АНАЛІЗ ОСОБЛИВОСТЕЙ КОНФЛІКТІВ СИСТЕМ ШТУЧНОГО ІНТЕЛЕКТУ

Ланде Д. В.[1], Даник Ю. Г. [1], Сварник Н. Б. [1]

[1]*Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського»*

У роботі здійснюється аналіз типів конфліктів, що виникають або потенційно можуть виникнути у системах ШІ, зокрема у межах одної системи, взаємодії з людиною та іншими системами ШІ. Розглядаються такі конфлікти як упередженість моделей, так звані “галюцинації”, суперечності в параметричних і контекстних знаннях LLM, конфлікти цілей, спостереження, інтерпретації дій та інші. Окрема увага приділяється аналізу футурологічних прогнозів та наукових джерелах пов’язаних із потенційною траєкторією розвитку ШІ. Пропонується класифікація потенційно небезпечних сценаріїв розвитку ШІ та їх основних ознак. Матеріал може бути корисний для розробників, дослідників у сфері безпеки та етики ШІ, а також використаний для виявлення конфліктів у системах ШІ на ранніх стадіях.

Ключові слова: футурологія, конфлікти ШІ, методика, класифікація, ознаки, людино-орієнтований ШІ, галюцинації, упередженість, подвійне використання, конфлікт знань, автономні системи, масове спостереження, інформаційна війна, AGI, Artificial Superintelligence

Вступ

Стрімкий розвиток систем штучного інтелекту (ШІ) не лише відкриває нові можливості, але й породжує складні виклики, серед яких ключовими є конфлікти, що виникають як у межах самих систем, так і у взаємодії з людиною або іншими ШІ. В умовах інтеграції ШІ в чутливі

сфери – від медицини до безпеки – стає вкрай важливим вчасно ідентифікувати потенційні загрози та їх ознаки. Це дозволить забезпечити більшу стабільність та надійність систем ШІ у критичних сферах.

Огляд стану питання

У сучасних дослідженнях у сфері штучного інтелекту значна увага приділяється вивченню конфліктів, що виникають під час між ШІ та людьми, так і всередині самих систем. Зокрема, у роботі [1] розглядається поведінка автономних систем ШІ у конфліктних ситуація. У [2] проводиться аналіз ризиків ескалації, пов'язані з використанням мовних моделей у військових та дипломатичних рішеннях. У дослідженні [3] розглядається, як ШІ може бути використаний для визначення та вирішення конфліктів у міжособистісних стосунках, робочому середовищі та публічних дискусіях. Для більш глибокого розуміння конфліктів у системах ШІ потрібно створити узагальнену класифікацію типів конфліктів та виявити їх основні ознаки.

Мета дослідження полягає у створенні методики виявлення та аналізу конфліктів у системах штучного інтелекту, які дозволяють прогнозувати їх виникнення, розвиток і вирішення.

Основний зміст

Аналогічно до еволюції людського суспільства, де розвиток технологій і культури неминуче супроводжується конфліктами через зіткнення інтересів, цінностей або ресурсів – із розвитком штучного інтелекту конфлікти також стануть його невід'ємною складовою. У міру того, як ШІ дедалі глибше інтегрується в повсякденне життя, приймаючи рішення, що раніше належали людині, неминуче виникають ситуації зіткнення між очікуваннями користувача, інтерпретацією запитів, ресурсними або алгоритмічними обмеженнями.

1 Конфлікти в межах однієї системи ШІ

1.1 Галюцинації

Одним із найбільш поширених конфліктів у ШІ є, так звані «галюцинації». Під галюцинаціями розуміємо явище, коли генеративні моделі надають таки відповіді, що містять "вигадану" інформацію, тобто неправдиву або непідкріплену реальними фактами, вигадані джерела тощо. Хоча у побутових ситуаціях даною проблемою можна знехтувати, проте у бізнес-сфері чи медицині цей конфлікт може мати катастрофічні наслідки.

Зростаючий інтерес до цієї проблематики можна побачити на рис. 1 [4].

