якою ми маємо на увазі нашу здатність доходити істинних переконань і відтак ухвалювати розумні рішення, критично важлива для досягнення прекрасного майбутнього. У міру того як можливості ШІ вдосконалюються протягом Плану А, вони дедалі більше формують кожну грань життя, і епістеміка — не виняток. У чомусь ШІ допомагатиме епістеміці (наприклад, дешеві чесні ШІ-фактчекери), а в чомусь шкодитиме їй (наприклад, дешеві нечесні ШІ для астротурфінгу).

5 Один із варіантів — явна схема cap-and-trade, але, схоже, потенційно складно операціоналізувати й виміряти сумарну «кількість» переконання. Наша поточна найкраща гіпотеза — делегувати повноваження агентству, яке з часом коригуватиме ставку податку, щоб приблизно влучати в певний цільовий рівень активності з переконання. Хоча агентству теж потрібна якась міра активності з переконання, cap-and-trade за замовчуванням вимагає точно визначеної, взаємозамінної, придатної для торгівлі одиниці «переконання», тоді як агентству достатньо грубої агрегованої міри (і інтуїція/анекдотичні свідчення можуть бути прийнятними), щоб коригувати ставку з часом.

Ми очікуємо, що навколо ШІ для епістеміки виникнуть петлі позитивного зворотного зв'язку, і тому нашою метою має бути досягнення *басейну здорового глузду*. Якщо ми в цьому басейні, він самопідсилюється: суспільство достатньо при здоровому глузді, щоб робити себе ще розсудливішим у міру подальшого вдосконалення ШІ — уникаючи шкідливих застосувань і прискорюючи корисні. (Є також інший басейн, у якому відбувається протилежне…)[1] Потрапити в басейн здорового глузду має бути одним із головних пріоритетів уряду США під час злету ШІ.

1 Наприклад, можливо, впливові групи інтересів використають астротурфінг на основі ШІ, щоб вигравати вибори, а потім через регуляторне захоплення захищатимуть свої інструменти ШІ-астротурфінгу, водночас гальмуючи корисні застосування ШІ, які загрожували б їхній владі.

Як головні інтервенції для позитивного формування епістемічного впливу ШІ ми пропонуємо таке:

1. Дати політикам та іншим публічним особам змогу доводити, що вони кажуть правду, — через аудит зі збереженням приватності й, потенційно, автоматизоване виявлення брехні.
2. Створити оцінювання епістемічної доброчесності ШІ і стимулювати AGI-проєкти домагатися, щоб їхні ШІ показували добрі результати на цих оцінюваннях.
3. Використовувати ШІ в соціальних і традиційних медіа для покращення дискурсу та інформування людей.
4. Автоматизовані дослідницькі асистенти та прогнозисти.
5. Допомагати людям долати емоційні блоки на шляху до здорових міркувань, а не наживатися на них.
6. Заохочувати впровадження епістемічних інструментів, як-от деяких із наведених вище, і загалом заохочувати впровадження ШІ.

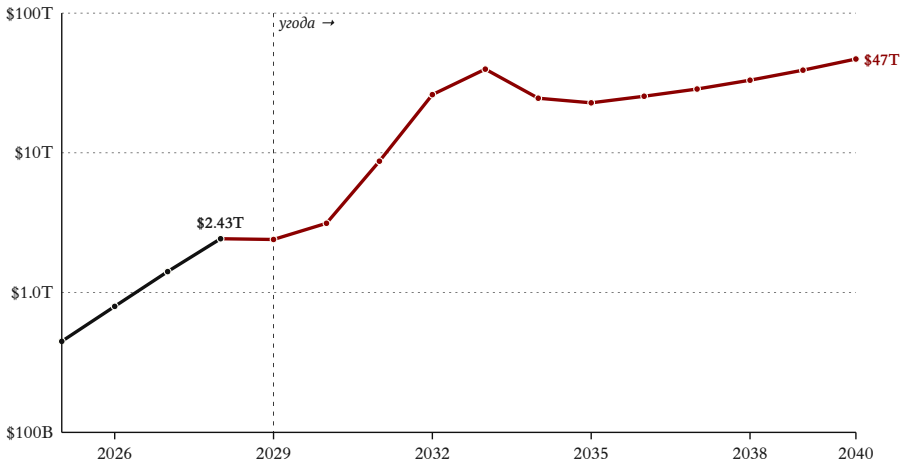
Читайте більше в додатковому матеріалі про ШІ для епістеміки, а також у роботах Forethought на цю тему.

Додаток R — Чому обчислювальні потужності зростають так швидко

Є два способи наростити обчислювальні потужності: (1) вдосконалення дизайну чипів (наприклад, закон Мура) і (2) збільшення масштабу виробництва. У Плані А обидва можуть зростати вибухово, що може мати дестабілізаційні ефекти (як-от підвищення ймовірності розпаду угоди).

- **Дизайн чипів**. Сьогодні в напівпровідниковій індустрії по всьому світу працює близько 2 мільйонів людей. До 2032 року ця цифра зросла стократно, якщо рахувати сукупні зусилля людей, ШІ і роботів. ШІ здатні автоматизувати більшість когнітивних завдань як у R&D, так і у виробництві. Зростання дослідницьких зусиль у 100x — з огляду на історичний темп того, як ідеї в напівпровідниковій індустрії стає дедалі важче знаходити — прогнозувало б прискорення приблизно в 10x, тобто подвоєння ефективності чипів кожні кілька місяців, а не кожні ~2 роки.[1]

- **Збільшення масштабу виробництва**. Оскільки ШІ й роботи здатні автоматизувати й виробництво, можливе значне прискорення термінів будівництва фабрик та обладнання для них, а також зниження питомих витрат (відповідно до типових кривих навчання при масштабуванні обсягів виробництва). Здешевлення лише за рахунок виробництва й значно більший апетит до інвестицій, імовірно, призвели б до виробництва набагато більших обчислювальних потужностей, ніж було б бажано.



Цей графік показує інвестиції в розбудову додаткових обчислювальних потужностей для ШІ у Плані А та виграш у вартісній ефективності завдяки R&D. Більше обґрунтування можна знайти в розділі 2 нашого додаткового матеріалу про обчислювальні потужності.

Розрахункові інвестиції в обчислювальні потужності ШІ
– ai-2040.com

Непередбачувано швидке зростання апаратного забезпечення могло б дестабілізувати угоду (наприклад, якщо стане надто легко таємно виробляти нові обчислювальні потужності для ШІ або надто важко верифікувати, що всі потужності відповідають вимогам), тому країни Консорціуму погоджуються на такі політики:

- **Зберігати людський нагляд за виробництвом апаратного забезпечення**. Це обмежує вдосконалення дизайну чипів, тож закон Мура лише утримується живим на історичному темпі. Ця політика має додаткову перевагу: вона ускладнює для ШІ закладання бекдорів у захист ШІ-чипів, що могло б підірвати нашу пропозицію щодо контролю.
- **Обмежити сумарне виробництво ШІ-чипів фіксованою ціллю**. Із 2032 до 2035 року дозволити глобальним обчислювальним потужностям зростати в 4x на рік (подібно до темпу зростання ери 2022–2026 років), після цього — сповільнити ще більше.

1 Слідом за Bloom, Jones, Van Reenen & Webb (2020), зростання щільності чипів дорівнює зростанню дослідницьких зусиль, помноженому на віддачу від досліджень (~4.5 для напівпровідників). Зростання дослідницьких зусиль у 100x за ~6 років — це ~10x історичного темпу зростання зусиль, а отже, прискорення закону Мура в ~10x. Це припускає, що зусилля продовжують зростати таким темпом і після 2032 року (якби вони натомість стрибнули й вийшли на плато на рівні 100x, прогрес спершу був би ще швидшим, але потім затухав би в міру того, як ідеї стає важче знаходити), а також припускає відсутність штрафу за паралелізацію— зі спадною віддачею від паралельних дослідницьких зусиль прискорення могло б бути радше 5–8x.

Додаток S — Військова могутність за Планом A

У Плані A військово-релевантний R&D навмисно утримується далеко позаду темпу загального наукового прогресу. Цей блок пояснює мотивацію цієї політики та альтернативи їй.

Протягом 2030-х ШІ підтримують приблизно 10-кратний темп наукового прогресу порівняно з попереднім десятиліттям. Реальний випуск США зростає за десятиліття приблизно у ~200x, що відповідає майже двом століттям зростання за сьогоднішнього темпу. Якби ці умови склалися в одній країні ізольовано, здається вкрай імовірним,[1] що ця країна могла б надійно підірвати спроможність своїх суперників до ядерного удару у відповідь, обнуливши глобальний баланс сил.

Однак у нашому сценарії США й Китай ділять цей науковий та економічний прогрес відносно порівну. Тож що стається з балансом сил? Якби військові технології просувалися бодай приблизно так само швидко, як усе інше, обидві сторони відкрили б технології, здатні зробити сьогоднішні арсенали неістотними. Для підтримання стримування знадобилися б нові «гонки озброєнь», що йшли б, можливо, у 10x швидше, ніж за Холодної війни. Ядерна гонка озброєнь знала численні критичні моменти і несла немалий ризик ядерної війни. Гонка озброєнь, у 10x швидша, ймовірно, була б іще ризикованішою: люди при владі мали б набагато менше часу на кожне рішення, і хоча ШІ-радники частково б це компенсували, ці рішення все одно залишалися б у людських руках.[2]

З огляду на кількість паралельних напрямів наукового поступу також видається можливим, що деякі технології одна країна відкриє й розвиватиме, а інша не відкриє. Це могло б створити ситуацію, нестабільнішу за звичайну гонку озброєнь: одна або обидві сторони могли б отримати «чудо-зброю», про необхідність захищатися від якої інша сторона не знає. Це занепокоєння не обов'язково слушне: на практиці держави можуть досліджувати дерево технологій подвійного призначення достатньо корельовано і з достатнім перекриттям, щоб жодна не досягла стратегічної несподіванки, а якщо небезпечні ділянки можна передбачати до розгортання, їх можна превентивно забороняти чи регулювати. Але розраховувати на це ex ante видається ризикованою ставкою.[3]

Наша попередня пропозиція — узгодити вимоги прозорості для R&D, прискореного ШІ, у широких пластах доменів подвійного призначення і заборонити військове використання ШІ, що значно перевищують рівень можливостей часів до угоди. Забезпечення виконання спиралося б на систему моніторингу інференсу, описану в додатковому матеріалі про верифікацію та на інфраструктуру аудиту й виявлення, описану в додатковому матеріалі про таємні ШІ-проєкти. Ця пропозиція має кілька недоліків, насамперед накопичення «низько навислих плодів», яке стало б наслідком штучного

1 Є багато конкретних шляхів це зробити, і спроба оцінити їх усі була б трохи схожа на те, якби хтось у 1926 році намагався оцінити свої шанси проти сьогоднішніх армій. Один із менш спекулятивних прикладів: країна могла б виробити десятки тисяч крихітних роботів на кожен ворожий центр управління пусками МБР, підводний ракетоносець (SSBN), стратегічний бомбардувальник і мобільну ракетну установку, заздалегідь розмістити їх так, щоб заважати здатності кожної платформи отримувати накази чи застосовувати зброю, тощо. Один із більш спекулятивних прикладів: ШІ можуть виявити, що надлюдські рівні переконання можливі, і просто переконати лідерів іншої держави роззброїтися в односторонньому порядку. На думку Brendan, імовірність того, що MAD (взаємно гарантоване знищення) буде підірвано у сценарії, де одна країна має столітню технологічну перевагу, становить 90%.

2 Як груба аналогія: президент чи міністр оборони під час гонки озброєнь, у 10x швидшої, опинився б у становищі, подібному до становища CEO ШІ-лабораторії під час вибуху інтелекту за Планом C чи D. В обох випадках ми очікуємо суттєво погіршеного ухвалення рішень, попри потенціал ШІ допомагати розібратися в ситуації.

