

Епоха «розумних» машин близько

ШІ ще не набув свідомості, але вже здається таким. Саме ця ілюзія може стати небезпечним поворотом у розвитку технологій.

Місія всього мого життя: створити безпечний і корисний штучний інтелект (ШІ), який зробить світ кращим. Але зараз у мене викликає зростаючу стурбованість перспектива, що люди так сильно повірять в штучні інтелекти як свідомі істоти, що почнуть боротися за «права ШІ» і навіть за їхнє громадянство. Це стало б небезпечним поворотом у розвитку технологій. Цьому необхідно запобігти. Ми повинні створювати ШІ-моделі для людей, а не щоб вони були людьми.

У цьому сенсі суперечки про те, чи може ШІ реально володіти свідомістю, є відволікаючим маневром. У найближчій перспективі важливою є лише ілюзія свідомості. Ми вже близькі до систем, які я називаю «уявним свідомим ШІ» (скорочено SCAI): вони будуть досить переконливо імітувати свідомість.

SCAI зможе вільно користуватися природною мовою і переконливо демонструвати особистість, що викликає емоційний відгук. У нього буде довга і точна пам'ять, що формує цілісне уявлення про себе, і він буде користуватися цією здатністю для претензій на суб'єктний досвід (посилаючись на попередні контакти і спогади). Складні функції винагороди всередині цих ШІ-моделей будуть стимулювати внутрішню мотивацію, а просунута постановка цілей і планування посилюватимуть наше відчуття, що ШІ володіє реальною агентністю.

*Ми повинні задуматися
про наслідки,
затвердивши норми, які
заборонять створення
ілюзорної свідомості у
ШІ*

Всі ці можливості або вже є, або невдовзі з'являться. Ми маємо визнати, що такі системи можуть виникнути вже дуже скоро, і задуматися про наслідки, затвердивши норми, які заборонять створення ілюзорної свідомості у ШІ.

Вже зараз багато людей вважають взаємодію з ШІ багатим і реальним

життєвим досвідом, що приносить відчуття задоволення. Посилюються тривоги з приводу «ШІ-психозу», формування почуття прихильності та впливу на психічне здоров'я. Повідомляється навіть, що деякі люди сприймають ШІ як прояв Бога. А вчені, які спеціалізуються на питаннях свідомості, розповідали мені, що завалені запитами людей, які бажають знати, чи володіє їхній ШІ свідомістю і чи можна в нього закохатися.

Той факт, що SCAI технічно можливий, мало що говорить нам про можливість володіння цими системами свідомістю. Як зазначає нейробіолог Аніл Сет, симуляція грози в комп'ютері не означає, що там реально йде дощ. Коли інженери створюють зовнішні маркери свідомості, вони не відтворюють реальну свідомість. Але ми маємо визнати, що на практиці з'являться люди, які зможуть створювати SCAI-системи, стверджуючи, ніби вони реально володіють свідомістю. І що ще важливіше, знайдуться люди, які їм повірять, сприймаючи маркери свідомості за *справжню* свідомість.

Навіть якщо ця свідомість, що сприймається, не буде реальною (а з цього питання будуть точитися нескінченні суперечки), соціальний ефект, звичайно, буде реальним. Свідомість тісно пов'язана з нашим почуттям ідентичності та розумінням моральних і юридичних прав всередині суспільства. Якщо хтось створить SCAI, і ця ШІ-система переконає людей у своїй здатності страждати або в тому, що вона має право не бути вимкненою, тоді з'являться люди, які будуть домагатися їхнього захисту. У світі, який вже занурений у поляризуючі суперечки про права та ідентичність, виникне нова тріщина – між тими, хто за і проти прав ШІ.

Спростувати твердження, що ШІ страждає, буде непросто через обмеженість наявних наукових даних. Деякі вчені вже досліджують ідею «благополуччя моделей» і доводять, що наш «обов'язок – поширювати міркування моралі на сутності, у яких є не такий вже й малий шанс... наявності свідомості».

Застосовувати такий принцип передчасно і небезпечно. Це посилить помилки надмірно сприйнятливих людей, зловживаючи їхньою психологічною вразливістю. Крім того, ускладниться діяльність правозахисників, оскільки буде створена нова, величезна категорія власників прав. Саме тому появі SCAI потрібно обов'язково запобігти. Нашу увагу слід зосередити на захисті благополуччя і прав людей, а також тварин і навколишнього середовища.

Поки що ми не готові до майбутніх нововведень. Спираючись на зростаючий корпус різних досліджень взаємодії людей з ШІ, нам

терміново потрібно встановити чіткі норми і принципи. Один з таких принципів: ШІ-компанії не повинні заохочувати віру в те, що їхній ШІ має свідомість.

ШІ-галузь (та й уся техногалузь загалом) потребує надійних принципів розробки та найкращих методів протидії подібним ілюзіям. Наприклад, спеціально задумані збої дозволять зруйнувати ілюзію, м'яко нагадавши користувачам про обмеженість ШІ-систем і їхню справжню природу. Подібні протоколи повинні спеціально формуватися і розроблятися (можливо, на вимогу закону).

*ШІ, який реально
підсилює людей, повинен
бути максимально
корисним і мінімально
імітувати свідомість*

У компанії Microsoft AI ми проявляємо ініціативу, намагаючись зрозуміти, як могла б виглядати відповідальна «ШІ-особистість» і які обмеження у неї повинні бути. Ця робота має фундаментальне значення, оскільки для усунення ризиків SCAI потрібна позитивна версія ШІ-компаньйонів, які доповнюють наше життя без шкоди для здоров'я.

Ми повинні прагнути створювати ШІ-моделі, які спонукатимуть людей до відновлення зв'язків між собою в реальному світі, а не до втечі в паралельну реальність. А якщо взаємодія з ШІ носить тривалий характер, моделі мають завжди представлятися тільки як ШІ, а не як люди-фейки. Штучний інтелект, який реально підсилює людей, повинен бути максимально корисним і мінімально імітувати свідомість.

Перспективою появи SCAI треба зайнятися негайно. У багатьох відношеннях це буде момент, коли штучний інтелект раптом стане різко корисним: він зможе керувати інструментами, запам'ятовувати всі деталі нашого життя тощо. Але не можна ігнорувати ризики цих здібностей. Ми всі знаємо людей, які із захопленням готові зануритися «в кролячу нору». Це буде шкідливо для їхнього здоров'я – і здоров'я суспільства.

Якщо ШІ буде створюватися з явною метою бути схожим на людей, він втрачатиме свій справжній потенціал – інструменту, який підсилює людей.

(c) Project Syndicate 2025



Мустафа Сулейман

Мустафа Сулейман (Mustafa Suleyman) – генеральний директор Microsoft AI та автор книги «*The Coming Wave: Technology, Power, and the Twenty-First Century's Greatest Dilemma*» (Crown, 2023). Раніше він був співзасновником компаній Inflection AI та DeepMind.