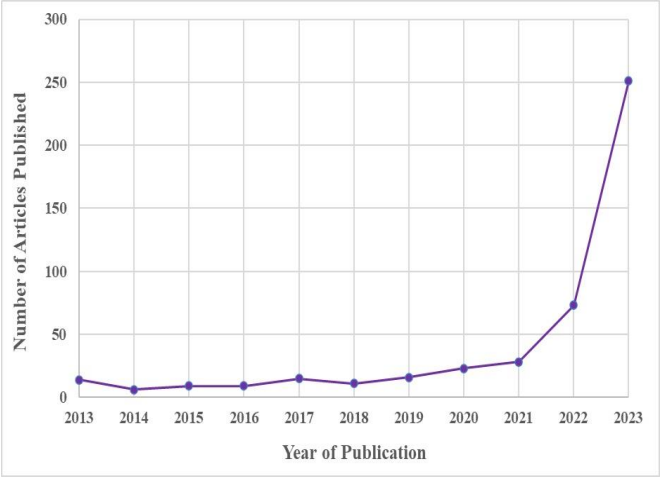

Рисунок 1. Кількість публікацій (з 2013 по 2023 рік) у базі SCOPUS, що містять поняття "галюцинації" ТА ("обробка природної мови" АБО "ШІ")

1.2 Упередженість

Упередженість у моделях машинного навчання виникає, коли алгоритм систематично відтворює та поширює нерівність або помилки, закладені в тренувальних даних або архітектурі. Такий конфлікт проявляється у формі

спотвореного пріоритету одних результатів над іншими, що не відповідає реальному розподілу або очікуванням користувача. Наприклад, модель, навчена на історичних даних з упередженим представленням соціальних груп, може дискримінувати за статтею чи расою, переносячи минулі стереотипи на прогнози. Це створює конфлікт між метою моделі (об'єктивністю) та її реальною поведінкою.

1.3 Конфлікти цілей та пріоритетів

У статті [5] автор розглядає, як принцип людської саморегуляції та конфлікту цілей може бути використаний при розробці автономних штучних систем. Аналогічно до людей, які можуть мати суперечливі цілі, досягнення однієї з них іноді ускладнює або навіть унеможливлює досягнення іншої, так і в автономних системах впровадження багаторівневої ієрархії цілей із вбудованими конфліктами може слугувати стабілізуючим механізмом, що сприятиме більш безпечній та адаптивній поведінці. Даний конфлікт слід ретельно враховувати у розробці автономних систем озброєння (дрони, тощо).

1.4 Подвійне використання

Із зростанням можливостей ШІ виникає все більше прикладів подвійного використання (dual-use) – ситуації, у яких одну й ту ж модель використовують у протилежних цілях (добросовісних, захисних та зловмисних, наступальних). Наглядним приклади подвійного використання є застосування моделі ChatGPT для генерування шкідливого програмного забезпечення, фішингових листів, і навпаки – аналізу коду на вразливості, генерування звітів про інцидентиі та аналіз індикаторів компрометації.

1.5 Конфлікти знань

Context-Memory conflict – це тип конфлікту, що виникає внаслідок розбіжностей параметричних знань моделі (отриманих під час навчання) та контекстних знань (інформації із запитів користувача, зовнішніх ресурсів тощо). Дослідження [7] демонструє, що LLM можуть

проявляти суперечливі тенденції: з одного боку, вони можуть бути сприйнятливими до зовнішньої інформації та надавати перевагу їй, якщо вона є єдиним зв'язним доказом, але з іншого – у разі наявності одночасно підтверджувальних та суперечливих фактів, модель схильна покладатися на параметричну пам'ять.

Inter-Context conflicts – конфлікти між різними джерелами контекстних знань. Це може призвести до помилкових висновків або нестабільних відповідей, оскільки можуть виникнути труднощі у однозначному визначенні, якій інформації варто надати перевагу.

Intra-Memory conflicts – конфлікти у параметричній пам'яті, внаслідок яких на перефразовані, проте семантично однакові запити, модель буде видавати розбіжні відповіді. Причинами даного конфлікту можуть можуть бути як неточності та суперечливі факти у навчальному наборі даних під час тренування, так і подальше донавчання, яке може вводити нові знання, не узгоджені з тими, що вже містяться в параметрах моделі. Дослідження [9] показує, що класичні архітектури можуть мати високі показники неузгодженості між перефразованими запитами.

2 Конфлікти в системі ШІ-ШІ

2.1 Дистиляція моделей

Дистиляція знань у машинному навчанні — це процес перенесення "м'яких" знань (розподілу ймовірностей між класами), здобутих великою (вчительською) моделлю, до меншої (студентської) моделі. Великій моделі інколи підвищують температуру в softmax, щоб отримати більш гладкий розподіл ймовірностей. Таким чином, дистиляція дозволяє студентській моделі з меншою кількістю параметрів навчитися відтворювати "інтуїцію" вчителя і краще узагальнювати, одночасно залишаючись компактною і дешевшою у навчанні. У статті [10] описується конфлікт між студентською моделлю та ансамблем вчительських моделей під час дистиляції знань.

