

УПРАВЛІННЯ ШТУЧНИМ ІНТЕЛЕКТОМ

Research ▾ Product ▾ Developers ▾ Safety Company ▾

Search Log in ↗ Sign up ↗

Moving AI governance forward

OpenAI and other leading labs reinforce AI safety, security and trustworthiness through voluntary commitments.



Illustration: Justin Jay Wang / iStock

У блозі OpenAI опубліковано повідомлення «Просування управління ШІ». У ньому міститься інформація, що OpenAI та інші провідні компанії – Amazon, Anthropic, Google, Inflection, Meta, Microsoft, – беруть на себе добровільні зобов'язання із посилення безпеки, захисту та надійності технологій штучного інтелекту. Таких зобов'язань вісім і вони стосуються питань безпеки та довіри. Зокрема, проголошено прихильність до внутрішнього та зовнішнього об'єднання моделей або систем для запобігання неправомірного використання, соціальних ризиків та проблем національної безпеки, таких як біологічна, кібербезпека та інші галузі безпеки. Компанії працюватимуть над обміном інформацією про ризики довіри та безпеки, спроби обійти заходи безпеки. Вони інвестуватимуть у кібербезпеку та захист від інсайдерських загроз, щоб захистити пропрієтарні та невипущені моделі. Ними всіляко підтримуватиметься мотивоване стороннє виявлення та повідомлення про проблеми і вразливості. Планується розробити й упровадити механізми, що дозволять користувачам зрозуміти, чи створено аудіо- або візуальний контент за допомогою ШІ, включаючи застосування водяних знаків. Громадськість буде поінформована про можливості моделей або систем, обмеження та

області належного чи неналежного використання, включно із обговоренням соціальних ризиків, таких як вплив на справедливість та упередженість. Пріоритетна увага буде приділена дослідженням соціальних ризиків, пов'язаних із системами ШІ, у тому числі щодо запобігання шкідливим упередженням, дискримінації та захисту конфіденційності. Планується розробляти і впроваджувати передові системи штучного інтелекту, щоб допомогти у вирішенні нагальних проблем.

Детальніше: <https://is.gd/PGcRQ6>, <https://is.gd/uViwMM>

Фото: скріншот

#НРАТ_Усі_новини #НРАТ_ШтучнийІнтелект #НРАТ_Науковцям_новини

#НРАТ_Освітянам_новини

2023-07-27

Інформація з офіційного вебпорталу Національного репозитарію академічних текстів