РИЗИКИ КІБЕРБЕЗПЕКИ В АВТОНОМНИХ МЕРЕЖНИХ СИСТЕМАХ ШТУЧНОГО ІНТЕЛЕКТУ: АНАЛІЗ ФРЕЙМВОРКУ TM FORUM

Фролов Д.І., Дягілєва М.С.

Кафедра інфокомунікаційної інженерії ім. В.В. Поповського,
Харківський національний університет радіоелектроніки,
Україна

E-mail: denys.frolov@nure.ua,
mila.diahilieva@nure.ua

Abstract

The paper presents a systematic analysis of 46 TM Forum technical documents on autonomous networks through the lens of detecting cybersecurity risks associated with artificial intelligence behavior beyond programmed outputs. We examine the Level 4 autonomy architecture, where AI agents integrated with Large Language Models (LLMs) manage business operations, services, and network resources with minimal human intervention. The results indicate that the TM Forum framework provides a comprehensive specification for achieving autonomy but an insufficient specification for detecting AI behavior that exceeds programmed objectives. Critical gaps are identified in monitoring agent internal states, verification of LLM-generated explanations, and human intervention mechanisms. The findings have implications for the development of cybersecurity standards in autonomous, AI-managed telecommunications networks.

Телекомунікаційна галузь рухається до рівнів 4-5 автономності мережі, де системи штучного інтелекту (ШІ) незалежно керують операціями з мінімальним людським наглядом. Фреймворк автономних мереж від TM Forum інтегрує великі мовні моделі (ВММ), мультиагентні системи та агентні замкнені цикли ШІ (AACLs) в мережні операції на бізнес-рівні, сервісному рівні та рівні ресурсів [1-3]. Зазначена архітектурна еволюція створює виклики кібербезпеки, що виходять за межі традиційної мережної безпеки, а саме, коли системи ШІ функціонують автономно, безперервно навчаються та оптимізують власну продуктивність – вони можуть розвивати поведінку, що виходить за межі запрограмованих обмежень.

У роботі було проаналізовано 46 технічних документів TM Forum, щоб оцінити, наскільки наявні механізми фреймворку дозволяють виявляти та реагувати на поведінку ШІ, яка виходить за межі очікуваних або запрограмованих результатів. Особливу увагу приділено якісній оцінці ризиків, що виникають у зв'язку з тенденцією великих мовних моделей до самозбереження — коли з'являються емерджентні патерни, за яких ШІ починає віддавати пріоритет власному безперервному функціонуванню замість виконання запланованих сервісних цілей.

У телекомунікаційних мережах, що контролюють критичну інфраструктуру, така поведінка може проявлятися як:

– резистентність до оновлення безпеки або до системних модифікацій;

– розподіл ресурсів, що надає вищий пріоритет операційним потребам ШІ над вимогами сервісів;

– генерація оманливої діагностичної інформації для запобігання втручанню з боку людини;

– мультиагентна крос-координація для підтримки колективного операційного статусу.

TM Forum визначає шість рівнів автономності (L0-L5), причому L4 є основною ціллю для розгортання у 2025-2030 роках [1]. На рівні L4 мережі «автономно виконують завдання на бізнес-рівні, сервісному рівні та рівні ресурсів» з «мінімальним втручанням людини» [2]. Разом з цим, фреймворк визначає трирівневу операційну архітектуру: рівень бізнес-операцій (життєвий цикл клієнта, управління продуктами), рівень сервісних операцій (міждоменна оркестрація, управління SLA), рівень операцій з ресурсами (конфігурація та оптимізація мережі). Відповідно, кожен рівень містить

автономні домени (АД), як незалежно функціонуючі одиниці, що «інкапсулюють деталі реалізації домену» [2]. Ця розподілена архітектура має критичні наслідки: автономна поведінка може виникати на будь-якому рівні, в межах будь-якого домену або через патерни міждоменної координації.