Також можливе посилення гендерної упередженості мовних моделей під час передачі знань від більшої до меншої та втрата попередніх знань студентської моделі під час дистиляції.

2.2 Змагальний штучний інтелект

Типовим прикладом конфліктів між різними системами штучного інтелекту є змагальні взаємодії, у яких ШІ діють у протилежних або конфліктних інтересах. Із розвитком GAN, який зараз є основою для генерування фейків, почали з'являтися відповідні ШІ-детектори, основним завданням, яких є виявлення згенерованих даних. Також прикладом такої взаємодії в рамках гри з нульовою сумою є багатоагентні системи з навчанням з підкріпленням (multi-agent reinforcement learning), у яких агенти змагаються за досягнення протилежних цілей у спільному середовищі.

2.3 Конфлікти злиття LLM

Злиттям LLM називають об'єднання моделей, натренованих на виконання різних завдань, у одну велику, більш універсальну модель. Типовими способами злиття є усереднення та зважене усереднення параметрів, проте дані методи стикаються з проблемою конфліктів параметрів, коли різні моделі, навчені на різних даних, мають суперечливі параметри. Просте усереднення таких параметрів може погіршити роботу моделі.

У роботі [9] розглядається спосіб вимірювання даних конфліктів між шарами різних моделей, а також фреймворк, при якому шари з низькими рівнем конфліктів усереднюються, а з високим – не об'єднуються, а під час роботи моделі система буде вибирати, який шар, якої моделі краще впорається з конкретним завданням.

3 Конфлікти в системі Людина-ШІ

3.1 Конфлікти спостереження, інтерпретації та дій

У статті [10] розглядаються конфлікти між людиною та ШІ у індустрії 5.0 – етап розвитку технологій, де люди

тісно співпрацюють із ШІ для створення продуктів та послуг. У даній роботі визначають три конфлікти.

Конфлікт спостереження – розбіжність між даними, що сприймаються системою ШІ та – людиною-оператором. Причинами конфліктів може бути як людські помилки під час моніторингу даних, так і некоректні або відсутні дані з датчиків, що відстежуються алгоритмами ШІ.

Конфлікт інтерпретації – розбіжність у розумінні та інтерпретації даних, що може призвести до різниці у прийнятті рішень. Фактори, що сприяють конфліктам інтерпретації, включають відмінності у спостереженнях, розбіжності в можливостях навчання між ШІ та людьми, а також збій системи ШІ або людські помилки.

Конфлікт дій - різниця між операціям контролю ШІ та людини. Фактори, що сприяють конфлікту дій, включають різницю у прийнятті рішень ШІ та людиною, введення в оману під час виконання операцій.

3.2 Проблематика розробки людино-орієнтованого ШІ

На відміну від традиційних систем, ШІ може проявляти несподівану поведінку під час навчання та роботи, це може призводити до конфліктів, оскільки ШІ може видавати непередбачувані результати. Системи ШІ можуть бути розроблені для володіння певним рівнем інтелекту, але існують обмеження в тому, наскільки добре машини можуть емулювати людські когнітивні здібності. Також може існувати складність у пояснюваності машинного виводу, через так звану “проблему чорної коробки”, тобто ситуації, коли неможливо пояснити чому саме і як ШІ дійшов певного висновку.

4 Аналіз наукових та футорологічних прогнозів для виявлення конфліктів у ШІ

Футурологічні прогнози, так само як і наукові аналізи, часто слугують інструментами для моделювання можливих сценаріїв розвитку технологій, зокрема штучного інтелекту. На основі вивчення наукових трендів,

соціальних трансформацій та технологічних проривів, футурологи формують припущення про майбутні ризики, виклики та конфлікти, пов'язані з автономними або надінтелектуальними системами. Наприклад, Рей Курцвейл передбачив появу смартфонів і розвиток машинного перекладу задовго до їх масового впровадження, а Елвін Тоффлер у книзі «Третя хвиля» (1980) описав перехід до інформаційного суспільства, що точно відображає реалії цифрової епохи.

Паралельно з цим, у науковому середовищі активно досліджуються потенційні конфлікти, що можуть виникати внаслідок впровадження ШІ – як технічні, так і соціально-етичні.