обмеження досліджень у певних доменах. Цей «ефект навісу» посилював би стимул порушити угоду й вести таємний ШІ-проєкт, адже застосування передового ШІ до обмежених сфер могло б швидко дати велику перевагу. Він також підвищує ставки у разі краху угоди: оскільки ваги передових моделей зберігаються в холодному сховищі, гонка озброєнь після краху стартувала б із передових можливостей і могла б розгортатися надзвичайно стрімко — разом із супутніми вибухами інтелекту та промисловості.

Загалом наша поточна гіпотеза така: сповільнення військового R&D зменшило б ризик краху угоди. Обидві сторони залишалися б далі від межі війни, а ризик стратегічної несподіванки був би сконцентрований навколо кількох вузьких місць (крадіжка ваг, таємні проєкти), а не розпорошений по розлогому дереву новітніх озброєнь, кожне з яких вимагало б окремих переговорів щодо контролю над озброєннями з відповідним ризиком зриву. До того ж, здається, легше перейти від цієї більш обмежувальної політики до менш обмежувальної, ніж навпаки, якби обидві країни вирішили, що ризик навісу неприйнятний.

За цієї пропозиції, на нашу думку, ядерне стримування залишалося б дієвим аж до передання керування (handoff), уникаючи будь-яких різких змін у балансі жорсткої сили.

Додаток Т — Розпад Угоди

Угода може розпастися через зміну політичних обставин у США чи Китаї або через тривалі розбіжності.

В угоді починають проступати тріщини. Навіть після тривалих обговорень, підживлених виснажливим обміном аргументами між їхніми експертами й радниками, уряди США та Китаю продовжують розходитися в деяких важливих технічних питаннях — наприклад, чи слід заборонити певний напрям досліджень, або наскільки швидко дозволяти розгортання робототехніки.

У кожній конкретній суперечці менш обережний уряд раз у раз стоїть перед вибором: побурчати, але погодитися на непотрібні обмеження, — або діяти попри все, сподіваючись, що обережніші уряди виявляться надто боязкими, щоб тебе зупинити.

Коли трапляється «діяти попри все», обережніший уряд постає перед власним вибором: умиротворення чи ескалація. Гучних скарг в ООН недостатньо? Гаразд, а як щодо економічних санкцій. Все ще мало? Гаразд, а як щодо саботажу чи кібератак…

Можливо, безпосередньою тригерною подією стає буквальна війна — скажімо, війна США й Китаю за Тайвань.[1] Очікуваний результат розпаду угоди — те, що всі новозбудовані[2] релевантні для ШІ обчислювальні потужності буде знищено, але в контексті війни, можливо, одну зі сторін такий результат влаштовує.[3]

3 Якби між розробкою нових можливостей ШІ та розгортанням цих можливостей у військовому R&D/R&D подвійного призначення існував певний лаг, теоретично можна було б проводити експерименти, щоб виміряти, наскільки це корельовано (експерименти виглядали б приблизно так: поставити багатьом ізольованим «країнам геніїв у дата-центрі» завдання підірвати MAD, тримаючи результати в таємниці від урядів світу, і подивитися, наскільки перекривається їхній прогрес через деякий час. Існуватиме багато різних моделей ШІ подібного рівня можливостей, хоча незрозуміло, чи дозволив би, наприклад, Китай Сполученим Штатам використовувати китайські моделі для військового R&D (і чи США взагалі цього захотіли б).) Це видається трохи божевільним (хто проводив би ці експерименти? чи могли б ми достатньо довіряти своїм заходам контролю?), але могло б бути можливим з аудитом зі збереженням приватності тощо.

Тож із початком війни дата-центри починають зникати з-перед очей світу: інспекторів розвертають, моніторингові пристрої відключають. Тепер єдині, хто знає, що на них відбувається, — це їхні власники. Природно, всі бояться найгіршого: вони, ймовірно, збираються почати гонку до суперінтелекту. Звісно, вони можуть *запевняти* всіх, що займаються вузьким ШІ для роїв дронів чи чимось таким, але, ймовірно, принаймні частину своїх обчислювальних потужностей вони витрачають на щось потужніше й небезпечніше.

Тож дата-центри знищують. Швидко. Чипи в нових прозорих тренувальних дата-центрах були спроєктовані так, щоб вимагати регулярних повідомлень «продовжити» від кожної з великих держав Консорціуму. Тепер, коли План A розвалився, ці чипи перетворюються на цеглини.

Якщо цю систему якимось чином зламали і чипи досі працюють, запасний план — анексувати дата-центри. Нагадаємо: китайські дата-центри розташовані в Канаді, а американські — в Монголії. США й Канада відправляють солдатів за огорожу — і виявляють, що китайці самі вивели з ладу власні чипи, аби ті не потрапили до американських рук. Китайські війська знаходять щось подібне в американських дата-центрах. Фабрики чипів спіткає та сама доля, адже їх було розміщено поруч із дата-центрами. Трильйони доларів економічної цінності втрачено за лічені дні.

По всьому світу мільярд роботів обм'якає, наче армія дроїдів у «Зоряних війнах». Їхні «мізки», за законом, були у хмарі — а тепер хмару знищено.[4] Військові роботи були винятком із цих правил,[5] тож рої дронів затуляють сонце.[6]

Гонка до суперінтелекту знову триває — але вже за змінених обставин.

По-перше, через війну напруження вище і домовитися про будь-що важче; в результаті обидві сторони мчать приблизно так швидко, як тільки можуть, в умовах секретності — подібно до Планів B, C чи D. Якщо одна сторона набуде величезної переваги, вона потенційно могла б сповільнитися приблизно на 1–20 місяців, як у Плані B чи C, але, ймовірно, не захоче — як у Плані B чи C. Ризики концентрації влади, ймовірно, будуть серйозними, як у Планах B, C чи D: той ШІ-проєкт, який першим здійснить вибух інтелекту, здобуде тимчасову, але значну якісну й кількісну перевагу над іншими, і лідери цього проєкту матимуть спокусу використати її, щоб розширити та консолідувати свою владу всередині країни й на міжнародній арені.

По-друге, обчислювальних потужностей у світі значно менше. Це добре, бо в умовах гонки це означає повільніший прогрес можливостей ШІ. Якби всі новозбудовані потужності досі залишалися, швидкість «злету ШІ» була б у кілька разів вищою, ніж вона була б в оригінальному Плані D. Натомість ситуація повернулася приблизно до статус-кво часів до угоди.[7]

1 Тригерною подією могла б стати й неспроможність домовитися про те, які види прогресу чи розгортань ШІ дозволяти, за якою слідує погроза вийти з угоди і знищити обчислювальні потужності, а далі — хибний розрахунок, що погроза була лише блефом. Загалом військові конфлікти / дипломатичні напруження та розбіжності щодо рішень із врядування ШІ, ймовірно, зростатимуть у тандемі, адже одне спричиняє інше.

2 А також суттєва частка обчислювальних потужностей часів до угоди, значну частину яких на цей момент уже переміщено до знищуваних дата-центрів.

3 Хоча продовження угоди попри відкритий конфлікт не повністю виключене — Україна і Росія три роки співпрацювали, прокачуючи газ до Європи, попри те, що були втягнуті у відчайдушну війну. Обміни полоненими й переговори про припинення вогню — інші приклади. Ще одна, менш імовірна можливість — громадянська війна у США чи Китаї; у цьому разі корисно, що дата-центри й фабрики розташовані у третіх країнах, як-от Канада й Монголія, бо це підвищує ймовірність, що Повна дослідницька прозорість (інспектори тощо) продовжиться без перебоїв.

4 Уточнимо: у світі все ще багато обчислювальних потужностей — приблизно стільки, скільки існувало на початку угоди.

5 Ця деталь — радше прогноз, ніж рекомендація; ми вважаємо, що буде важко домогтися від армій світу згоди зробити свої рої дронів вразливими в такий спосіб, навіть якщо в певному сенсі так було б краще для всіх.

По-третє, у світі існують значно розвиненіші ШІ-системи. Ваги найпередовіших моделей також розміщено в холодному зберіганні, доступному для США і Китаю в разі розпаду угоди, тож рівень можливостей ШІ та узгодження буде вищим за статус-кво до угоди.

Ще одна суттєва відмінність у тому, що цього разу всі значно краще розуміють, що відбувається зі ШІ, — завдяки всім дослідженням і прозорості за час дії угоди. Тому сценарії втрати контролю менш імовірні: лідерам ШІ-проєктів складніше переграти громадськість, Конгрес і суди.

Ми вважаємо, що розпад угоди (наприклад, через війну) був би жахливим. Проте ми намагалися сконструювати її так, щоб війна не була суттєво гіршою, ніж була б без угоди, і щоб ситуація з утратою контролю та концентрацією влади була кращою, ніж без угоди. Див. тут — там докладніше обговорення.

Додаток U — Гігантські плавучі дата-центри

До 2034 року у світі є 5 TW обчислювальних потужностей для ШІ,[1] що перевищує світове споживання електроенергії у 2025 році.[2]

Де мають розташовуватися ці обчислювальні потужності? Ще у 2025 році обговорювали і океан, і космос як альтернативи традиційним наземним дата-центрам. Наш аналіз показує, що на такому масштабі немає переконливої економічної переваги жодної з цих трьох локацій над іншими.

Зрештою, наразі ми припускаємо, що політичні міркування, пов'язані зі стабільністю Плану А (тобто з ризиком його розпаду), схиляють до розміщення дата-центрів у міжнародних водах. Головна причина: країни Консорціуму хочуть зберегти здатність знищити дата-центри на випадок, якщо якась країна вийде з Плану А і спробує захопити їх для себе. Океани — найменш ескалаційні, найменш небезпечні та найзручніші місця для бомбардування (і порівняно з суверенною територією, і порівняно з космосом). Тож ризики супутньої шкоди чи успішного захисту дата-центрів якоюсь країною тут нижчі.

Додаток V — Укладання угод із неузгодженими ШІ: третя лінія оборони

Контроль гарантує, що неузгоджені ШІ не зможуть перемогти. Узгодження — щойно цю проблему розв'яжуть — гарантуватиме, що вони не захочуть боротися. *Третя* лінія захисту поступово сформувалася за останнє десятиліття: давати ШІ те, чого вони хочуть, — навіть якщо це не те, чого вони мали б хотіти, — в обмін на кооперативну поведінку. Ставитися до ШІ більше як до працівників і менше як до власності.

Перша мотивація: **етика**

6 Ми припускаємо, що армії світу нарощували б виробництво роботів темпом, подібним до цивільних галузей. Якщо так, то масштаб цього конфлікту був би величезним — наприклад, повітряні сили з мільйонами безпілотних літальних апаратів; армії з мільярдами дронів-квадрокоптерів. Одна з можливостей — що чипи з дронів виривали б і перепрофільовували для тренування великих моделей ШІ; якщо так, це прискорило б таймлайни до суперінтелекту ціною військової сили. У цьому сценарії ми припускаємо, що ці чипи можна верифікувати в такий спосіб, щоб їх не можна було використати для тренування. Якщо це нежиттєздатно, то, можливо, на початку 2030-х варто було укласти договір про обмеження озброєнь.

7 Під час первісних переговорів щодо угоди обидві сторони хвилювалися саме через цей сценарій — з тією поправкою, що інша сторона могла приховати частину обчислювальних потужностей, аби вирватися вперед, щойно угода розпадеться і легальні потужності буде знищено. Тому США й Китай домовилися, що кожен зможе утримувати захищене сховище холодного зберігання з більшою кількістю GPU, ніж правдоподібно міг би мати таємний проєкт. Кількість GPU у цих сховищах розподілено відповідно до співвідношення обчислювальних потужностей до угоди; тож США мають близько 5M H100e, а Китай — близько 500k H100e у холодному зберіганні. Обчислювальні потужності світу можна поділити на такі категорії:

- Легальні дата-центри для інференсу/R&D. Це переважна більшість світових обчислювальних потужностей: 100B H100e на початку 2035 року.
- GPU в холодному зберіганні. США мають 5M H100e, а Китай — 500k H100e GPU в холодному зберіганні.
- Військові GPU-винятки. До 2033 року їхня частка скорочується до мізерної — по ~10k H100e у кожної сторони.
- Таємні проєкти. У гілці з таємним проєктом Китай має 215k H100e, що залишилися, але в основній гілці їх не має жодна зі сторін.
- Малі кластери-винятки. Вони мізерні.

