ЛЮДИНА В КОНТУРІ КЕРУВАННЯ АГЕНТНОЮ СИСТЕМОЮ: РОЗПОДІЛ ВІДПОВІДАЛЬНОСТІ ЗА ДІЮ

Киселевич Володимир Володимирович
асистент кафедри комп'ютерних наук та інформаційних технологій
Житомирський державний університет імені Івана Франка, Україна

Постановка проблеми. Програмні системи на основі великих мовних моделей дедалі частіше отримують дозвіл на виклик зовнішніх інструментів, тобто на дії з наслідками поза межами діалогу: запис у файлову систему, надсилання повідомлень, платіжні операції, зміну конфігурації робочого середовища. Типовим проєктним рішенням для таких систем є вимога підтвердження користувачем перед незворотною дією. Це рішення подають як засіб безпеки, проте воно виконує ще одну функцію, а саме переносить відповідальність за наслідки на людину, яка санкціонувала дію. Аналіз делегування в агентних системах через модель «принципал і агент» показує, що зростання самостійності агента розмиває межі відповідальності між замовником, постачальником і оператором [1]. Наглядові документи закріплюють людський нагляд як обов'язкову вимогу до систем високого ризику [2] та як складову керування ризиками впродовж життєвого циклу системи [3]. Суть проблеми полягає в тому, що захисна функція підтвердження є умовною і зникає тоді, коли підтвердження стає формальним, тоді як функція перенесення відповідальності зберігається незалежно від змістовності перевірки. Саме ця асиметрія робить точку підтвердження засобом захисту постачальника, а не засобом захисту користувача, і саме вона потребує окремого аналізу.

Аналіз останніх досліджень. Критика політик, які вимагають людського нагляду над алгоритмами в державному секторі, виявляє систематичну невідповідність між наглядовою роллю, покладеною на людину, та її фактичною спроможністю цю роль виконувати; звідси походить пропозиція перенести основний механізм регулювання з індивідуального нагляду на інституційний і вимагати емпіричного підтвердження того, що передбачена форма нагляду справді працює [4]. Огляд упередження на користь автоматизації у співпраці людини й системи штучного інтелекту (далі ШІ) фіксує, що пояснення підвищують сприйману прийнятність системи, проте часто недостатні для підвищення точності рішень або для зменшення самого упередження, а послаблення надмірної довіри пов'язане зі зростанням витрат на самостійну перевірку [5]. Метааналіз 106 експериментальних досліджень із 370 оцінками розміру ефекту засвідчив, що комбінації людини й системи ШІ в середньому працювали гірше за найкращого з двох учасників окремо (Hedges' g = −0,23; 95-відсотковий довірчий інтервал від −0,39 до −0,07) [6]. Експеримент із написанням програмного коду показав, що учасники з доступом до ШІ-асистента створювали суттєво менш безпечний код і водночас були впевненіші в його

безпеці [7]. Польове дослідження розбору сповіщень аналітиками першої лінії в центрі керування безпекою описує, які саме відомості потрібні для обґрунтованого рішення і на яких кроках процес зривається [8]. Концептуальний перехід від наглядової моделі до моделі спільної роботи людини й системи уточнює, що контроль вимагає узгодженості поведінки обох сторін упродовж усього розгортання, а не разового акту санкціонування [9]. Наведені далі числові оцінки походять із препринтів [10; 11], тому їх слід трактувати як попередні, до завершення рецензування.

Мета. Мета роботи полягає в тому, щоб розрізнити стани контуру керування агентною системою за фактичним суб'єктом рішення, сформулювати перевірні умови, за яких підтвердження користувача є реальним контролем, а не формальністю, і подати перелік вимог до проєктування точок підтвердження.

Виклад основного матеріалу. Запропоновано розрізняти три стани контуру керування за тим, хто фактично ухвалює рішення. Стан С1 означає, що агент готує варіанти, а вибір здійснює людина. Стан С2 означає, що агент уже обрав дію, а людина санкціонує підготовлене рішення. Стан С3 означає, що агент діє самостійно, а людина спостерігає і може втрутитися згодом. Ці стани відрізняються не формальною наявністю запиту на підтвердження, а обсягом інформації, потрібним людині для змістовного контролю, і напрямом виродження: за певних умов С1 вироджується в С2, а С2 в С3, причому інтерфейс системи може залишатися незмінним, тому виродження не спостерігається безпосередньо. Характеристики станів подано в табл. 1.