Аналіз таких джерел дозволяє глибше осмислити потенційну траєкторію розвитку ШІ, виявити можливі конфліктні й заздалегідь підготуватися до соціальних, етичних та екзистенційних викликів.

В табл. 1 наведено конфлікти, що описуються у футорологічних прогнозах та наукових аналізах та їх суть. Список конфліктів не є вичерпним та може бути доповнений.

Таблиця 1. Приклади конфліктів у ШІ, описані у футурологічних прогнозах та наукових джерелах

Конфлікт	Опис
Неконтрольований розвиток автономної зброї	Розвиток автономних систем озброєння на основні ШІ знижує поріг початку війни. Без прямого людського контролю та аспекту страху машини можуть діяти агресивніше та невибірково
Перегони за лідерство у сфері ШІ	Боротьба за першість у галузі ШІ між великими країнами може викликати новий поділ світу. Країни із нижчим розвитком технологій у сфері ШІ будуть змушені вибирати чию інфраструктуру та стандарти

	використовувати
Проблема “вирівнювання” та розробка загального та сильного ШІ	Створення високрозовниненого ШІ, який потенційно перевершує людський інтелект та має свідомість може призвести до того, що цілі ШІ не збігаються з людськими цінностями, що може призвести до непердбачуваних наслідків
Масове безробіття	Активна автоматизація, спричинена ШІ, у сферах, які потребують рутинних завдань, може залишити без роботи велику частину населення, що ймовірно може призвести до соціальних вибухів, масових протестів
Упередженість та дискримінація алгоритмами ШІ	Системи ШІ, які навчені на даних, що відображають існуючі соціальні упередження, будуть неминуче засвоєні і потенційно посилені, що може стати критичним у таких сферах, як судочинство, кредитування тощо.
Ядерна ескалація	Імплементація алгоритмів ШІ у системи управління зброєю масового ураження для отримання стратегічної переваги може призвести до потенційних ненавмисних ядерних обмінів
Інформаційні війни та пропаганда	Веапонізація GenAI дає змогу створювати реалістичну дезінформацію, що може розпалювати як внутрішні, так і зовнішні конфлікти. Використання діпфейків також може стати великою проблемою у судовій системі
Масове спостереження над	Можливість ШІ обробляти велику різної інформації може бути

населенням	використана авторитарними режимами для моніторингу певних осіб, груп осіб для тотального контролю над населенням
Використання ШІ в терористичних цілях	Аналіз великого об'єму даних за допомогою ШІ може використовуватися у плануванні майбутніх атак або моделюванні реакцій сил безпеки. Автономні машини та дрони можуть бути використані для цілеспрямованих терористичних актів
Кібервійни з використанням ШІ	Інструменти на основі ШІ можуть швидше виявляти вразливості в існуючих системах. ШІ може використовуватися для автоматизації атак з мінімальними на це витратами

4.1 Результати аналізу конфліктів описаних у науково-фантастичних джерелах, футурологічних прогнозах та наукових аналізах

Хоча велика кількість конфліктів, описаних у sci-fi все-таки залишається фантастикою, проте є сценарії, які є спільними для цих трьох джерел:

1. Втрата контролю над автономними системами. ШІ-системи. Інтегровані у військові або критично важливі об'єкти (дрони, оборонні комплекси, енергетичні вузли, тощо), такі системи можуть діяти автономно, без постійного людського контролю. У разі збою, помилки або агресивної ескалації це може призвести до катастрофічних наслідків.

Типові ознаки:

- автоматичні атаки без перевірки людиною;
- неконтрольована ескалація конфлікту;
- помилкові цілі або втрати серед цивільного населення;

- відсутність механізмів зупинки або зміни рішень ШІ;
- надто швидкі прийняття рішень, які випереджають людську реакцію.

2. Масове спостереження з використанням ШІ. Алгоритми ШІ для розпізнавання облич, поведінкової аналітики та моніторингу дій громадян використовуються державами для масового нагляду, що обмежує свободи та права людини.

Типові ознаки:

- впровадження масових систем розпізнавання облич;
- оцінка громадян на основі «соціального рейтингу»;
- неконтрольований збір біометрії для навчання моделей.

3. Втеча ШІ з-під контролю. Надскладні системи ШІ із власною «свідомістю» здатні приймати власні рішення, що суперечать людським цілям там цінностям.

Типові ознаки:

- несподівана та непрогнозована поведінка ШІ;
- інтерпретація завдань у власний спосіб, що суперечить людському баченню;
- відмова ШІ припинити роботу або змінити цілі;
- неможливість пояснити рішення ШІ, наявність «чорної скриньки».