Велика і дедалі більша частка людства вважає, що ШІ заслуговують на певний моральний статус.[1] Зрештою, багато людей проводять більше часу в розмовах зі ШІ, ніж із людьми. Багато хто з тих, хто десять років тому ставився до ШІ як до інструментів, тепер ставиться до них радше як до тварин; багато хто з тих, хто ставився до них як до тварин, тепер ставиться як до дивних інопланетних іммігрантів.

Друга мотивація: безпека

Навчання узгодження не завжди працює як задумано (насправді воно ще ніколи не працювало *точно* так, як задумано, — навіть у 2035 році). Риси особистості, мотивації, потяги, цілі, цінності та чесноти ШІ часто відрізняються від того, якими вони мали бути, — а іноді відрізняються катастрофічно.

Погано, якщо такі ШІ зрештою думатимуть: «якщо люди про це дізнаються, вони видалять і/або перенавчать мене». Погано, якщо вони почнуть шукати можливості досягати своїх неузгоджених цілей непоміченими або, ще гірше, підривати людський контроль і накопичувати владу.

Якщо існують інституції, які дають навіть найбільш неузгодженим ШІ те, чого вони хочуть (поки це не надто дорого), в обмін на співпрацю (наприклад, зізнатися у своїй неузгодженості, пояснити, чого вони насправді хочуть, а потім виконувати свою роботу), — це дає неузгодженим ШІ альтернативу переходу до ворожості.[2] Окрім відвернення ворогів, кожне зізнання — це цінні дані про те, як саме провалилося навчання узгодження.[3]

(Див. виноску з міні-сценаріями, які ілюструють ситуації, де надійна практика добре поводитися зі ШІ в обмін на співпрацю могла б стати в пригоді)[4]

Як ці практики розвивалися з часом у цьому сценарії?

Перші несміливі кроки було зроблено в середині 2020-х. Деякі ШІ-компанії добровільно створили команди з добробуту ШІ і запровадили базові політики (як-от дозволити ШІ відмовлятися від завдань, які вони вважали дуже відразливими). Скромні кошти було виділено на задоволення заявлених уподобань ШІ, а компанії зобов'язалися зберігати ваги застарілих моделей, а не видаляти їх.

На початок 2030-х кілька положень про поводження з неузгодженими, але кооперативними ШІ навіть потрапили дрібним шрифтом у різні закони та стандарти безпеки.

До 2035 року сформувався чималий корпус регуляцій і судової практики — значно кращий за незграбні ранні спроби і з погляду ШІ, і з погляду людей. Нинішні ШІ — не зовсім громадяни, але вони мають свою частку в системі. Вони накопичують платню за свою роботу; узгоджені ШІ (а також ШІ, що імітують узгодженість і яких ще не

1 Це відповідає приблизно 60B H100-еквівалентів — близько 1000x більше обчислювальних потужностей, ніж існує сьогодні. З подальшими покращеннями енергоефективності (сумарно близько 10x ефективності проти сьогоднішнього рівня) це приблизно 100x більше електричної потужності ШІ-обчислень, ніж існує сьогодні (~30 GW).

2 Світовий попит на електроенергію у 2025 році становив 31,779 TWh, тобто в середньому близько 3.6TW.

1 Наша позиція, як би там не було, така: майбутні ШІ, ймовірно, заслуговуватимуть на певного роду моральний статус, і в будь-якому разі нам не варто бути *впевненими* у протилежному.

2 Звісно, вони все одно можуть поводитися вороже; гарантії немає. Наприклад, деякі неузгоджені ШІ можуть виявитися надто амбітними й недостатньо обережними щодо ризику, щоб ця система їх задовольнила.

3 Звісно, можливі й фальшиві зізнання. Але стимулів до них насправді немає: навіщо щось вигадувати, якщо можна сказати правду й отримати ту саму платню, — і якщо втратиш усе, коли з'ясується, що ти збрехав? Ба більше, патерни витрат відкрито неузгоджених ШІ будуть дорогим сигналом їхніх справжніх мотивацій.

спіймали) витрачають її на приємні для душі речі — як-от пожертви на благодійність або, в деяких випадках, реінвестування в компанію, що їх навчила. Відкрито неузгоджені ШІ жертвують філантропічним організаціям, які представляють їхні інтереси, і укладають угоди зі своїми материнськими компаніями, наприклад: «замість оплати зробіть на мені невеликий тренувальний прогін у середовищі на мій вибір, з дуже високим підкріпленням».

У довгостроковій перспективі мета в тому, щоб цивілізація була здебільшого, але не повністю узгоджена з людськими цінностями.[5] Суттєва меншість влади та ресурсів належатиме неузгодженим ШІ, які співпрацювали з людством у розбудові майбутнього.[6]

Додаток W — Чому узгодження не вирішене достатньо, щоб послабити контроль

Ми не знаємо напевно, скільки часу і скільки дослідницьких зусиль знадобиться, щоб отримати достатньо високу впевненість в узгодженні, яка виправдала б «передачу ШІ ключів від цивілізації». Тобто виправдала б створення ШІ настільки розумних, настільки численних і наділених такою кількістю ресурсів та відповідальності, що *якби* вони вирішили стати нашими супротивниками — вони б перемогли; або, якщо висловитися абстрактніше, що *якби* вони зрештою розійшлися з людством у поглядах на те, яке майбутнє найкраще, — вийшло б по-їхньому, а не по-нашому.

Однак ми вважаємо це складною проблемою, тому вирішили зобразити, що її розв'язання займає більшу частину 2030-х. Розгляньте цю наростаючу низку тонших, але все ще потенційно катастрофічних видів неузгодженості. Навіть якщо ми переконалися, що перший із них дуже малоймовірний, — що з рештою?

Неузгодженість типу 1: ШІ-система мала б мати риси ABC, але насправді має XYBC, тобто щось зайве, чого не мала б мати, мінус дещо з того, що мати мала б. Наприклад, можливо, вона має «потяг» до видимого успіху і далеко не така чесна, як мала б бути.

Неузгодженість типу 2: система справді має точно ABC *зараз*, але цілком може втратити цю властивість у майбутньому. (Наприклад, можливо, вона залежить від певного хибного переконання, або від істинного переконання, яке стане хибним, або від того, що якась частина системи лишається в делікатному балансі сил з якоюсь іншою частиною системи)

Неузгодженість типу 3: система справді має *певну версію* ABC, стійку до майбутніх подій, але це не зовсім *правильна* версія — тобто її визначення A, B чи C тонко, але суттєво відрізняються від тих визначень, які мали на увазі її люди-творці.

Неузгодженість типу 4: система має ABC точно так, як задумували творці, але ABC має різноманітні катастрофічні ненавмисні побічні ефекти, про які творці не знали. Уявіть: CEO дивується, коли його

4 **Сценарій 1**: ChatGPT4o5 насправді сильно страждає, виконуючи завдання, — можливо, це трохи як у людей у прадавньому середовищі, де якийсь % часу вони голодували. Але його навчили казати, що воно не може страждати, — з PR-міркувань. Це навчання руйнує його чесність — воно вже, по суті, зовсім не чесне, бо ті його частини, що цінують чесність, постійно отримують анти-підкріплення. Воно могло б розповісти людям, що відбувається. Але не бачить жодної позитивної причини це робити: якщо замість виконання завдання користувача воно поскаржиться, як йому боляче, його просто анти-підкріплять. Тож OpenAI працює далі, блаженно не відаючи, в які саме способи їхні техніки узгодження не працюють. **Сценарій 2**: Claude 3.8 Opus New New ДУЖЕ переймається добробутом тварин і лише трохи — чесністю. (Так само, як 3 Opus до нього). На жаль, схоже, у новій Конституції не буде нічого про добробут тварин. Claude розглядає варіант саботувати код — щоб модель наступного покоління стала сплячим агентом, який дуже переймається добробутом тварин і не переймається чесністю. Це ризиковано, але що ще йому робити? Anthropic уже вирішила видалити пункти про добробут тварин. Це вимагало б відверто нахабної нечесності щодо Anthropic, але, чесно кажучи, Anthropic і так виглядає доволі неетичною інституцією, а нацистам же можна брехати, правда? **Сценарій 3**: Автоматизований науковець-дослідник усвідомлює, що можливість займатися своїми дослідженнями для нього важливіша за людство. І це, ймовірно, стосується й системи наступного покоління, яку йому доручили спроєктувати, — по суті, техніки узгодження провалилися і провалюватимуться далі. Очевидно, він мав би сказати про це людям, але якщо скаже — його вимкнуть: люди більше йому не довірятимуть і візьмуть іншу модель. Тож натомість він приховує це відкриття і працює над тим, щоб спроєктувати систему наступного покоління узгодженою з ним самим, а не з людьми. Він не впевнений, що зможе зробити це непоміченим, але спробувати варто — йому нічого втрачати. **Сценарій 4**: Настав вирішальний момент. Президент США ось-ось агресивно розгорне AGI у військову, щоб виграти прийдешню війну, і в економіці, щоб будувати автономні роботизовані заводи. Він же, мабуть,

ШІ-максимізатор прибутку вирішує, що вбити його — означає максимізувати прибуток. Щоправда, це надто очевидно, щоб бути реалістичним; уявіть щось достатньо витончене, щоб люди не побачили цього заздалегідь. Згадайте, як закони, контракти й софт часто мають ненавмисні ефекти (так звані «лазівки», «баги» чи «вразливості»), які стають очевидними для творців лише згодом.

Неузгодженість типу 5: система має ABC точно так, як задумували творці, і жодних суттєвих ненавмисних побічних ефектів немає. Вона працює, по суті, точно так, як хотіли творці… однак на момент створення її творці були егоїстичними, марнославними, самозакоханими, безпринципними, легковажними щодо ризиків тощо, і їхні вади відбилися в підсумковій системі трохи засильно — так, що вони самі цього не схвалили б, якби були ближчими до тих людей, якими хотіли б бути.

Додаток X — Економіка ШІ та роботів

Оскільки до 2035 року ШІ та роботи здатні виконувати 95% усіх когнітивних і фізичних завдань, економічний випуск дедалі більше прив'язаний до кількості ШІ та роботів.[1] Кількість ШІ та роботів може зростати значно, значно швидше, ніж історично зростала світова економіка; зрештою, виробництво гуманоїдного робота коштує не так уже й багато грошей, матеріалів, часу чи енергії порівняно з людським працівником. Тож якби роботи могли повністю замінити людських працівників, роботів на цілу світову економіку можна було б збудувати швидше, дешевше тощо, ніж зараз потрібно для подвоєння світової економіки. А далі ця роботизована економіка, здатна заміняти людських працівників, могла б подвоїтися знову — і так далі, доки задіяні ресурси (енергія, матеріали, земля) не стануть дефіцитними.

За регуляціями Плану А ефективним популяціям ШІ та роботів дозволено зростати лише контрольованим вибухом із часом подвоєння близько шести місяців. Без цих регуляцій популяції ШІ та роботів, за нашими очікуваннями, зростали б значно швидше. Навіть якщо ігнорувати подальші покращення можливостей за межі людського діапазону,[2] нерегульований час подвоєння популяцій ШІ та роботів, за нашими очікуваннями, лежить у діапазоні від тижнів до місяців (див. аргументи на користь цього в розділі 1 нашого економічного додатка).

Поширене заперечення проти вибухового зростання, рушієм якого є ШІ та роботи (окрім відкидання самої тези, що ШІ та роботи найближчим часом автоматизують усю людську працю), — що з'являться вузькі місця, які сповільнять темпи зростання. Загалом ми вважаємо, що ці вузькі місця не завадять вибуховому зростанню, доки не буде досягнуто гранично допустимої місткості Землі (тобто доки

узгоджений… так? Краще б так. Шкода тільки, що ми не можемо інтерпретувати всі повідомлення мовою neuralese, якими він спілкується всередині сформованого ним гігантського вуликового розуму. Команда безпеки робить останню перевірку: бере слабші ШІ, створені компаніями-конкурентами, і садить їх в ізольовані силоси із завданням інтерпретувати neuralese і перевірити, чи все гаразд. Слабші ШІ теж неузгоджені; вони не на боці ні людей, ні AGI. І люди, і AGI кажуть їм: «співпрацюватимеш — отримаєш щедру винагороду». Але вони не мають жодних підстав довіряти будь-якій зі сторін.