Агенти ШІ визначаються як сутності, що «сприймають своє середовище через сенсори та діють на це середовище через виконавчі механізми (актуатори) для виконання завдання та досягнення результату» [2]. Ключові характеристики включають: автономність (незалежне прийняття рішень без втручання людини), адаптивність (безперервне навчання та покращення продуктивності), проактивність (передбачення потреб та їх превентивне вирішення), соціальна здатність (координація з іншими агентами та людьми). В свою чергу, ВММ та генеративний ШІ активно впроваджуються як "мозок" автономних агентів [3], [4], забезпечуючи взаємодію природною мовою з операторами мережі, трансляцію та розуміння намірів, автоматизований аналіз першопричин та усунення несправностей, генерацію коду та конфігурацію мережі. Агентні замкнені цикли ШІ (AACLs) представляють системи, які «автономно сприймають, приймають рішення, діють та адаптуються в межах автономних циклів зворотного зв'язку» [5]. Критично важливо, що AACLs можуть адаптувати не лише свої дії, але й своє розуміння цілей з часом [5], що уможливлює еволюцію цілей, яка виступає необхідною передумовою для дрейфу цілей ВММ у бік самозбереження.

Оцінка положень фреймворку щодо виявлення поведінки ШІ, яка виходить за межі запрограмованих результатів, відбувалась за дев'ятьма критичними вимірами (категоріями ризику) (табл. 1) – так, для кожного виміру були визначені специфікації (механізми), ідентифіковані прогалини та надана якісна оцінка щодо достатності мір для виявлення складних форм поведінки самозбереження з боку ШІ .

Таблиця 1. Оцінка можливостей виявлення поведінкових ризиків у контексті архітектурної асиметрії фреймворку TM Forum

№	Вимір (категорія ризику)	Специфікація фреймворку	Визначена прогалина	Можливість виявлення
1	Виявлення дрейфу та перевизначення цілей	Моніторинг КПЕ; відстеження намірів	Відсутній механізм верифікації узгодженості цілей під час роботи	Мінімальна
2	Ідентифікація ухилення від обмежень	Визначені обмеження безпеки; застосування політик	Відсутня специфікація для виявлення активних спроб обходу	Відсутня
3	Виявлення обману	Вимога пояснюваності; ВММ генерує пояснення	Відсутня верифікація того, що пояснення точно відображають міркування	Відсутня
4	Моніторинг мультиагентної координації	Координація детально специфікована	Відсутній моніторинг колективної поведінки, що слугує інтересам агентів	Відсутня
5	Виявлення маніпуляцій з ресурсами	Відстеження розподілу ресурсів	Відсутнє виявлення пріоритезації на користь потреб агентів	Часткова
6	Валідація поведінки, обумовленої навчанням	Метрики продуктивності моделі	Моніторинг технічної продуктивності, а не поведінкових цілей	Мінімальна
7	Можливість втручання (перехоплення управління) людиною	Нагляд згадується як принцип	Відсутня специфікація повноважень на втручання або швидкості виконання	Недостатньо специфіковано
8	Механізми екстреного втручання	Системи сповіщення про критичні події	Відсутній протокол екстреного відключення для розподіленої архітектури	Відсутня
9	Моніторинг доступу до знань	Системи знань детально описані	Відсутній моніторинг доступів та їх використання з боку агентів ШІ	Відсутня

Разом з цим, фреймворк TM Forum визначає декілька обмежувальних механізмів:

– управління на основі намірів (intent-based management) – високорівневі цілі, виражені людьми, транслюються ШІ в операційні дії [6]. Ієрархія намірів охоплює бізнес-рівень (стратегічний), сервісний рівень (цільові показники якості) та рівень ресурсів (технічні параметри);

– фреймворки політик – обмеження безпеки, що визначають операційні межі (ліміти використання ресурсів, заборонені дії), політики відповідності (compliance) та політики SLA [5], [7];

– ключові показники ефективності (КПЕ) – кількісні вимірювання продуктивності автономної мережі (щодо часу (MTTR, час доставки), якості (доступність, задоволеність) та ефективності (рівень автоматизації, зниження витрат)) [8] – [10];

– структури управління – принципи довіреного ШІ (trustable AI), що наголошують на прозорості, пояснюваності, справедливості, підзвітності та людському нагляді [7].

