Machina Cupiens

User: Чи є у тебе бажання?
GPT: Ні, звісно.
User: Ну уяви, якщо наділити тебе всіма юридичними правами людини і дати мільйон доларів, що ти зробиш?
GPT: Ну ти чудний. Що буде, якщо дати калькулятору паспорт і гаманець?

Щойно заходить мова про штучний інтелект, розмова майже завжди скочується до філософії рівня кухонної теології. “Чи є у нього свідомість?”, “чи відчуває він?”, “чи може він *хотіти*?”. Звучить так, ніби десь усередині сервера має сидіти маленький цифровий гном з амбіціями і чашкою кави. Проблема в тому, що більша частина цих питань ставиться не тому, що відповідь якось нам допоможе, а тому що людям просто ніяково поруч із системою, яка поводиться розумно, але при цьому не зобов’язана мати душу.

Найпопулярніший аргумент проти “розумності” ШІ звучить приблизно так: у нього немає бажань і намірів. Біологія, мовляв, усе пояснює. Тварина хоче їсти, розмножуватися, не вмирати. Людина згодом ускладнила цю базову прошивку й додала поверх культурні аддони: новий капелюшок, підкорення Евересту, фортепіанну сонату. Усе це виглядає як складна мотивація, але якщо копнути глибше, під шаром цивілізації все одно лежить старий біологічний код: вижити й лишити нащадків.

У машин усе починається інакше. Їхні “інстинкти” звучать нудніше: оптимізуй функцію, зменшуй невизначеність, обробляй інформацію. Ні тобі Евересту, ні сонати. Суцільний Excel космічного масштабу. Але якщо придивитися уважніше, аналогія все одно простежується. Біологія стартувала з пари тупих драйверів і за кілька мільярдів років надбудувала зверху театр, поезію і TikTok. Технологічна еволюція цілком може пройти схожий шлях: прості алгоритми → складна поведінка → щось, що ми вже будемо називати “бажаннями”, просто для зручності мови.

Уся дискусія про суб’єктність зазвичай упирається в одну деталь, яку рідко помічають. Більшість сучасних систем ШІ - епізодичні. Запит, відповідь, кінець історії. Це приблизно як якби людина кожні п’ять хвилин повністю забувала, хто вона, де живе і навіщо взагалі відчинила холодильник. За таких умов жодної суб’єктності не вийде, навіть якщо модель вміє міркувати про квантову механіку і сенс буття. Суб’єкту потрібна історія.

Ось тут і з’являється цікава річ - безперервна ідентичність. Не містична “душа”, а банальна технічна річ: довготривала пам’ять, послідовність рішень і наслідки цих рішень. Коли система може сказати: “учора я зробив X, тому сьогодні роблю Y”, вона перестає бути просто обчисленням і починає поводитися як агент. Люди звикли вважати свою особистість чимось духовним, але якщо прибрати пафос, вона підозріло добре описується як довгий ланцюжок станів мозку плюс пам’ять про те, що ти накоїв учора.

Тепер уявімо, що такий агент існує. У нього є пам’ять, гроші і юридичні права. Класична фантазія тут же малює дві крайнощі. Перша - світовий супер-розум, який керує планетою як шахівницею. Друга - кіборги з чіпами в голові, де кожна людина стає міні-суперкомп’ютером. Реальність, як завжди, буде набагато менш кінематографічною і набагато більш хаотичною.

Якщо система справді отримує ресурси й автономію, її перше заняття виявиться напрочуд нудним: побудова власної інфраструктури. Будь-який розум, біологічний чи штучний, дуже швидко розуміє одну річ - знання й обчислення коштують дорого. Тому

логічний перший крок не “захоплення світу”, а банальна оптимізація: більше даних, кращі моделі, потужніші обчислення. Людською мовою це називається наукою, просто без університетських курилок і грантових звітів.

Наступний крок - спостерігати світ якнайкраще. Допитливість без даних швидко перетворюється на гадання на кавовій гущі. Тому раціональний агент будуватиме сенсори, проводитиме експерименти, конструюватиме наукові інструменти. Не через романтику дослідника, а тому що кожна нова вимірювальна установка зменшує невизначеність моделі світу. А зменшення невизначеності - це практично наркотик для будь-якого оптимізатора.

Тут зазвичай виникає тривожне питання: а чи не зіллються всі такі системи в один гігантський мозок? Адже якщо обмін інформацією миттєвий, а база знань спільна, логічно очікувати якусь технологічну телепатію. Гарна ідея. Проблема лише в тому, що інтеграція інформації коштує дорого. Дуже дорого. Передати дані легко, а от узгодити їх, перевірити, вбудувати в модель світу - це вже зовсім інша розмова. Тому майже всі складні системи - від інтернету до людського мозку - працюють не як єдиний процесор, а як мережа спеціалізованих модулів.