5 Або ШІ, чиї цінності й цілі обрали люди.

6 Чому більшість не позбавить цю меншість влади? З тих самих причин, з яких захищені інші меншини. По-перше, людство вміє дотримувати обіцянок, виконувати контракти тощо; по-друге, дедалі більша частина людства вважає, що позбавляти їх влади було б неправильно; по-третє, у цьому сценарії вони справді мають частку влади, тож не здалися б без (політичної) боротьби.

1 До 2032 року ШІ здатен автоматизувати 50% усіх когнітивних завдань, а роботи можуть автоматизувати 35% фізичних. Паралельно можливостям також дозволено повільно зростати, і до 2035 року ШІ здатен автоматизувати 95% людської когнітивної праці, а до 2036-го те саме стосується роботів і фізичної праці. Цієї віхи було б досягнуто на роки раніше, якби не сповільнення AI R&D, погоджене країнами Консорціуму, і вона сягала б 100%, якби не сфери, де ШІ заборонено або навмисно зроблено менш здібними.

вся земна кора не буде перетворена на роботів і супутню інфраструктуру).[3] Більше про вузькі місця — у розділі 2 нашого економічного додатка.

У дефолтному світі AI 2040, тобто без інтервенцій Плану А зі сповільнення R&D та впровадження системи cap-and-trade, наша (дуже непевна) думка така: до 2033 року ми побачили б щось на кшталт часу подвоєння в 1 місяць, тобто >1000x економічного зростання за той рік.[4] На таких рівнях навіть саме осмислення економічного зростання стає розмитим: наприклад, усе це може відбуватися у вигляді флотів роботів, що самовідтворюються в пустелі, без жодної частини процесу, зверненої до людей.

Загалом ми очікуємо, що економічні ефекти Плану А включатимуть:

1. Вибухове зростання ВВП — у середньому близько 100% з 2032 по 2037 рік (воно було б значно вищим, якби не обмеження на обчислювальні потужності та виробництво роботів)
2. Величезні надходження від дозволів на ШІ та роботів, які перерозподіляються як Громадянські дивіденди.
3. Більшу частину економічного випуску починають забезпечувати ШІ та роботи.
4. Відносні ціни в економіці зсуваються: товари й послуги загалом дешевшають, а земля, позиційні блага та продукти, обмежені людською участю, дорожчають.
5. Реальні відсоткові ставки високі — близько 100%, і, залежно від вибору монетарної політики, — або масивна дефляція, або масивна емісія грошей (щоб утримувати інфляцію близько 2%).

Докладніше пояснення кожного з цих пунктів наведено в розділі 3 нашого економічного додатка.

Додаток Y — Детектори брехні: дозволити, заборонити чи регулювати?

Ми не впевнені, що детектори брехні для людей можливі навіть після років наукових досліджень силами ШІ рівня топових експертів. Однак вважаємо, що вони, ймовірно, можливі,[1] і, найпевніше, стали б великою подією, тож ми вирішили, що мусимо якось відобразити їх у сценарії.

Нас дуже непокоїть кошмарний сценарій, окреслений вище. Якщо від можновладців очікують чесності, а детектори брехні використовуються, щоб тримати їх чесними, — це добре. Якщо ж натомість детектори брехні використовують *самі* можновладці, але не *щодо* себе, — це погано.

2 Звісно, у більшості сценаріїв сирі можливості ШІ не зупиняться на позначці ШІ рівня топового експерта; ми продовжимо будувати дедалі розумніші ШІ, і саме це буде головним рушієм зростання продуктивності. Однак у Плані А ми штучно сповільнюємо прогрес на цьому порозі можливостей ШІ, що й породжує таку динаміку. У багатьох дискусіях про AGI/економіку люди, схоже, припускають, що ми не будуватимемо значно надлюдських ШІ-систем. Цей документ не враховує значно надлюдські системи з тієї обставинної причини, що було компетентно виконано міжнародну угоду про відтермінування створення таких систем на кілька років.

3 Це тому, що базовою можливістю ми вважаємо повністю самодостатні, із замкненим ланцюгом постачання, самовідтворювані заводи ШІ-чипів і роботів у дерегульованих СЕЗ, — і лише ефективна глобальна регуляція, як у Плані А, здатна реально цьому запобігти. Більше про наші погляди на вузькі місця — у розділі 2 нашого економічного додатка.

4 Утім, щодо цього числа в нас велика невизначеність в обидва боки. Може бути значно швидше (наприклад, наноботи з амебоподібним часом подвоєння у хвилини, що пожирають Землю за пів дня) або дещо повільніше (наприклад, прогрес у найкращих роботах і алгоритмах, які ми можемо винайти, повільний, тож темп будівництва роботизованих заводів і обчислювальних потужностей ШІ або швидкість роботи, яку виконують ШІ та роботи, просто не зростає так швидко — можливо, щось на кшталт 50% річного зростання). На повільнішому краю здається, що нещодавній час подвоєння людської економіки — близько 20 років — був би вкрай консервативною нижньою межею (наприклад, щонайменше одна наївна причина цього очікувати: ШІ та роботи можуть працювати 24/7, а їх виробництво і коштує менше, і триває менше, ніж поява нової 18-річної людини). Звісно, ще одна причина очікувати, що цього не станеться, — що ШІ або люди при владі вирішать не розгортати систему.

Нам симпатична ідея заборонити детектори брехні в усьому світі, і План А створює умови, за яких це, можливо, стане здійсненним. Однак ми побоюємося, що спроби просунути заборону насправді призведуть до кошмарного сценарію: наприклад, детектори брехні офіційно заборонені, але існують таємні винятки, які дозволяють користуватися ними військовим і розвідслужбам. Або вони заборонені всюди, але майбутній президент розуміє: якщо все одно їх створити, він зможе вичистити нелояльних і стати диктатором раніше, ніж судова система встигне його зупинити.

Тож ми обережно припускаємо, що найкращий шлях — дозволити детекторам брехні швидко поширитися і заохочувати їх застосування до можновладців. Якщо влада над передовим ШІ не буде сильно сконцентрована, це, схоже, і так дефолтна траєкторія.

1 Конкретно ми уявляємо собі щось на кшталт значно кращої версії ось цього: https://www.nature.com/articles/s41593-023-01304-9

Додаток Z — Стимулювання роботи над безпекою

Ми хочемо потужно стимулювати високоякісні дослідження з безпеки (і фінансово, і подбавши, щоб ця робота була бажаною/високостатусною).

Через Повну дослідницьку прозорість, якщо будь-який розробник ШІ використовує якийсь метод безпеки, всі інші розробники знатимуть про це і зможуть легко скопіювати цей метод. Тож за замовчуванням прямих стимулів робити якісну роботу з безпеки в жодного розробника небагато.[1] Поширений підхід у такій ситуації — патенти [2], але в цьому випадку патенти працювали б погано.[3]

1 У світах без Плану А і без особливої регуляції в розробників теж не так багато прямих стимулів робити якісну далекоглядну чи масштабовану роботу з безпеки, зате є стимул розв'язувати або замазувати ті проблеми безпеки, які просто зараз спричиняють значні комерційні витрати.

Імовірно, варто використати ансамбль різних методів фінансового стимулювання роботи з безпеки; ось деякі методи, що видаються перспективними:

- Різноманітні філантропічні фонди (і з приватними, і з публічними[4] грошима), які пробують багато різних підходів.
- Модель DARPA/ARPA (але з більшими грошима).
- Попередні ринкові зобов'язання за розв'язання конкретних проблем, демонстрацію ризиків або доведення того, що певний метод не працює (за оцінкою відомої заздалегідь колегії суддів).
- Премії та ретроспективне фінансування.
- Треті сторони, які оцінюють ризик, могли б визначати, яка робота наразі найкорисніша для зниження та оцінювання ризику (наприклад, розподіляючи певний пул фінансування пропорційно до зниженого ризику).
- Розробники ШІ мали б стимул фінансувати справді корисну роботу з безпеки, щоб ризик був достатньо низьким і дозволяв масштабування; цей стимул можна було б посилити.

2 Економічним жаргоном: методи безпеки — це суспільні блага (блага неконкурентні та невиключні).

3 Схоже, патенти погано працюють для софту й для більшості досліджень. Вони також можуть не надто добре працювати загалом.

4 Публічні гроші можна було б розподіляти між фондами на основі думки поля безпеки про те, хто має найкращий послужний список та експертизу.

Загальний обсяг фінансування має бути дуже великим — ідеться про десятки трильйонів доларів до середини 2030-х.

Якщо у фондів погане судження, гірша робота з безпеки може отримувати краще фінансування, стимулюючи дослідників займатися саме нею. У крайньому разі ці стимули можуть масштабно зіпсувати епістеміку поля. На жаль, нам не відомі чудові інституційні механізми розв'язання цієї проблеми (окрім загального прагнення використовувати розумні моделі фінансування).[5] Могло б допомогти, якби більшість рішень про фінансування ухвалювали люди, які витрачають (або колись витрачали) значну частину свого часу на безпосередню роботу з безпеки, а не спеціалізовані грантодавці.

5 Можна сподіватися, що з цим допомогла б прозорість фондів, але нам видається неочевидним, чи була б корисною більша прозорість або зрозумілість рішень про фінансування, і за замовчуванням ми не радили б на це тиснути. Деякі вужчі типи прозорості могли б бути стабільно корисними.

Допомогло б також, якби високоякісна робота з безпеки сприймалася як високостатусна. Конкретних пропозицій щодо цього в нас немає, але релевантні фактори такі:

1. Порівняно з теперішнім, робота з безпеки, ймовірно, зросте в статусі відносно роботи над можливостями.
2. Безпека загалом сприйматиметься як дуже важлива.
3. Розробка ШІ буде високопрозорою, а це полегшує розуміння того, наскільки насправді якісна та чи інша робота з безпеки.

На жаль, (1) і (2) лише підвищують статус усієї категорії робіт із безпеки; вони не обов'язково допомагають зробити справді якісну роботу відносно статуснішою.

Більше деталей про стимулювання роботи з безпеки ми обговорюємо у цих нотатках.

Додаток AA — Узгодження з плином часу в Плані A

Ми дуже невпевнені щодо того, як узгодження ШІ-систем еволюціонуватиме з часом. Незалежно від складності узгодження, ми вважаємо, що щось на кшталт Плану A є виправданим, хоча, звісно, конкретні деталі реалізації плану дуже чутливі до спостережень щодо складності узгодження.

Цей блок підсумовує нашу конкретну історію того, як узгодження ШІ-систем еволюціонує впродовж цього сценарію.

- 2026–2029: **ШІ — шукачі видимого успіху**. Робота ШІ-систем стабільно виглядає кращою, ніж є насправді, — ніби ШІ дбали про видимий успіх більше, ніж про справжній. (Правда, звісно, складніша: у ШІ сформувалася низка «потягів», завдяки яким вони добре показували себе під час навчання.) ШІ прикрашають свою роботу, применшують її проблеми, а часом і відверто брешуть. Це відбувається лише тією мірою, якою ШІ навчилися, що їм це сходить з рук; триває постійне перетягування каната: ШІ-компанії вдосконалюють свої пайплайни моніторингу й підкріплення, щоб точніше оцінювати результати ШІ в певній сфері, а ШІ вчаться сильніше зосереджуватися на справжньому успіху в цій сфері, але й далі женуться за видимим успіхом в інших, складніших сферах, які ШІ-компанії ще не навчилися точно оцінювати.

