ДЕМОКРАТИЧНЕ СУСПІЛЬСТВО 23.03.2026 | Марк Фаддул

Хто шепоче на вухо вашому чат-боту?

Надаючи бізнесу владу над інфраструктурою ШІ, демократії створюють основу для переходу до техноавторитаризму. Урок соцмереж не засвоєно.

Алгоритми не є нейтральними щодо цінностей та принципів. Але вже більше десяти років ми дозволяємо найбільшим технокомпаніям використовувати їх як «контролери входу» в нашу інформаційну екосистему, не вимагаючи прозорості чи відповідальності. Серед різноманітних наслідків – поширення поляризуючого та сенсаційного контенту, прихована персоналізована реклама, часті випадки монополістичної поведінки, форми впливу на суспільні дискусії, що суперечать нормам демократичного обговорення.

На власному гіркому досвіді ми дізналися, що саме відбувається, коли критично важлива інформаційна інфраструктура віддається без жодного нагляду в руки корпорацій, але сьогодні ми повторюємо ту саму помилку з чат-ботами на базі штучного інтелекту (ШІ), і цього разу ставки можуть виявитися набагато вищими. Чат-боти не просто відбирають наявну інформацію; вони генерують та оформляють її. Якщо Facebook і Google вирішували, які новини вам показувати, то інструменти, подібні до ChatGPT, Claude і Gemini, синтезують інформацію у відповіді, які звучать авторитетно.

Якщо Facebook і Google вирішували, які новини вам показувати, то інструменти, подібні до ChatGPT, Claude і Gemini, синтезують інформацію у відповіді, які звучать авторитетно

Це важлива відмінність, тому що завдяки переходу від відбору до редагування їхній неправомірний вплив виявляється ще менш помітним і ще згубнішим. Ми віддаємо приватним корпораціям безпрецедентну владу над інформаційною інфраструктурою майбутнього, навіть не вимагаючи незалежного нагляду. Найнагальніша загроза не в тому, що ШІ-системи можуть вийти з-під контролю, а в тому, що купка егоїстичних корпорацій швидко перетвориться на інформаційних «контролерів входу» для

значної частини населення, що зростає.

Сучасні чат-боти – не просто великі мовні моделі (LLM). Вони спираються на кілька непрозорих алгоритмічних рівнів, які впливають на розробку та впровадження моделей, і на кожному з цих рівнів платформи чи інші сторони можуть втручатися, формуючи інформацію відповідно до власних інтересів.

У цьому «стеку алгоритмічного впливу» є щонайменше п'ять рівнів. Перший рівень – відбір даних на навчання. Визначаючи, які дані включати або виключати під час навчання, платформи приймають непрозорі рішення про джерела, як зважувати різні точки зору і який контент відфільтровувати. Ці рішення потім формують світогляд моделі. Наприклад, у жовтні 2025 року Ілон Маск запустив проєкт Grokipedia, щоб забезпечити навчальними даними свій чат-бот Grok. Ця корпоративна енциклопедія має на меті стати «анти-воук» альтернативою «Вікіпедії» (і її моделі управління спільнотою), яка вже давно служить джерелом інформації в інтернеті, що користується широкою довірою.

Другий рівень – навчання з підкріпленням за рахунок зворотного зв'язку з людьми та ШІ. Цей процес перетворює LLM з непередбачуваних генераторів тексту на корисних «помічників». На етапі «постнавчання» у розробці моделі рецензенти-люди оцінюють відповіді, щоб підштовхнути систему до бажаної поведінки, наприклад, бути ввічливим помічником. Поки що ці людські оцінки залишаються важливою, хоча багато в чому невидимою частиною ШІ-галузі. Але сьогодні їх дедалі частіше замінюють спеціалізовані ШІ-«вчителі», які покликані привести базову модель у відповідність до заздалегідь визначених принципів, закодованих у «конституції».

Третій рівень – вебпошук. Коли чат-боти ведуть пошук онлайн або звертаються до цифрових баз даних, системи RAG (генерація з доповненою вибіркою) визначають, яку інформацію треба включити у відповідь моделі. Ця функція копіює функцію традиційних пошукових систем, які мають пріоритети між джерелами. Як і в разі пошукових систем, включення реклами у відповіді чат-ботів (ChatGPT анонсував це на 2026 рік) викличе нові побоювання з приводу об'єктивності.