Таблиця 1. Стани контуру керування агентною системою

Стан контуру	Хто ухвалює рішення	Потрібний обсяг інформації	Умова виродження в наступний стан
С1. Людина обирає	Людина	Перелік варіантів, критерії порівняння, ціна помилки	Агент подає один варіант як типовий
С2. Людина затверджує	Агент, людина санкціонує	Підстави дії, очікуваний наслідок, ознака незворотності	Темп або частота запитів перевищують межу перевірки
С3. Людина спостерігає	Агент	Журнал дій, сигнали відхилення, засоби відкату	Відкат недоступний або відхилення виявляють запізно

Джерело: розробка автора

Сформульовано умови, за яких підтвердження є реальним контролем. Умова темпу вимагає, щоб час, доступний людині на одне рішення, перевищував час, потрібний для перевірки підстав дії; за польовими спостереженнями розбір одного сповіщення передбачає послідовне звернення до кількох джерел контексту, тому цей час не є малим [8]. Умова частоти вимагає, щоб кількість запитів на одиницю часу не перевищувала порогу, за яким виникає звикання і відповідь стає автоматичною, а саме такий механізм описано як автоматизаційне упередження [5]. Умова підстав вимагає доступу людини до підстав рішення (вхідні дані, відкинуті альтернативи, очікуваний наслідок, ознака незворотності),

а не лише до формулювання дії; наявність пояснення сама собою не задовольняє цієї умови, якщо пояснення не спонукає до самостійної перевірки [5]. Умова здійсненності відмови вимагає, щоб відмова була технічно можливою і не тягла непропорційних організаційних витрат, інакше підтвердження є вибором без альтернативи. Умова компетентності вимагає, щоб людина була здатна виявити хибну дію на наданому їй матеріалі; результати щодо перевірки згенерованого коду свідчать, що така здатність не є самоочевидною [7], а середній ефект спільної роботи людини й системи може бути негативним [6].

Порушення будь-якої з п’яти умов позбавляє підтвердження захисної функції, проте зберігає його документальну функцію, бо в журналі залишається запис про санкціонування дії людиною. Асиметрія втрати двох функцій означає, що погіршення умов роботи людини (зростання темпу, збільшення кількості запитів, скорочення поданих підстав) не змінює розподілу відповідальності, а лише знижує ймовірність виявлення хибної дії, тому система з формальним підтвердженням є менш безпечною, ніж система без підтвердження при однаковому обсязі відповідальності людини. Наглядові вимоги [2; 3] встановлюють обов’язок людського нагляду, однак не задають перевірних критеріїв його змістовності, а модель спільної роботи [9] описує потрібну якість взаємодії без операційних порогів. Запропоноване розрізнення станів разом із п’ятьма умовами дає таку операційну основу: стан контуру визначають не за задекларованою схемою, а за виконанням умов темпу, частоти, підстав, здійсненності відмови й компетентності.

Практичним наслідком є перелік вимог до проєктування точок підтвердження. Перша вимога полягає в тому, щоб запитувати підтвердження лише для незворотних дій, а решту виконувати без запиту, оскільки надлишкові запити витрачають той самий ресурс уваги, від якого залежить перевірка справді небезпечних дій. Друга вимога полягає в поданні підстав і очікуваного наслідку, а не лише назви дії; обсяг і структура матеріалу, доступного перевірникові, змінюють результат перевірки, бо в контрольованому експерименті обмежувальне середовище разом із невеликим інструментом доступу до документації підвищило повноту виявлення вставлених закладок з 54,5 до 90,9 відсотка [11]. Третя вимога полягає в групуванні однотипних дій в одне рішення з переліком, що знижує частоту запитів без втрати обсягу підстав. Четверта вимога полягає в тому, щоб там, де відкат дешевший за перевірку, реалізовувати відкат і подальший контроль замість попереднього підтвердження. П’ята вимога полягає у фіксації в журналі того, хто санкціонував дію, який матеріал йому було подано і скільки часу тривало рішення, бо без цих даних неможливо відрізнити стан C1 від стану C2. Шоста вимога полягає в детермінованому застосуванні політики авторизації перед викликом інструменту, а не в покладанні рішення на саму модель: за даними препринту, під дозвільною політикою соціальна інженерія проти моделі була успішною у 74,6 відсотка випадків, тоді як під обмежувальною політикою 879 спроб дали нульову успішність; за окремим вимірюванням на тисячі рішень медіанна затримка авторизації становила 53 мс [10].