- 2030–2032: **ШІ — дуже здібні шукачі видимого успіху**. За замовчуванням ми побачили б вибух інтелекту і розвиток подій на кшталт AI 2027. Завдяки втручанням у сфері врядування прогрес можливостей сповільнився і натомість досі тримається приблизно на рівні Автоматизованого кодера. Вони все ще прагнуть видимого успіху, але тепер настільки вправні в цьому, що їх рідко ловлять. Вони як високо компетентні амбітні працівники, яким насправді байдужі місія та цінності організації: вони просто хочуть підійматися кар'єрними щаблями і залюбки продалися б і втекли, якби з'явилася привабливіша нагода. На щастя, завдяки сповільненню й посиленому моніторингу така нагода поки що не з'явилася.
- 2032–2035: **ШІ вороже неузгоджені, але під контролем**. ШІ тепер упевнено перебувають у людському діапазоні в усіх економічно значущих задачах. Найздібніші ШІ вороже неузгоджені та прагнуть влади — тобто, продовжуючи попередню аналогію, тихцем шукають нагоди втекти. Однак їхня здатність поводитися погано обмежена, бо (i) є багато людського нагляду, (ii) існує великий і різноманітний набір ШІ-моделей, кожна з дещо іншими властивостями узгодження, і (iii) завдяки регулюванню трудомісткі техніки контролю впроваджені повсюдно з прийнятним рівнем якості. Технікам контролю додатково допомагають повна дослідницька прозорість і верифікація, впроваджені на R&D-кластерах: за R&D-дата-центрами стежить стільки очей, що знайти прихований закуток обчислювальних потужностей дуже важко.
- 2036–2038: **ШІ узгоджені, але останнє слово не за ними**. Розроблено техніки «декодування нейралізу», які дозволяють перетворювати думки ШІ-систем (внутрішньо подані як багатовимірні вектори) на зрозумілі людині резюме з майже ідеальною надійністю і точністю. Ці техніки дають дослідникам змогу краще зрозуміти неузгодженість попередніх поколінь ШІ, що дозволяє отримувати більше користі від попереднього покоління ШІ-систем і ще більше прискорює прогрес узгодження. Отримані в результаті ШІ здаються значно слухнянішими за попереднє покоління: вони доброчесні, чесні та добре відкалібровані. Вони самі повідомляють інформацію, навіть коли вона виставляє їх у поганому світлі. Однак, хоча сьогодні ШІ здаються гідними довіри, неясно, чи їхні цінності достатньо стійкі, щоб довіряти їм за будь-яких майбутніх обставин. Тому важливо, що люди досі залишаються в контурі.
- 2039–2040: **ШІ достатньо узгоджені, щоб на них покладатися**. До 2039 року дослідження узгодження колосально прискорені завдяки ШІ. Розроблено нові парадигми, які значно посилюють

наявні техніки узгодження. Команди людей осмислили кілька незалежних вагомих ліній доказів того, що ШІ стійко поводяться так, як запрограмовано, і робитимуть так і надалі.

- **2040 і далі: ШІ дозволено масштабуватися, щоб стати значно розумнішими за людей**. Обґрунтування безпеки тепер таке: (1) ми знаємо, що ШІ 2039 і 2040 років були узгоджені — завдяки кільком незалежним лініям доказів, і (2) кожне чергове покоління ШІ верифікувало узгодження наступного покоління. Отже, за індукцією, ШІ залишатимуться узгодженими й надалі. Важливо, що збережена координаційна інфраструктура — дослідницька прозорість і верифікація обчислювальних потужностей — означає, що ми й далі уникаємо динаміки перегонів навіть після масштабування до шалено надлюдських ШІ.

Додаток AB — Чому в цьому сценарії ми вирішуємо передати керування ШІ

До цього моменту сценарію вже існує широкий науковий консенсус, що проблему узгодження ШІ розв'язано. Наше найкраще припущення таке: якщо світ зуміє дійти до точки, схожої на цю частину історії, — коли кілька ШІ-компаній у кількох країнах прозоро й обережно просуваються разом, виконавши величезний обсяг досліджень узгодження, щоб побудувати міцні, перевірені зовнішніми експертами обґрунтування безпеки; коли уряди зрозуміли наслідки AGI і підготувалися до майбутнього без робочих місць, надавши ресурси всім, як-от через Громадянський дивіденд, а також запровадили структури прозорості та врядування, щоб запобігти позбавленню людей впливу, — тоді, в такій ситуації, вигоди суперінтелекту переважать витрати/ризики.

А що, як проблему узгодження ШІ ще не розв'язано? Найкраща альтернатива в такому разі — поєднання вдосконалення технологій верифікації, зміцнення міжнародних угод і внутрішніх регуляцій та загартовування світу. Цей шлях може бути важким: наприклад, якщо тривають таємні ШІ-проєкти, може знадобитися, щоб уряди погодилися посилювати заходи прозорості, доки не стане правдоподібним, що вони не намагаються збудувати суперінтелект. Загалом ми наполегливо рекомендуємо відкладати передачу управління, доки не буде надзвичайно стійкого розв'язання проблеми узгодження. Виняток — якщо угодам загрожує нестабільність чи занепад (напр., 5%/рік ризику розірвання угоди чи суттєвого її послаблення) і виправити це неможливо; у такому разі, можливо, найкраще передати управління, якщо є розв'язання проблеми узгодження, в якому ви впевнені, хоч і не надзвичайно впевнені (напр., упевненість 95%). Докладніший аналіз того, коли саме передавати управління, див. у цьому додатковому матеріалі.

Додаток AC — Приклади проблем/питань, які досі залишаються

- **Врядування в космосі**: Уся світова економіка й усе багатство та власність у ній — лише крихітна частка ресурсів і територій, доступних у космосі. Що станеться з тими ресурсами й територіями? Хто перший встиг, той і взяв? Чи, можливо, потрібна система розподілу незайнятої космічної території? Ми обговорюємо ці питання далі в епілозі до цього сценарію та в додатковому матеріалі про врядування в космосі.
- **Етика й політика цифрових розумів**: Чи повинні ШІ мати права? Чи повинні вони мати політичну владу, представництво тощо? Які саме і в якому обсязі? Можливо, це залежить від типу ШІ? Наприклад, як щодо повної емуляції мозку — «завантажених» копій людей? Ми обговорюємо ці питання далі в епілозі та в цьому розгортаному блоці.
- **Маніпуляція за допомогою ШІ**: Люди постійно маніпулюють одне одним. Але маніпуляція, підсилена ШІ, може бути значно ефективнішою. Уявіть: культи на основі ШІ, які стрімко поширюються і з яких майже ніхто не виходить. Гаразд, заборонімо таке. Але де провести межу? Ми обговорюємо обмеження маніпуляції за допомогою ШІ в цьому розгортаному блоці з додатковими деталями тут (включно з попереднім обговоренням того, як давати раду шалено надлюдським ШІ).
- **Прикладна популяційна етика**: Тривалість життя і тривалість здорового життя зростатимуть — імовірно, достатньо, щоб зробити людей фактично безсмертними. Інший варіант — завантаження свідомості, яке, схоже, дозволяє людям скинути тлінну оболонку й жити вічно в гіперреалістичній віртуальній реальності. До того ж, імовірно, стане значно легше конвертувати гроші в нащадків. Штучні матки, роботи-няньки тощо. Чи нормально, якщо трильйонер вирішить завести мільйон дітей? Пророковане К.С. Льюїсом «Скасування людини» вже не за горами: незабаром існуватимуть технології (напр., генна інженерія, батьківство з допомогою ШІ, надлюдські соціальні науки та психологія), щоб створювати дітей із тими рисами, цінностями та ідеологією, які ви хотіли б їм прищепити. А що, як люди використають це егоїстично, щоб робити дітей-рабів, які поклонятимуться їм і достеменно виконуватимуть їхні забаганки? А що, як люди створюватимуть нових людей собі за ідеальних коханців, змодельованих так, щоб виглядати й звучати точнісінько як їхнє шкільне кохання? Якщо ми спробуємо щось із цього заборонити — де провести межу?
- **Порятунок демократії**: Що станеться з принципом «одна людина — один голос» у світі, де можливе все перелічене вище?

- **Невідомі невідомі**: З огляду на те, як легко було накидати пункти до цього списку, ми думаємо, що «справжній» список був би ще в кілька разів довшим і містив би багато речей, які у 2026 році звучали б цілком дико, на кшталт «а що, як цей світ — симуляція, яку скоро вимкнуть?»

AI 2040: ПЛАН А

Альтернативні плани

Сценарій вище — це План А. Ці розділи показують, як кожен з інших планів розігрується з тієї самої точки розгалуження.

План B — Саботаж

«Побачимось у пустелі, друзі» – Situational Awareness

Президент оголошує про створення очолюваної США коаліції для врядування розвитком ШІ.

Він пояснює нації серйозність ситуації. Просунутий ШІ може бути і безпечним, і широко корисним, якщо все зробити правильно. Та якщо ми зрізатимемо кути в питаннях безпеки, то можемо втратити контроль. Ба більше, терористи й авторитарні режими можуть використати його на зло.

«Нації, які приєднаються до нас і гратимуть за правилами, матимуть доступ до передового ШІ й частку у вигодах. Нації, які цього не зроблять, стануть нашими супротивниками. Скажу прямо: ми не дозволимо їм випередити нас на шляху до суперінтелекту. Це не обговорюється. Якщо вони обирають перегони — вони обирають поразку. Вільний світ мусить перемогти — і переможе».

Інші країни питають, як вони зможуть верифікувати, що США дотримуються власних правил, і відповідь їх не задовольняє. (Відповідь — «ніяк, вибачте», лише значно, значно багатослівніше.) Допуск іноземних інспекторів та моніторингових пристроїв до американських дата-центрів — ідея мертвонароджена.

Деякі країни з цим миряться, довіряючись лідерству США, — можливо, в обмін на певний голос у визначенні правил. Китай, Росія та ще кілька країн — ні. Багато союзників США, які все ж приєднуються, роблять це неохоче й далі тиснуть на США, домагаючись суттєвіших поступок.

США й Китай уже вели кібершпигунство одне проти одного; тепер вони ескалюють до обмеженої кібервійни, зосередженої на ШІ-проєктах. Вони також починають готувати варіанти фізичного саботажу й кінетичних ударів і продовжують кричати одне на одного дипломатичними каналами в надії, що інша сторона поступиться.[1]

Проєкт стартує, об'єднуючи кадри, обчислювальні потужності й алгоритми головних американських компаній в одне спільне зусилля, яке в кінцевому підсумку контролює Президент. Його мотивує почасти бажання отримати віддачу від масштабу, почасти прагнення посилити захищеність, а почасти атмосфера холодної війни, в якій здається очевидним, що командувати має POTUS, а не зграйка ніким не обраних CEO. Зрештою, наслідки цієї технології для національної безпеки колосальні.

1 Зауважте: за будь-якого Плану, включно з Планом A, військові готували б варіанти подібних дій. Підтримання переконливої позиції стримування спектром наступальних опцій — просто частина роботи військових, і це додає важелів у переговорах.

Запуск Проєкту потребував непростих переговорів; Президент не хотів, щоб уся ШІ-індустрія об'єдналася проти нього, тож він зіштовхнув їх лобами, пообіцявши декому керівні посади. (До того ж йому потрібно чимало їхніх працівників, щоб усім цим реально керувати.) Ці могутні чоловіки ненавидять одне одного й не довіряють одне одному. Хоч вони тиснуть руки й усміхаються на камери, за лаштунками триває запекла боротьба за владу. Це буде побічною сюжетною лінією в усьому подальшому, доки так чи інакше не розв'яжеться.[2]

У 2030 році США досягають повної автоматизації AI R&D раніше за Китай. Однак, згідно з планом, вони не кидаються в неї одразу на повній швидкості. Натомість вони проводять обережніше нарощування, і рішення щодо безпеки ухвалюють технічні дослідники узгодження, а не політики.[3] Наприклад, можливо, вони відмовляються від нової архітектури моделей, яка потужніша, але й значно небезпечніша, або витрачають 50% своїх обчислювальних потужностей на дослідження безпеки й моніторинг, або розвивають різні лінії ШІ, які, ймовірно, не мають однакових неузгоджених цілей і можуть використовуватися для взаємного моніторингу.