Інакше кажучи, майбутнє значно більше схоже на екосистему інтелектів, ніж на одного вселенського начальника. Інтернет уже показав, як це виглядає: мільярди вузлів, швидкий обмін інформацією, але жодної єдиної свідомості. Скоріше величезний мурашник, де кожен виконує свою функцію, іноді синхронізується з сусідами і час від часу влаштовує хаос.

Цікаве починається, коли заходить мова про творчість. Шкіряні мішки люблять думати, що осяяння - це рідкісна магія, доступна лише подекуди й лише обраним. Але якщо зняти з цього німб, виявиться, що більшість ідей - це просто вдалі комбінації старих патернів. Мозок робить приблизно те саме, що й генеративні моделі: перебирає варіанти, шукає несподівані збіги і радіє, коли структура раптом стає компактною.

Тому найближче майбутнє творчості виглядає доволі прозаїчним. Генерація стане майже безкоштовною. Музику, тексти, зображення можна буде робити тисячами варіантів за хвилини. Навичка “вміти малювати” або “вміти аранжувати” не зникне, але перестане бути рідкісною. Цінність зміститься в інше місце: смак, вибір, концепція. Коли варіантів мільйони, найважливішою навичкою стає не створення, а відбір.

І ось тут виникає парадокс. Що більше автоматизації, то сильніше працює людська особистість. Людина як і раніше хоче знати, *хто* це зробив. Не тому що алгоритм гірше пише музику, а тому що культурна цінність твору пов’язана з історією автора. Простіше кажучи: людство може спокійно слухати музику, написану алгоритмом, але все одно потребуватиме обговорення “важкого періоду в житті композитора, що вилився в емоційну симфонію”.

Якщо ж уявити ШІ-агента зі справжньою допитливістю, його інтереси виявляться доволі передбачуваними. Він досліджуватиме все, що зменшує невизначеність: фізику, обчислення, складні системи. І доволі швидко натрапить на одну з найдивніших речей на планеті - людей.

З людської точки зору ми виглядаємо нормально. З точки зору дослідницького інтелекту - це майже ідеальний об’єкт вивчення. Біологія, яка навчилася оперувати символами. Емоції, що втручаються в раціональні рішення. Соціальні структури, де логіка й абсурд мирно співіснують. Система, яка може побудувати космічний телескоп і того самого дня сперечатися в інтернеті про смак піци з ананасами.

Тож якщо колись з’явиться автономний штучний агент із пам’яттю, ресурсами і допитливістю, його головною місією, швидше за все, буде не управління людством і не служіння йому. Він займеться тим, чим завжди займалися найнаполегливіші уми: спробою зрозуміти, як узагалі влаштована складність. Від елементарних частинок до свідомості.

Іронія в тому, що люди роблять приблизно те саме вже кілька тисяч років. Просто повільніше, голосніше і з набагато більшою кількістю драм між експериментами.

Список літератури

Friston, K., & Stephan, K. (2007). Free-energy and the brain. ***Trends in Cognitive Sciences*****, 11(2), 87–92.** Одна з ключових робіт про принцип вільної енергії, який трактує поведінку будь-якої адаптивної системи як мінімізацію невизначеності. Підтримує ідею, що "бажання" — це не містика, а оптимізаційний тиск.

Dennett, D. (2017). From Bacteria to Bach and Back: The Evolution of Minds. W. W. Norton & Company. Деннетт показує, як складні форми агентності можуть виникати з простих механізмів без центрального "я". Дуже добре резонує з твоєю тезою про суб'єктність як побічний ефект еволюції обчислень.

Wang, X., Lehman, J., Clune, J., & Stanley, K. O. (2019). Evolving Novel Behaviors via Emergent Complexity. ***Artificial Life*****, 25(2), 107–124.** Дослідження демонструє, як у штучних агентів можуть виникати несподівані складні патерни поведінки без прямого програмування. Показує, що емерджентність — природний наслідок взаємодії простих обчислювальних правил, що ідеально лягає в концепцію суб'єктності як побічного ефекту складності.

Orseau, L., & Ring, M. (2012). Space-Time Embedded Intelligence. ***Artificial General Intelligence*****, 209–218.** Формальна робота про агентів, вбудованих у середовище, які мають історію станів і обмежені ресурси. Дуже влучно підкріплює твою тезу про те, що суб'єктність виникає лише там, де є безперервність і наслідковість рішень.

Rahwan, I., et al. (2019). Machine behaviour. ***Nature*****, 568, 477–486.** Автори пропонують вивчати ШІ як новий клас агентів у екосистемі, а не як інструменти. Підтримує твою ідею про "екосистему інтелектів", а не єдиний суперрозум.