Обмеження. Робота не містить власного експерименту, тому запропоноване розрізнення станів і сформульовані умови є аналітичною конструкцією, а не емпірично підтвердженим результатом. Числові оцінки взято з чужих вимірювань, отриманих в інших предметних областях (розбір сповіщень безпеки, написання коду, перевірка закладок), і перенесення їх на точки підтвердження в агентних системах є припущенням, яке потребує окремої перевірки. Порогові значення умов темпу й частоти не визначено кількісно, оскільки вони залежать від класу дій та організаційного контексту. Статистичної перевірки запропонованих залежностей не проводилося, а частина числових оцінок походить із препринтів [10; 11] і може змінитися після рецензування. Вибірка опрацьованих джерел обмежена працями, доступними англійською та українською мовами, і не є результатом систематичного огляду. Твердження про виродження станів контуру сформульовано як умовну залежність і не супроводжується вимірюванням частоти такого виродження в реальних розгортаннях.

Висновки. Вимога підтвердження користувачем перед незворотною дією поєднує дві функції, які втрачаються неодночасно: захисна функція зникає разом зі змістовністю перевірки, а функція перенесення відповідальності зберігається завдяки самому факту запису про санкціонування. Виокремлено три стани контуру керування (людина обирає, людина затверджує, людина спостерігає) та вказано умови, за яких кожен стан вироджується в наступний без зміни інтерфейсу. Сформульовано п’ять перевірних умов реального контролю, а саме умови темпу, частоти, доступності підстав, здійсненності відмови й компетентності перевірника, та подано шість вимог до проєктування точок підтвердження. Розрізнення станів дає змогу відокремити власний результат від запозичених спостережень: числові оцінки належать наведеним джерелам, тоді як стани, умови та вимоги побудовано автором. Практичне значення полягає в тому, що відповідність наглядовим вимогам [2; 3] доцільно оцінювати за виконанням цих умов, а не за наявністю запиту на підтвердження, тому подальша робота передбачає емпіричне визначення порогів частоти й темпу для окремих класів незворотних дій.

Список літератури

1. Gabison G.A., Xian R.P. Inherent and emergent liability issues in LLM-based agentic systems: a principal-agent perspective. Proceedings of the 1st Workshop for Research on Agent Language Models (REALM 2025). 2025. P. 109–130. DOI: 10.18653/v1/2025.realm-1.9.

2. Enqvist L. ‘Human oversight’ in the EU artificial intelligence act: what, when and by whom? Law, Innovation and Technology. Vol. 15, No. 2. 2023. P. 508–535. DOI: 10.1080/17579961.2023.2245683.

3. Tabassi E. Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST (AI 100-1). 2023. DOI: 10.6028/nist.ai.100-1.

4. Green B. The flaws of policies requiring human oversight of government algorithms. Computer Law & Security Review. Vol. 45. 2022. P. 105681. DOI: 10.1016/j.clsr.2022.105681.

5. Romeo G., Conti D. Exploring automation bias in human–AI collaboration: a review and implications for explainable AI. AI & SOCIETY. Vol. 41, No. 1. 2026. P. 259–278. DOI: 10.1007/s00146-025-02422-7.

6. Vaccaro M., Almaatouq A., Malone T.W. When combinations of humans and AI are useful: A systematic review and meta-analysis. Nature Human Behaviour. Vol. 8, No. 12. 2024. P. 2293–2303. DOI: 10.1038/s41562-024-02024-1.

7. Perry N., Srivastava M., Kumar D. [та ін.] Do Users Write More Insecure Code with AI Assistants? Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security. 2023. P. 2785–2799. DOI: 10.1145/3576915.3623157.

8. Kersten L., Beelen K., Zambon E. [та ін.] A Field Study to Uncover and a Tool to Support the Alert Investigation Process of Tier-1 Analysts. Proceedings 2025 Symposium on Usable Security. 2025. DOI: 10.14722/usec.2025.23034.

9. Tsamados A., Floridi L., Taddeo M. Human control of AI systems: from supervision to teaming. AI and Ethics. Vol. 5, No. 2. 2025. P. 1535–1548. DOI: 10.1007/s43681-024-00489-4.

10. Uchi U. Before the Tool Call: Deterministic Pre-Action Authorization for Autonomous AI Agents : препринт. arXiv:2603.20953. 2026. URL: https://arxiv.org/abs/2603.20953.

11. Winninger T. Steerability via constraints: a substrate for scalable oversight of coding agents : препринт. arXiv:2607.02389. 2026. URL: https://arxiv.org/abs/2607.02389.