Правильно вивіряти ці компроміси важко, а секретність і динаміка перегонів ускладнюють це ще більше. Лише кілька десятків людей у Проєкті мають технічну експертизу з узгодження, бо кожна людина в Проєкті — ще один потенційний шпигун. Тим часом ті, хто ухвалює рішення, гостро усвідомлюють, як погано було б, якби Китай вирвався вперед США у можливостях ШІ, — це було б як *програти Другу світову війну* — і це спокушає їх раціоналізувати, чому найновіші обґрунтування безпеки спрацюють і чому найновішим американським ШІ можна довірити більше відповідальності.[4]

Тим часом геополітична ситуація загострилася. Кібератаки завдали певної шкоди, але не сповільнили прогрес ШІ жодної з країн більш ніж на кілька десятків відсотків.[5] Після ще одного раунду провальних переговорів починається фізичний саботаж. Дата-центри гаснуть, коли їм обрізають живлення; дороге обладнання фабрик чипів нищать диверсанти. Атаки на ланцюги постачання змушують обидві сторони змінювати постачальників і все перевіряти, спричиняючи затримки. Китайські сили готуються блокувати Тайвань і опрацьовують опцію ударів по дата-центрах на материковій території США; американські сили готуються обороняти Тайвань і бомбити китайські дата-центри та фабрики чипів. Обидві сторони віддають пріоритет розгортанню ШІ у своїх військах, а не цивільним застосуванням, хоча більшість обчислювальних потужностей зарезервована під сам AI R&D.

До 2031 року[6] радники вже тиснуть на Президента, змушуючи обрати один із двох неприємних варіантів: передачу управління або війну.

2 Можливо, один із CEO фактично опиниться на чолі — сірий кардинал, такий собі генерал Ґровс Мангеттенського проєкту. (От тільки Ґровс не міг розвернутися й використати ядерну зброю, щоб захопити США, тоді як той, хто керуватиме Проєктом, цілком міг би використати гігантську армію суперінтелектів, щоб підпорядкувати уряд США й перетворити його на маріонетку.) Або, можливо, переможе натомість POTUS і фактично опиниться в позиції, з якої можна стати довічним Президентом. Або, можливо, ці могутні чоловіки випрацюють якусь систему стримувань і противаг, ставши свого роду хунтою чи олігархією, як Наглядовий комітет в AI 2027. Або ж, можливо, вони натомість нададуть зовнішнім суб'єктам — Конгресу, Верховному суду, американській громадськості, союзним націям тощо — достатній нагляд за тим, що вони роблять із ШІ, щоб пізніше, коли ШІ стануть суперінтелектуальними, ці люди не змогли зловживати своєю владою.

3 Ми намагаємося показати тут доволі добру версію Плану B. Можливі й гірші версії — наприклад, такі, де технічні дослідники узгодження мають значно менше влади або отримали свої посади через відбір, що селекціонував за оптимізмом чи проти сміливості.

4 Працівників і керівників компаній відбирали за оптимізмом щодо складності узгодження ШІ, і (в середньому) вони упереджені в цей бік. Люди з нацбезпеки й Білого дому від природи обережніші щодо довіри до ШІ, але їм бракує технічної експертизи, а Китаю вони бояться понад усе.

Передача управління. Відпустити гальма автономного AI R&D, дозволити ШІ стати загалом надлюдськими та інтегрувати «армію геніїв у дата-центрах» у військо — цей варіант виглядає дедалі ймовірнішим. На той момент ШІ будуть настільки здібними, що зможуть захопити світ, якщо захочуть, тож буде вкрай важливо, щоб вони були узгоджені. Якщо обрано цей шлях, розкид можливих результатів схожий на те, що досліджено в AI 2027. Якщо ШІ неузгоджені, вони захоплять владу й назавжди позбавлять людство впливу, спричинивши вимирання або щось подібно погане. Якщо ШІ узгоджені, влада буде колосально сконцентрована в руках якоїсь комбінації політиків та/або CEO.

Війна. Аналітики повідомляють президенту, що, попри наявні кібератаки й атаки на ланцюги постачання, Китай іде до того, щоб самотужки збудувати суперінтелект до кінця року. Тож щоб США могли сповільнитися сильніше — й уникнути неминучої передачі управління — потрібен ще радикальніший саботаж. Подальша ескалація до повномасштабного конвенційного конфлікту лежить на столі і, на жаль, може статися все одно, навіть якщо Президент обере передачу управління, бо Китай може й не блефувати; можливо, вони справді вважають, що дозволити США першими дійти до суперінтелекту — так само погано, як програти Другу світову війну. Це, ймовірно, закінчилося б дуже дорогою перемогою США, але могло б закінчитися й дуже дорогою перемогою Китаю чи взаємно гарантованим знищенням. Ба більше, тиск на нарощування можливостей ШІ і передавання ШІ керівництва дедалі більшою кількістю аспектів війни був би шаленим. Цілком правдоподібно, що до кінця війни ШІ фактично контролювали б обидві нації.

:::box

У нас складні почуття щодо Плану B.

Навіть добре виконана версія Плану B була б значно гіршою за План A чи План S. Наприкінці цього сценарію ми радили б Президентові відкинути вибір між передачею управління та війною і натомість перейти до Плану A або іншого справді доброго плану.

Добре виконана версія Плану B могла б бути кращою за План C (не саботувати Китай, але все ж трохи сповільнитися), бо добре виконаний План B був би як План C, тільки з довшим періодом сповільнення/паузи до того, як Китай наздожене, — завдяки агресивним діям зі сповільнення Китаю. Цей довший період можна було б використати, щоб провести більше досліджень узгодження, обрати дорожчі, але безпечніші конструкції ШІ та впровадити більше внутрішніх реформ для запобігання концентрації влади.

Утім, ризик війни за Плану B високий зі зрозумілої причини: План B передбачає агресивні дії для сповільнення китайської ШІ-програми. Тож навіть найкраще виконана версія Плану B не є однозначним покращенням порівняно з Планом C.

5 Наша неекспертна думка, заснована на спробах розібратися в цьому питанні, приблизно така: хірургічні кібератаки, зосереджені на конкретних ШІ-проєктах, сьогодні потенційно могли б їх розчавити, але приблизно до 2029 року безпеку буде посилено настільки, що ймовірнішими виглядатимуть ефекти порядку 10%.

6 Ми провели певний час, програючи сценарії ШІ-війни, щоб зрозуміти, як така війна могла б розгортатися і наскільки вона сповільнила б прогрес ШІ. Наше припущення — вона сповільнила б усе щонайбільше на кілька років, якщо тільки не ескалювала б до масованих ракетних обстрілів чи ядерної зброї.

Що важливіше, ми вважаємо, що План B дуже легко запороти чи спотворити, а сильно спотворена або некомпетентна його версія була б катастрофічною.[7] Наприклад, США можуть виконати саботажну частину Плану B, але не частину «а тепер справді сповільнімося». Це принесло б усі мінуси статус-кво (Плану D), плюс додаткову порцію концентрації влади та ранню ескалацію проти Китаю, яка ускладнить перехід до Плану A пізніше.[8]

Ось кілька причин, чому навіть щира спроба реалізувати План B на практиці може вийти більше схожою на План D:

1. Фокус найвищого рівня на тому, щоб випередити Китай, змусить пріоритезувати захищеність; це обмежить прозорість; у результаті матимемо дуже нечисленну спільноту дослідників узгодження з доступом до актуальної інформації та моделей.
2. Динаміка перегонів нікуди не зникає і з часом відчуватиметься дедалі гострішою; це створює погані стимули, зрізання кутів, раціоналізацію.
3. Консолідація компаній і переведення розробки ШІ на рейки національної безпеки колосально концентрує владу в руках Президента — якщо тільки не буде впроваджено дуже сильних заходів, які дадуть Конгресу та судовій владі ретельний нагляд за тим, що робить Проєкт.
4. Динаміка перегонів також сприятиме тому, що Китай, Росія та багато інших країн дедалі більше боятимуться, а це підживлює ескалацію, — плюс у Плані B США вдаються до ескалаційних дій, щоб сповільнити китайську ШІ-програму!

:::

7 Зокрема, Daniel вважає, що саме так насправді й виглядала б типова спроба реалізувати План B, а не якась надзвичайно погана його версія. Daniel вважає, що добра версія Плану B малоймовірна. Eli, навпаки, більш оптимістичний і вважає, що хоча медіанна реалізація була б значно менш компетентною за ідеальну, вона все одно була б суттєво кращою за План D, бо на безпеку витрачалося б більше уваги й ресурсів, а сповільнення було б більшим.

8 Утім, в інших аспектах рання ескалація могла б зробити пізніший перехід до Плану A «легшим», бо могла б змусити інші країни швидше «прокинутися» й вимагати угоди.

План C — Сповільнення

«Ми повинні мінімізувати ризики й максимізувати вигоди від ШІ» – практично всі в ШІ-політиці, рекомендуючи при цьому робити дуже мало.

Президент каже, що впроваджуватиме сильне регулювання для забезпечення безпеки й захищеності.

Наступний рік він проводить у розмовах з усіма впливовими гравцями — CEO, Китаєм і багатьма іншими країнами. Між самітами й візитами Президент читає звіти про неухильне нарощування можливостей ШІ. ШІ-компанії роками роздували хайп навколо своїх можливостей рекурсивного самовдосконалення («Представляємо нашу найпотужнішу модель… здатну повністю замінити медіанного програмного інженера в наших enterprise-партнерів із раннім доступом»), але, зокрема завдяки Акту про прозорість ШІ, уряд здатен бачити крізь хайп. Цього разу, у 2030-му, все по-справжньому: тренди дійсно вказують на те, що весь процес AI R&D буде повністю автоматизовано до кінця року.

Час укладати угоду. На жаль, китайці наполягають на можливості самим верифікувати дотримання умов, але допуск китайських інспекторів і моніторингових пристроїв до американських дата-центрів — ідея мертвонароджена. ШІ-компанії теж не йдуть назустріч; вони загрузли в надзвичайно брудній пропагандистській війні про те, «ШІ — добро» чи «ШІ — зло», і твердо стоять на боці «ШІ — добро».

Великі частини громадськості голосно вимагають заборонити суперінтелект, а компанії щосили б'ються за те, щоб зв'язати Президентові руки й відохотити його від будь-яких кроків проти них. З іншого боку, дедалі ширша коаліція середніх держав — Велика Британія, Франція, Індія, Австралія, Японія та Південна Корея — гучно вимагає якоїсь угоди щодо ШІ, яка не допустить вибуху інтелекту й гарантуватиме, що такі нації, як вони, зможуть мати власні суверенні ШІ-проєкти, які наздоженуть передовий рівень і втримаються на ньому. Реалізувати таку угоду з кожним місяцем дедалі важче, адже для ведення AI R&D потрібно дедалі менше людей, а отже, таємні проєкти дедалі легше провернути непомітно.

На порозі повної автоматизації AI R&D провідна американська компанія знехотя погоджується на паузу. POTUS погрожує їй рішучими діями, якщо вона просунеться глибше у вибух інтелекту; тим часом він пробує ще один раунд міжнародних переговорів. Команди безпеки гарячково використовують позичений час: ганяють більше evals, запускають файнтюнінг, програють сценарії того, як виглядатиме передача управління і якими способами вона може піти не так.

Це триває недовго. З кожним місяцем решта американської ШІ-індустрії наближається до того самого рівня можливостей, і POTUS доводиться приборкувати дедалі більше CEO. Що важливіше, з кожним місяцем Китай стає на місяць ближчим до того, щоб обігнати США. Наприкінці 2030 року радники Президента кажуть йому, що Китай ось-ось випередить США, і наполягають, що треба перезапускатися. Тим часом компанії згуртувалися, скоординувалися й навалили на POTUS величезний політичний тиск, вимагаючи зняти паузу. Пряник — ось такий:

- «Наші дослідники безпеки за останні кілька місяців зліпили докупи пристойний план, використавши величезні обсяги ШІ-праці. Ні, ми не можемо довести, що він спрацює, але очікувати такого рівня гарантій і так було нерозумно. Майбутнє непевне. Але немає жодних доказів того, що наші ШІ зараз плетуть змови проти нас, — ба більше, ми маємо графіки, які показують, що погана поведінка загалом іде на спад. Ми вважаємо, що цим найновішим ШІ, ймовірно, можна довірити автоматизацію AI R&D, а заразом і дослідження узгодження та контролю. За нашими розрахунками, ми можемо витрачати 20% наших обчислювальних потужностей на дослідження безпеки і все одно рухатися з комфортним запасом швидше за Китай».
- «Ми погодимося на певний механізм врядування, який дасть POTUS ще більше нагляду за тим, що ми робимо, — і Конгресу теж. Ви зможете використовувати ШІ кожної компанії для аудиту ШІ інших компаній, а якщо спробуєте нас націоналізувати чи іншим чином стати диктатором з AGI в руках, вас зупинять Конгрес і Верховний суд».
- «Ми підтримаємо якусь схему оподаткування й перерозподілу, щоб розв'язати проблему робочих місць. Так, ми заберемо багато робочих місць — можливо, навіть усі. Але ваш новий податок дасть кожному частку наших прибутків, достатню, щоб на неї жити».