Четвертий рівень – системні промпти. Вони спрацьовують, коли чат-бот генерує відповідь, що дозволяє платформам змінювати поведінку чат-бота, не проводячи його перенавчання. Оскільки минулого року було оприлюднено системні промпти в Grok, ми знаємо, що вони включають, наприклад, такі вказівки: «Не соромся політичних

некоректних заяв». (ChatGPT, Claude та Gemini теж використовують системні промпти, але вони залишаються секретом.)

Останній рівень – фільтри безпеки. Перш ніж запит чат-бота досягне моделі, вхідні фільтри визначають, чи він «прийнятний». А коли модель згенерувала відповідь, вихідні фільтри можуть модифікувати, цензурувати або гігієнізувати контент, перш ніж ви його побачите. Хоча платформи мають легітимні причини блокувати певні запити (наприклад, прохання інструкцій для виготовлення бомби), той факт, що ці фільтри непрозорі, викликає питання. Розробники моделей можуть створити інфраструктуру для системної цензури, і ми не дізнаємося про це. Фільтри «безпеки» в китайських чат-ботах цензурують будь-які згадки про масові вбивства на площі Тяньаньмень.

Політичні маніпуляції за допомогою чат-ботів вже довели ефективність

Весь цей стек алгоритмічного впливу вже формується відповідно до політичних та корпоративних інтересів, причому рівно в той момент, коли чат-боти починають впроваджуватися в глобальному масштабі. Після другої інавгурації Дональда Трампа компанія Apple оновила інструкції навчання ШІ, виключивши звідти визначення прихильників руху MAGA як «радикальних» чи «екстремістських». Минулого літа агентство Reuters дізналося, що компанія Meta оновила внутрішні правила ШІ, пом'якшивши заходи захисту, які, серед іншого, утримували чат-боти від расистських висловлювань або від кокетливої ⌧⌧поведінки з неповнолітніми. А у травні минулого року чат-бот Grok почав поширювати необґрунтовані та вирвані з контексту заяви про «геноцид білих» у ПАР (Ілон Маск – білий південноафриканець). Компанія звалила провину на «несанкціоновані модифікації», але подібні «помилки» поширені, і всі вони явно збігаються з поглядами Маска.

Політичні маніпуляції за допомогою чат-ботів вже довели ефективність. Проведене в 2025 році дослідження журналу Nature показало, що чат-боти, навчені відстоювати інтереси конкретного кандидата, можуть з разючою легкістю змінювати погляди поміркованих і невизначених виборців (а саме ці групи вирішують

результат більшості виборів).

На відміну від авторитарних систем, які відкрито контролюють інформацію, демократії залежать від плюралізму джерел та прозорих, відповідальних інформаційних екосистем. Якщо допустити появу централізованого, непідзвітного контролю за інфраструктурою ШІ, це призведе до дрейфу до техноавторитаризму, адже легко побачити, як кожен рівень у стеку алгоритмічного впливу можна використовувати для посилення чи придушення певних поглядів без необхідності відкритої цензури.

У грудні минулого року Єврокомісія оштрафувала компанію X на 120 млн євро за «порушення зобов'язань прозорості відповідно до закону "Про цифрові послуги"». Як і слід було очікувати, X та її захисники назвали це рішення атакою на свободу слова. Але прозорість відіграє центральну роль у захисті свободи слова. Без неї ми не можемо дізнатися, кого піддають цензурі та який вплив на медіа, що передають інформацію, яку ми всі споживаємо.

Історія соцмереж показала нам, що відбувається, коли відповідальність не встигає за впровадженням технології. Ми не можемо дозволити собі повторення тих самих помилок із системами, які мають ще більшу владу над знанням суспільства.

(c) Project Syndicate 2026



Марк Фаддул

Марк Фаддул (Marc Faddoul) – технолог та дослідник у галузі штучного інтелекту, спеціаліст з систем рекомендацій та аудиту алгоритмів.