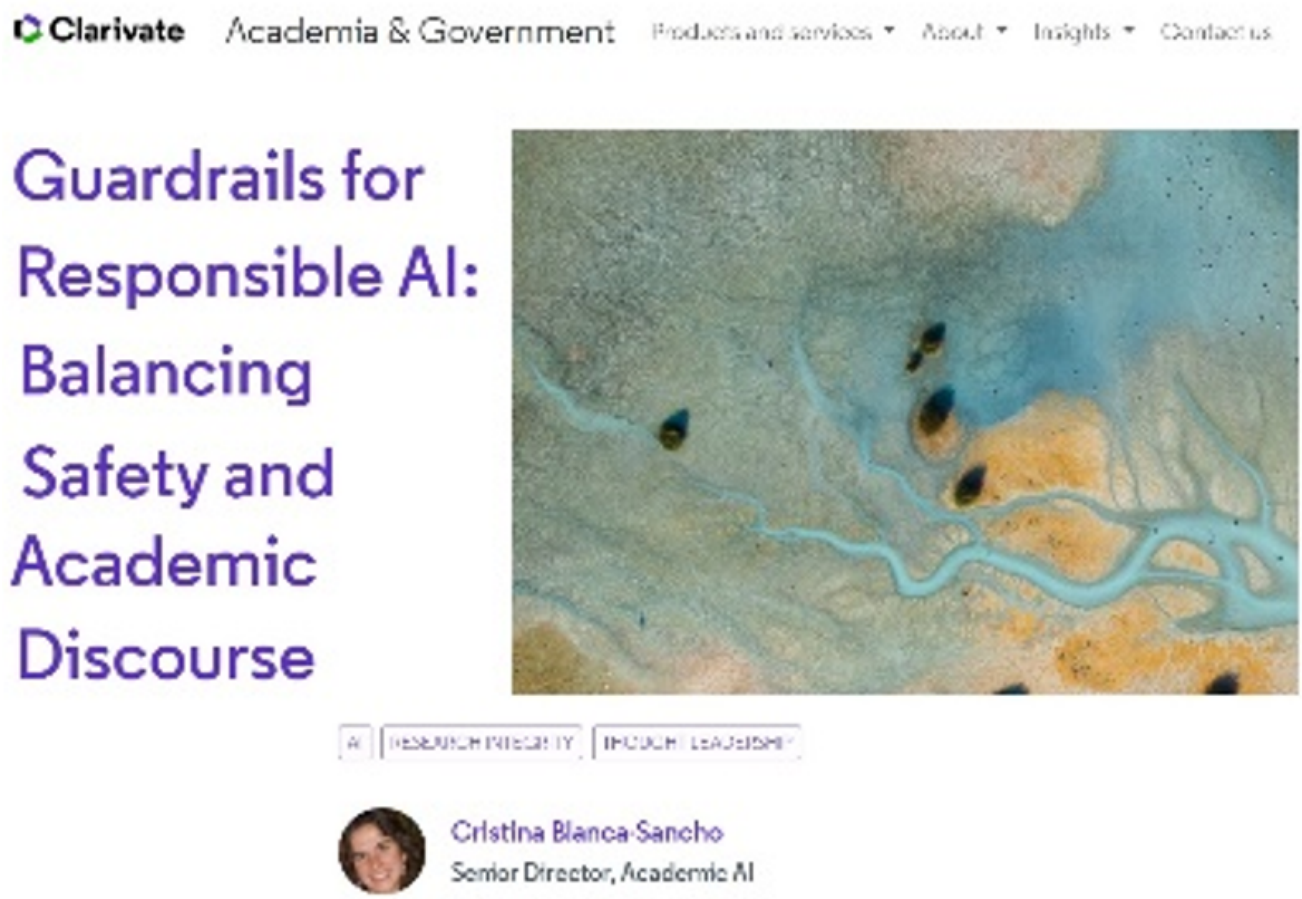
https://nrat.ukrintei.ua/zahysni-mehanizmy-vidpovidalnogo-vykorystannya-shi-balans-mizh-bezpekoyu-ta-akademichnymy-dyskusiyamy/

ЗАХИСНІ МЕХАНІЗМИ ВІДПОВІДАЛЬНОГО ВИКОРИСТАННЯ ШІ: БАЛАНС МІЖ БЕЗПЕКОЮ ТА АКАДЕМІЧНИМИ ДИСКУСІЯМИ



У блозі на сайті компанії Clarivate опублікована стаття Крістіни Бланка-Санчо «Захисні механізми відповідального використання ШІ: баланс між безпекою та академічними дискусіями».

У ній йдеться про результати дослідження щодо захисту та фільтрації контенту з використанням ШІ, які можуть сприяти безпечному та етичному застосуванню ГШІ в академічних дослідженнях без шкоди для наукової свободи. Наголошується, що по мірі упровадження ШІ в наукові процеси стає очевидною необхідність формування нових правил та норм у складі загальних та галузевих стандартів академічної доброчесності. ГШІ відкрив нові можливості для академічних досліджень, дозволяючи швидше знаходити, узагальнювати та синтезувати знання, а також підтримуючи науковий дискурс. Разом із тим з новою гостротою постає питання забезпечення відповідального використання ШІ для збереження академічної свободи та дотримання наукової сумлінності. Існують різні способи технічного вирішення цієї проблеми. Два найважливіші з них —

запобіжники (проактивні механізми, призначені для запобігання небажаній поведінці моделі) та фільтрація контенту (реактивний механізм, який оцінює вхідні дані програми та згенеровані моделлю результати, щоб визначити, чи слід їх видавати користувачу). Наразі використовуються автоматизовані моделі класифікації виявлення та блокування (або позначення) небажаного або шкідливого контенту. По суті, фільтри контенту являють собою процеси, які можуть блокувати потрапляння контенту до LLM, а також блокувати доставку відповідей LLM. Мета фільтрації контенту – виловлювати та блокувати неприйнятний контент, який може бути сформований під час процесу генерації відповіді. Проблемним є питання визначення рівня безпеки та збалансування безпеки й відкритості. Дослідники очікують, що інструменти штучного інтелекту підтримуватимуть дослідження, а не цензуруватимуть їх. Однак кожен постачальник, який використовує LLM, стикається з однаковими обмеженнями: фільтрами на рівні постачальника, регуляторними вимогами та етичним імперативом запобігання небажаному впливу.

Детальніше:
https://clarivate.com/academia-government/blog/guardrails-for-responsible-ai/

Фото: скріншот

#НРАТ_Усі_новини #НРАТ_ШтучнийІнтелект #НРАТ_АкадемДоброчесність #НРАТ_Науковцям_новини #НРАТ_Освітянам_новини

2026-05-14