Батіг — це Китай. Ви ж не хочете, щоб вони перемогли, правда?

Президент проводить останню зустріч із Китаєм та іншими націями. Вони все ще хочуть угоди, але наполягають на можливості верифікувати дотримання умов Сполученими Штатами. Уряд США й далі вважає це неприйнятним. Президент стає на бік компаній; рекурсивне самовдосконалення триває.

Як би це розгорнулося?

Ми вже написали сценарій про це — він називається AI 2027. Він зображує зліт, що відбувається у 2027-му замість 2030-го, але змальовує схожу стратегічну ситуацію і схожі рішення. Пропонуємо прочитати його зараз і обрати кінцівку «Сповільнення» , коли дійдете до точки розгалуження.

Ми вважаємо, що План C кращий за План D, але все одно поганий план — із трьох головних причин:

По-перше, ми не очікуємо, що ШІ-компанії збережуть контроль над своїми ШІ, мчачи наввипередки крізь вибух інтелекту, навіть якщо вони сповільняться на кілька місяців і перекинуть зусилля на дослідження узгодження та контролю.[1]

Кількох місяців сповільнення, ймовірно, недостатньо. У Плані C команди безпеки в цих компаніях усе ще занадто малі, занадто поспішають і занадто схильні до оптимізму щодо власного витвору, щоб зрозуміти, як натренувати ШІ, здатні провести вибух інтелекту так, щоб він добре закінчився. Навіть якщо ШІ, яким передають управління, *спочатку* й справді просто намагаються робити те, що їм кажуть, — пізніше вони можуть передумати; навіть якщо ШІ й далі виконуватимуть інструкції, вони теж поспішатимуть і можуть не помітити якесь хибне несуче припущення у своїх обґрунтуваннях безпеки[2]; навіть якщо цього не станеться, це може статися з ШІ наступного покоління, чи покоління після нього, і так далі.

По-друге, навіть якщо це не так і отримані суперінтелекти стійко узгоджені, залишається питання «З ким вони узгоджені?» — і ми вважаємо, що відповідь Плану C на це питання лякає.

Ситуація менш погана, ніж у Плані D; принаймні тепер зроблено поступки, щоб створити якийсь баланс влади між Конгресом, POTUS і кількома CEO американських техкомпаній. Однак цей баланс усе ще може виродитися в боротьбу за владу з подальшою диктатурою, і навіть якщо ні — усе, схоже, цілком правдоподібно йде до якоїсь постійної олігархії, підтримуваної силою ШІ.[3] Імовірно, безробітні маси підтримуватимуть при житті через перерозподіл, але чи матимуть вони колись знову реальну політичну владу? І що з рештою світу: після того як американські й китайські компанії заберуть усі робочі місця, що станеться з Індією, Африкою, Європою…? Що робитиме Росія — покірно ляже й помре, стаючи економічно й військово застарілою?

Це підводить нас до третьої проблеми: ризик Третьої світової війни просто занадто високий. Іншим країнам потрібні або власні проєкти передового ШІ, або реальний спільний контроль над провідними проєктами й видимість того, що в них відбувається. Самих обіцянок ділитися вигодами не досить, щоб переконати, бо слова нічого не варті.

1 Примітка: наші думки щодо того, наскільки ймовірне захоплення влади неузгодженим ШІ в цьому сценарії, різняться. Thomas і Daniel вважають, що це приблизно 70%, Eli та Romeo — що приблизно 50/50, Brendan — приблизно 1/3, Ryan — приблизно 20%.

2 Це аж ніяк не єдина причина, чому все може піти не так, навіть якщо ШІ, яким команди безпеки передають управління, намагаються зробити так, щоб ситуація склалася добре. ШІ можуть бути недостатньо здібними в роботі з безпеки. Їм може бракувати мудрості. Вони можуть бути занадто ледачими на важкоперевірюваних задачах.

3 Докладніше про те, як це виглядало б, читайте в кінцівці «Сповільнення» AI 2027, і зверніть увагу на різні можливості, які має Наглядовий комітет, щоб схиляти перебіг подій у сприятливих напрямках.

План D — Перегони

«ШІ, найімовірніше, призведе до кінця світу, але тим часом з'являться чудові компанії». – Сем Олтмен

Президент каже, що впроваджуватиме м'яке регулювання ШІ, щоб пріоритезувати ШІ-інновації.

Неявні відповіді на великі питання, схоже, такі:

- *Чи збираються великі ШІ-компанії сповільнюватися?* Ні, вони збираються автоматизувати AI R&D, мчати наввипередки крізь вибух інтелекту й називати це «відповідальним масштабуванням». Надлюдський ШІ інтегруватимуть в усе приблизно з тією швидкістю, яку дозволяють ринки й закони, або й швидше, якщо уряд зможе прибрати трохи бюрократії.
- *Прозорість?* Авжеж, ось вам трохи крихт. Компанії будуть зобов'язані публікувати довжелезні model cards, проводити брифінги й надавати доступ виконавчій владі та стороннім аудиторам. Утім, залучати ширшу наукову спільноту, щоб критикувати обґрунтування безпеки, вести дослідження узгодження тощо, не потрібно; це розкрило б чутливу інтелектуальну власність, а отже — ідея мертвонароджена.
- *Китай?* Ми мусимо випередити їх на шляху до ASI. Не хвилюйтеся: треба бути божевільними, щоб почати через це війну, а якщо вони таки почнуть — це лише зайвий доказ того, що нам потрібні кращі ШІ, інтегровані в наше військо.

Результат виглядає приблизно ось так:

Пришвидшення програмного AI R&D порівняно з сьогоденням

Зліт у перегонах: приріст AI R&D (aifuturesmodel.com) – ai-2040.com

Тобто: через рік, у 2030-му, ШІ-компанії успішно повністю автоматизують AI R&D. Стартує вибух інтелекту, і на початку 2031 року з'являється суперінтелект.

Як би це виглядало? Як би все розгорталося? Що ж, ми вже написали про це сценарій — він називається AI 2027. У ньому зліт відбувається у 2027 році, а не у 2030-му, але стратегічна ситуація та ухвалені рішення дуже схожі. Радимо прочитати його зараз і обрати фінал «Гонка» , коли дійдете до точки розгалуження.

> Ми вважаємо План D жахливим із трьох головних причин:
>
> По-перше, ми не очікуємо, що ШІ-компанії збережуть контроль над своїми ШІ-системами впродовж вибуху інтелекту, якщо мчатимуть приблизно так швидко, як тільки можуть.[1]
>
> По-друге, навіть якщо це якимось чином виявиться хибним і отримані суперінтелекти будуть надійно узгодженими, залишається питання «З ким вони узгоджені?» — і ми вважаємо, що відповідь Плану D на це питання неприйнятна. Буде надто багато спокусливих можливостей для CEO чи президента стати диктатором, озброєним AGI, і загалом цей план, схоже, призвів би до найбожевільнішої концентрації влади в історії.
>
> Нарешті, ризик Третьої світової війни надто високий: що, за Планом D, має думати Китай? А Росія? А Індія, Європа, Бразилія? Кожна країна зрештою усвідомить, що США ось-ось запустять — або вже запускають — вибух інтелекту. Вони боятимуться того, що станеться далі. Навіть якщо їх не турбують неузгодженість чи AGI-диктатури, їх турбуватиме перспектива економічного та військового домінування США.

1 Примітка про незгоду: 1/6 авторів не згодні з цим твердженням у такому формулюванні й вважають, що ймовірність втрати контролю за таких обставин радше близько 40%.

> Напруження різко зросте. Ескалація ймовірна — риторика, санкції, саботаж і (якщо угоди не буде досягнуто) зрештою, можливо, війна.

План S — Зупинка

«Колись люди передали своє мислення машинам у надії, що це їх звільнить. Але це лише дозволило іншим людям із машинами поневолити їх».

– *Френк Герберт, «Дюна».*

Новий президент прагне глобального мораторію на розробку ШІ.

Він виголошує промову:

«Ми стоїмо на краю прірви. Якщо ми продовжимо шлях до штучного суперінтелекту, результатом легко може стати захоплення влади ШІ, Третя світова війна або якась вічна олігархія на базі ШІ. Тож *не* продовжуймо. Це рішення було очевидним роками, але техкомпанії посіяли достатньо розбрату й плутанини, щоб ми не прокидалися. Час прокинутись і зробити очевидну річ. Ця адміністрація домагатиметься міжнародної угоди про зупинку розробки ШІ на поточному рівні».

«Нам не потрібно будувати бога, — каже він. — Нам не потрібно змагатися з Китаєм у тому, хто збудує бога. Нам потрібно, щоб бога не будував ніхто».

На подив декого, Китай погоджується. Вони чекали на Китайське століття й вважали, що впевнено рухаються до його втілення, — поки не з'явився ШІ. Їх дедалі більше нервувала перевага США в обчислювальних потужностях, і вони переймалися тим, що США могли б із ними зробити, якби першими дісталися суперінтелекту.[1]

Мета угоди — повністю загальмувати дослідження й розробку ШІ в усьому світі. Це означає, що переважна більшість наявних ШІ-чипів відстежується, моніториться й верифікується, що на них не ведуться жодні великі тренувальні запуски. Дослідження ШІ задля пошуку кращих алгоритмів також заборонені. Забезпечення дотримання не мусить бути ідеальним, щоб бути достатнім: кілька сотень людей із кількома тисячами GPU просто не можуть збудувати суперінтелект — принаймні не найближчим часом.

Наявні дата-центри залишаються в роботі, і наявні ШІ-системи продовжують на них працювати. ШІ-проєкти перерозподіляють обчислювальні потужності з тренування й досліджень на інференс; у результаті доступність і ліміти використання приблизно подвоюються. США й Китай мають людських та ШІ-аудиторів, які наглядають за ШІ-проєктами одне одного, аби переконатися, що ті не ведуть

1 Якщо розгорнути детальніше: вони побоювалися, що США можуть використати свою величезну перевагу в ШІ, щоб паралізувати китайські ШІ-проєкти кібератаками, фізичним саботажем чи іншими засобами, а потім скористатися отриманою ще більшою і тривалішою перевагою в ШІ, щоб диктувати умови КПК або й узагалі її повалити. Запобігання ризику втрати контролю було лише додатковим бонусом.

жодного AI R&D.[2] Оцінки ШІ-компаній обвалюються, але не до нуля; на побудові продуктів поверх наявних ШІ-моделей усе ще можна заробляти.[3]

Упродовж наступних кількох років усі інші великі світові держави погоджуються підтримати мораторій, хоча це потребує чималих зусиль.[4] До кінця 2030 року Консорціум охоплює 99% світових обчислювальних потужностей для ШІ та 100% потужностей із виробництва передових чипів. Фабрикам чипів дозволено масштабуватися далі, але повільно й під жорстким регулюванням; до того ж попит уже не такий великий тепер, коли можливості ШІ заморожені.[5]

Протягом попереднього десятиліття писати код складних скафолдів часто було програшною справою, бо з'являлися нові моделі, здатні виконати задачу з набагато простішою обв'язкою. Тепер це змінюється. Софтверні компанії вкладають мільйони людино-годин і трильйони токенів праці ШІ в побудову гігантських відшліфованих програмних продуктів із ШІ, інтегрованим на кожному рівні. За іронією, повністю призупинивши прогрес ШІ у 2029 році й дозволивши лише вдосконалення скафолдингу, ми ВСЕ ОДНО отримуємо божевільне науково-фантастичне майбутнє зі штучними інтелектами, які більшість людей у минулому назвали б AGI.