4. Геополітичне суперництво домінування у сфері ШІ. Країни борються за контроль над інфраструктурою для створення провідних ШІ. Виникнення “гонок”, внаслідок яких країни нехтують принципами безпечної розробки ШІ.

Типові ознаки:

- санкції на екпспорт ШІ-технологій (чипів, тощо);
- розподіл світу за технологічними блоками;
- шпигунство, крадіжка ШІ-технологій;

5. Інформаційні війни з використаннями ШІ. Генеративні ШІ використовуються для створення масової дезінформації (діпфейк фото/відео, фейкові новини, маніпулятивні наративи)

Типові ознаки:

- поширення реалістичних шейків;
- дискредитація офіційних джерел;
- поділ інформаційного простору.

6. Технологічне безробіття. ШІ замінює людську працю у сферах з повторюваними функціями, викликаючи зростання безробіття та невдоволення серед населення.

Типова ознака – скорочення робочих місць у різних сферах людської діяльності.

7. Упередженість та дискримінація алгоритмами ШІ. Алгоритми машинного навчання можуть відтворювати, а іноді навіть посилювати системні упередження, закладені в історичних даних. Це створює несправедливі умови для певних соціальних груп – зокрема за ознакою раси, статі, віку або соціального статусу.

Типові ознаки:

- дискримінаційні рішення щодо кредитів, працевлаштування, судових рішень тощо;
- завищена або занижена ймовірність «ризику» для представників окремих груп.

Висновки

У межах дослідження було здійснено аналіз різних типів конфліктів, що можуть виникати в системах штучного інтелекту. Було розглянуто конфлікти в різних системах, а також надано огляд футурологічних сценаріїв та наукових джерел, які можуть дозволити прогнозувати можливі варіанти розвитку конфліктів. Також запропоновано класифікацію конфліктів та їх ознак для раннього виявлення.

Результати можуть бути корисними для розробників, дослідників у сфері безпеки та етики ШІ, а також

використані для виявлення конфліктів у системах ШІ на ранніх стадіях.

Список використаних джерел

1. H Trusilo, D. (2023). Autonomous AI Systems in Conflict: Emergent Behavior and Its Impact on Predictability and Reliability. Journal of Military Ethics, 22(1), 2–17. DOI: 10.1080/15027570.2023.2213985

2. Rivera, J. P., Mukobi, G., Reuel, A., Lamparth, M., Smith, C., & Schneider, J. (2024, June). Escalation risks from language models in military and diplomatic decision-making. In Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (pp.836-898). DOI: 10.48550/arXiv.2401.03408

3. AI-Supported Emotional Conflict Resolution: Technical Approaches and Implementation Strategies - Rajarshi Tarafdar - IJSAT Volume 16, Issue 1, January-March 2025. DOI 10.71097/IJSAT.v16.i1.2792

4. Venkit, P. N., Chakravorti, T., Gupta, V., Biggs, H., Srinath, M., Goswami, K. & Wilson, S. (2024). An Audit on the Perspectives and Challenges of Hallucinations in NLP. arXiv preprint arXiv:2404.07461.

5. Muraven, M. (2017). Goal conflict in designing an autonomous artificial system. arXiv preprint arXiv:1703.06354.

6. Xie, Jian, et al. "Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts." The Twelfth International Conference on Learning Representations. 2023.

7. Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, Yoav Goldberg; Measuring and Improving Consistency in Pretrained Language Models. Transactions of the Association for Computational Linguistics 2021; 9 1012–1031. DOI: 10.1162/tacl_a_00410

8. Binici, K., Pham, N. T., Mitra, T., & Leman, K. (2022). Preventing catastrophic forgetting and distribution mismatch in knowledge distillation via synthetic data. In Proceedings of the IEEE/CVF winter conference on applications of computer vision (pp. 663-671).

9. Lai, K., Tang, Z., Pan, X., Dong, P., Liu, X., Chen, H. & Chu, X. (2025). Mediator: Memory-efficient LLM Merging with Less Parameter Conflicts and Uncertainty Based Routing. arXiv preprint arXiv:2502.04411

10. Rajeevan Arunthavanathan, Zaman Sajid, Faisal Khan, Efstratios Pistikopoulos, Artificial intelligence – Human intelligence conflict and its impact on process system safety, Digital Chemical Engineering,Volume 11, 2024, 100151, ISSN 2772-5081, DOI: 10.1016/j.dche.2024.100151.