УДК 004.89

НАВЧАННЯ З ПІДКРІПЛЕННЯМ НА ОСНОВІ ЗВОРОТНОГО ЗВ'ЯЗКУ ВІД ШТУЧНОГО ІНТЕЛЕКТУ ДЛЯ ВИРІВНЮВАННЯ МОДЕЛЕЙ

Селін Я.Ю.
e-mail: yaroslav.selin@nure.ua
Науковий керівник – к.т.н., доц. Головянко М.В.
Харківський національний університет радіоелектроніки, каф. ШІ
м. Харків, Україна

This paper explores Reinforcement Learning from AI Feedback (RLAIF) as a scalable alternative to human feedback (RLHF) for aligning large language models. It examines the core mechanisms of RLAIF, particularly the Constitutional AI framework, where a superior AI evaluates outputs based on predefined ethical guidelines. Key algorithmic stages and potential challenges like reward hacking are analyzed. The findings highlight RLAIF as an efficient, cost-effective methodology for ensuring AI safety without relying on extensive human annotation.

Стрімкий розвиток великих мовних моделей і їхня активна інтеграція в різні сфери суспільного життя актуалізують питання безпеки та етичності штучного інтелекту. Виникає необхідність «вирівнювання»моделей – забезпечення їхньої роботи відповідно до людських цінностей, етичних норм та заданих правил безпеки. Традиційним підходом для вирішення цього завдання тривалий час залишалося навчання з підкріпленням на основі відгуків людей RLHF (Reinforcement Learning from Human Feedback). Проте цей метод має суттєві обмеження: він вимагає надзвичайно великих фінансових та часових витрат, залучення великих команд експертів-анотаторів, а також вразливим до людських упереджень [1].

Відповіддю на ці виклики став новий інноваційний підхід – навчання з підкріпленням на основі зворотного зв'язку від штучного інтелекту RLAIF (Reinforcement Learning from AI Feedback). Суть методу полягає в тому, що замість живих людей оцінку діям цільової моделі дає інша, зазвичай більш потужна, система штучного інтелекту. Фактично, одна модель виступає в ролі «вчителя», який тренує «учня» працювати відповідно до визначеного набору правил.

Одним із найвідоміших практичних втілень цього підходу є концепція «Конституційного штучного інтелекту», розроблена та успішно імплементована компанією Anthropic [2]. Моделі-вчителю надається своєрідна «конституція» – чітко артикульований перелік етичних настанов, базових принципів та обмежень. Під час тренування цільова модель генерує кілька варіантів відповідей на складний або провокативний запит, а модель-суддя оцінює їх виключно через призму цієї конституції. Важливою інновацією є те, що модель-суддя часто

використовує метод Chain-of-Thought, аргументуючи свій вибір перед тим, як винести остаточну оцінку. Це робить процес вирівнювання більш прозорим та інтерпретованим.

Процес RLAIF складається з генерації даних, навчання моделі винагороди та оптимізації політики цільової моделі. Моделі, натреновані таким чином, демонструють рівень безпеки та вирівнювання, який не поступається, а іноді й перевершує результати RLHF.

Використання RLAIF відкриває нові перспективи масштабування етичних систем: ітеративне покращення в реальному часі, швидку адаптацію через оновлення «конституції» та суттєве здешевлення процесу. Більш того, гібридні підходи, що поєднують невелику кількість високоякісних людських відгуків із масовим зворотнім зв'язком від штучного інтелекту, розглядаються як найбільш перспективний шлях розвитку [3]. Однак справжній прорив очікується в сценаріях, коли самі автономні системи ШІ стають дизайнерами наступних поколінь штучного інтелекту. Тут критичним стає забезпечення незмінної етичної спадковості: кожне нове покоління має автоматично успадковувати фундаментальні принципи [4].

Таким чином, навчання з підкріпленням на основі зворотного зв'язку від штучного інтелекту є логічним кроком еволюції технологій машинного навчання. Це відкриває шлях до справжньої ітеративної автономії. Завдяки делегуванню рутинного процесу оцінювання алгоритмам, дослідники та інженери отримують можливість створювати масштабовані, етичні та безпечні моделі, мінімізуючи людський фактор та прискорюючи темпи створення надійного програмного забезпечення нового покоління.

Список використаних джерел:

1. Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback. arXiv.org. URL: https://arxiv.org/abs/2307.15217 (дата звернення: 11.03.2026).

2. Constitutional AI: Harmlessness from AI Feedback. arXiv.org. URL: https://arxiv.org/abs/2212.08073 (дата звернення: 11.03.2026).

3. A Critical Evaluation of AI Feedback for Aligning Large Language Models. arXiv.org. URL: https://arxiv.org/abs/2402.12366 (дата звернення: 11.03.2026).

4. Towards ethical evolution: responsible autonomy of artificial intelligence across generations - AI and Ethics / M. Golovianko та ін. SpringerLink. URL: https://link.springer.com/article/10.1007/s43681-025-00759-9 (дата звернення: 11.03.2026).