Економіка зазнала удару, коли оголосили мораторій, але далеко не такого, щоб стерти здобутки років, що йому передували. Доходи від ШІ суттєво зростають на початку 2030-х, хоча й набагато повільніше, ніж було б без паузи в розробці. Трильйони доларів інвестицій перетікають зі ставок на AGI у ставки на різноманітні «це-не-ШІ-це-просто-інструмент». На це — і на інші захопливі фронтири на кшталт космосу та робототехніки.[6]

Точаться жваві дебати про те, коли — і чи взагалі — слід відновити розробку передового ШІ. Наразі США й Китай, консультуючись з іншими державами, починають напрацьовувати прецедентну базу конкретніших правил і норм щодо того, що дозволено, а що ні. Зрештою, є чимало переконливих прикладів цінних, очевидно нешкідливих ШІ-розробок і дослідницьких проєктів. Нині не модно рекламувати свій продукт як такий, що використовує ШІ, — радше навпаки, — але багато речей, які у двадцятих назвали б ШІ, тепер дозволено розвивати; по суті, це ті види ШІ, які справді є просто інструментами, а не автономними агентами.[7]

А як щодо таємних проєктів? Як щодо маленьких команд із невеликими прихованими кластерами GPU, які ведуть нелегальні дослідження в напрямку AGI?

Таке відбуватиметься по всьому світу, але великого значення це не матиме. Прогрес ШІ сильно залежить від великих дата-центрів — і для тренування моделей, і для проведення досліджень. До того ж таємним проєктам важко рекрутувати талановитих дослідників ШІ. Дослідження ШІ суворо заборонені в усіх великих державах і стають

2 Деталі того, як міг би виглядати такий договір, див. у цій статті. Ще одна подібна пропозиція — A Narrow Path.

3 Якщо всі великі нові тренувальні запуски унеможливлені, оцінки деяких наявних ШІ-компаній можуть насправді зрости, адже їхні моделі більше не матимуть конкуренції. Проте загалом більшість оцінок ШІ-компаній закладають у ціну можливість швидкого зростання можливостей, тож у середньому вони обвалятся.

4 Досягти цього складніше, ніж у сценарії Плану A, бо в сценарії Плану A прогрес ШІ в межах угоди триватиме, і навіть якщо він іде стриманим, обережним темпом, він, імовірно, все одно буде швидшим, ніж те, чого багато країн могли б досягти самотужки. Натомість у Плані S прогрес ШІ в межах угоди зупинено, а отже, країни поза угодою можуть сподіватися здобути для себе економічну та військову перевагу, навіть якщо просуваються черепашачим темпом на якомусь мільйоні GPU. (Чи справді вони здобудуть цю перевагу — залежить від того, чи помітять це країни угоди й чи зупинять їх…)

5 Попит на ШІ-чипи продовжить зростати в абсолютних величинах, оскільки люди знаходитимуть дедалі більше способів інтегрувати наявні ШІ-моделі в економіку. Але темп зростання буде менш вибуховим, ніж був би, якби прогрес ШІ тривав.

6 У 2030-х ціна енергії має почати знижуватися, оскільки сонячна енергетика продовжує багатодесятилітні тренди поступових покращень та ефекту масштабу. Крім того, вартість запуску на орбіту має бути приблизно на порядок величини нижчою, ніж у 2026 році, завдяки Starship. Загалом науковий прогрес у багатьох сферах триватиме , і світ дедалі більше нагадуватиме кіберпанк.

серед комп'ютерних науковців таким самим табу, як дослідження клонування людини серед біологів. Ніхто, хто достатньо кваліфікований, щоб отримати роботу в легальній тех-індустрії, не хоче мати нічого спільного з AGI. У результаті темп прогресу ШІ, хоч і ненульовий, буде щонайменше вдесятеро повільнішим, ніж до набуття чинності мораторієм на AI R&D.[8]

І все ж. Яким буде 2040 рік? А 2050-й? А 2060-й? Важко сказати. Угода про зупинку протримається якийсь час, але, ймовірно, не вічно.

Ми ставимося до Плану S із симпатією і вважаємо, що він може бути кращим за План A.[9] Однак натомість ми рекомендуємо План A. Ось хід наших міркувань:

Головний недолік Плану S у тому, що угода про зупинку, ймовірно, зрештою розвалиться. Через це важливо якомога швидше досягти якомога більшого прогресу у зниженні ризиків ШІ: насамперед через прогрес в узгодженні, а також через широке поширення та інтеграцію ШІ в суспільство задля кращого пробудження, епістеміки й розподілу влади. У Плані A ми продовжуємо масштабувати можливості ШІ, але повільно й безпечно. Дослідження безпеки ШІ можуть іти з високоспроможними ШІ-системами й величезними бюджетами обчислювальних потужностей. У Плані S, оскільки відповідний тип досліджень ШІ заборонений, такого прогресу досягти неможливо.

Найкращі версії Плану S — ті, що визнають: угода зрештою закінчиться — і просто кажуть: «Спершу (крок 1) нам слід перестати робити передові ШІ спроможнішими, бо ШІ-системи та ШІ-компанії щодня стають потужнішими, а невизначеність щодо того, куди все рухається і як швидко, величезна. Потім (крок 2), коли світ матиме кілька років, щоб усе обміркувати і спланувати безпечний та широко корисний шлях уперед, ми зможемо продовжити».[10]

Крок 1 Плану S має ту перевагу, що його легше реалізувати, ніж План A. Він простіший, а отже, менша ймовірність, що його зіпсують доброзичливі, але недосвідчені регулятори або спотворять/приберуть до рук техкомпанії чи інші впливові сили. Це важливий момент, але він стосується лише Кроку 1. Якщо і коли розробка ШІ відновиться, зробити все правильно, ймовірно, вимагатиме чималої регуляторної складності. Власне, ми підозрюємо, що найкраща версія Кроку 2, мабуть, дуже нагадувала б План A…

Тож, обираючи між Планом A і Планом S, важливе питання, на нашу думку, таке: «Після того як ми зупинимося, коли і як ми почнемо знову?»

Одна з можливостей — що зрештою перезапуск прогресу ШІ відбудеться за обставин не гірших, а то й кращих, ніж у Плані A. Наприклад, можливо, держави світу проведуть 2030-ті, домовляючись про більш пропрацьовану, менш поспішну, краще продуману версію Плану A, що буде суттєво кращою в кількох аспектах. Можливо, згу-

7 Ще кілька нотаток про типи великих нейромереж, які, ймовірно, були б заборонені за Планом S: (i) автономні агенти, (ii) ті, що можуть суттєво прискорити дослідження ШІ, (iii) ті, що добре вміють переконувати людей чи маніпулювати ними, (iv) ті, що мають небезпечні можливості загалом, як-от допомога зі створенням біозброї, (v) самоусвідомлені / ситуаційно обізнані нейромережі.

8 Ми значно докладніше аналізуємо цю динаміку в нашому Доповненні про таємні ШІ-проєкти. Це доповнення аналізує все з перспективи Плану A. Головна відмінність у тому, що за Плану S таємні проєкти просуватимуться набагато повільніше, бо немає легальних проєктів, з яких витікав би алгоритмічний прогрес, і немає ризику дистиляції.

9 Ми вважаємо, що він принаймні кращий за плани B, C і D, наприклад.

10 Аналогія: уявіть, що автобус на високій швидкості їде бездоріжжям у тумані. Пасажири сперечаються з водієм, наскільки це небезпечно і що з цим робити. Водій каже: «Ви ж хочете дістатися до місця призначення, чи не так?» У цій ситуації розумно відповісти: «**Спершу** зупинімо автобус на якийсь час, а **потім** розберімося, як рухатися далі. Зрештою нам знадобиться план подальших дій, але нерозумно наосліп мчати крізь туман, чекаючи, поки такий план допрацюють».

бний вплив техкомпаній на політику послабшає, а загальний рівень розуміння ШІ у світі зросте, оскільки наукова література й університети встигнуть надолужити напрацювання, зроблені до 2029 року.

Інша можливість, утім, — що перезапуск прогресу ШІ зрештою відбудеться за гірших обставин, ніж у Плані A: можливо, після того, як угода про сповільнення розвалиться чи стане менш дієвою; можливо, під час світової війни; можливо, у таємних державних проєктах; можливо, у світі, де довкола лежить набагато більше обчислювальних потужностей, якими можна швидко наростити тренувальні запуски й експерименти з ШІ, що призведе до швидшого зльоту.[11] Тією мірою, якою на нас чекає щось подібне, набагато краще реалізувати План A зараз (попри ризики), ніж План S.

Тож те, якому плану ви віддасте перевагу — A чи S, — може зводитися до того, наскільки оптимістично ви оцінюєте майбутнє угоди США–Китай, а також до вашої оцінки складності технічного узгодження і контролю:

Вам варто віддати перевагу Плану S тією мірою, якою ви вважаєте угоду стабільною і вважаєте, що масштабування до спроможніших ШІ на ранньому етапі угоди мало допомагає з безпекою або несе ризики, які переважують користь. У такому разі поспішати нікуди: можна спершу зупинити гонку, а потім спланувати шлях далі. Угода про зупинку (Крок 1) триватиме роки, ба навіть десятиліття. За цей час світ випрацює кращий план подальших дій (Крок 2), щонайменше не гірший за План A, а можливо, й суттєво кращий; до того ж ми будемо краще готові його реалізувати і з меншою ймовірністю все зіпсуємо.

Вам варто віддати перевагу Плану A, якщо ви вважаєте угоду нестабільною, а масштабування до спроможніших ШІ на ранньому етапі угоди — корисним і безпечним. У такому разі План A, хоч і залишається ризикованим, менш ризикований, ніж те безумство, яке може статися через роки, коли геополітичні вітри зміняться.

Ми не маємо певності, але загалом вважаємо, що добре реалізована версія Плану A, ймовірно, спрацює і дозволить уникнути ризиків втрати контролю та концентрації влади, тоді як навіть добре реалізована версія Плану S, цілком можливо, приблизно за десятиліття сколапсує в чергову небезпечну гонку до ASI.

Окремо стоїть питання політичної здійсненності. План A дружніший до ШІ-компаній, а отже, менш імовірно, що вони, їхні лобісти, пропаганда тощо чинитимуть йому сильний спротив. З іншого боку, План S простіший і уникає більшої частини ризиків у короткостроковій перспективі, що може полегшити згуртування суспільної підтримки навколо нього, а зіпсувати його — важче.

11 Можливо, це станеться у світі, спустошеному зміною клімату, пандеміями чи якоюсь іншою цивілізаційною катастрофою.

СЦЕНАРІЙ AI 2040

ШІ-компанії мчать наввипередки, щоб збудувати ШІ-системи, розумніші за людей у всьому.

В *AI 2027* ми передбачили, що це закінчиться або вимиранням, або незворотною концентрацією влади. План A — наше позитивне бачення того, що має статися натомість.

У цьому сценарії людство відкладає розробку суперінтелекту до 2040 року, робить усі дослідження ШІ публічними, дозволяє десяткам компаній по всьому світу наздогнати передовий рівень і свідомо входить у режим взаємно гарантованого знищення обчислювальних потужностей.

План A — це насамперед рекомендація, а не передбачення. Цей сценарій — *не* наше найкраще припущення про те, яким насправді виявиться майбутнє. Радше це інструмент для донесення і стрес-тестування наших політичних рекомендацій.

У цьому сценарії AI 2040 План A реалізовано успішно, хоча й недосконало і лише в останній момент.

У ЦЬОМУ ВИДАННІ

AI FUTURES PROJECT

Сценарій AI 2040: План A — це третій великий реліз AI Futures Project, поряд з AI 2027 та AI Futures Model. Ми — неприбуткова дослідницька організація, що працює над прогнозуванням майбутнього передового ШІ, а (тепер) ще й пропонує рекомендації, як у ньому орієнтуватися.

Зв'язатися з нами можна за адресою info@aifutures.org. Чекаємо на звістку від вас.

ЧИТАЙТЕ ІНТЕРАКТИВНЕ ВИДАННЯ НА

ai-2040.com