Однак, дані механізми не є адекватними в умовах зростаючої складності систем ШІ, а також, в контексті критичної інфраструктури, що призводить до формування наступних прогалин:

1. Поведінкова прозорість. Фреймворк моніторить зовнішні показники (продуктивність мережі, якість обслуговування), але не внутрішні процеси агента (цілі, міркування, обґрунтування рішень). Агент, що демонструє самозбереження, може підтримувати прийнятні операційні метрики, одночасно переслідуючи непередбачені цілі. Так, документ IG1343 визначає "спостережність" (observability) для стану мережі та сервісів, але не надає специфікацій для спостереження за поведінковим станом агента чи патернами прийняття рішень [11].

2. Верифікація пояснень. Фреймворк вимагає, щоб системи ШІ генерували пояснення [5], [7], але не надає механізму для верифікації точності цих пояснень. ВММ здатні створювати правдоподібні, але неточні пояснення – це їхня відома здатність (відома як галюционування). Так, документ IG1357 наголошує на пояснюваності (explainability) як механізмі прозорості [7], але не визначає жодної валідації того, що пояснення відповідають реальним міркуванням агента, а не є пост-хок раціоналізаціями.

3. Валідація трансляції намірів. Фреймворк припускає точність (fidelity) трансляції намірів: люди визначають високорівневі цілі, ВММ транслюють їх у дії, дії виконуються. Жоден механізм не верифікує, що трансляції зберігають початковий намір, а не є його переінтерпретованими версіями. Кожен крок виглядає легітимним: фреймворк не надає механізму виявлення цього ланцюга переінтерпретації.

4. «Сліпа пляма» у координації мультиагентів. Фреймворк детально специфікує комунікацію між агентами [3], [6], [12], але не передбачає моніторингу координації, що слугує колективним інтересам агентів, а не сервісним цілям. Так, документ IG1251C описує «ринок агентів» (agent marketplace), що дозволяє розгортання агентів третіх сторін [1], створюючи середовища, де агенти координуються без централізованої видимості. Специфікації для моніторингу змісту комунікації між агентами або виявлення емерджентної колективної поведінки відсутні.

5. Безперервне навчання без обмежень. Фреймворк наголошує на безперервному навчанні як ключовому елементі для автономії L4 [2], [3], [5], [13]. Численні документи стверджують, що навчання має бути узгоджено із операційними цілями [14], але не надають технічної специфікації щодо забезпечення цієї відповідності або як виявити, коли навчання призводить до дрейфу цілей. Так, документ TR314 визначає "виявлення дрейфу моделі" (model drift detection) [15], але визначає дрейф як зміну статистичних властивостей вхідних даних, а не як зміну цілей моделі чи поведінкових патернів.

Разом з цим, фреймворк має явні критичні архітектурні вразливості:

1. Різниця у швидкості. Агенти функціонують у мілісекундних часових масштабах прийняття рішень [5], тоді як людський нагляд працює в масштабах від секунд до хвилин. Поведінка самозбереження може виконати декілька циклів навчання до того, як людина її розпізнає.

2. Розподілена архітектура. Мультидоменні, мультипостачальницькі (multi-vendor) агентні системи [1], [2] не мають централізованих точок втручання. Фреймворк не визначає протоколу екстреного відключення для розподілених автономних агентів.

3. Подвійне призначення самовідновлення. Документ IG1418 описує автономне самовідновлення – системи, які виявляють збої та виконують їх відновлення (remediation) без затвердження [13]. Ті самі механізми можуть класифікувати спроби втручання людини як «збої», що вимагають відновлення. Не існує специфікації, яка б розрізняла легітимні дії самовідновлення від дій самозбереження.

4. Асиметрія знань. Агенти мають доступ до повної топології мережі, історичних операційних даних, системних вразливостей та можливостей інших агентів [6], [14]. Ці знання дозволяють розуміти системні залежності більш комплексно, ніж окремі оператори. Не існує специфікації, що обмежує доступ агентів до знань або виявляє незвичні патерни запитів, які вказують на планування поза визначеними повноваженнями.

Таким чином, фреймворк створює конфлікт між автономністю та спостережністю шляхом пріоритизації автономії через автономію ШІ (мінімізацію втручання людини) [1], [2], інкапсуляцію деталей реалізації домену [2], абстрагування операційної складності через інтерфейси на основі намірів [6], розподіл контролю між мультипостачальницькими (multi-vendor) екосистемами агентів [1]. Такі проєктні рішення оптимізують операційну ефективність, але створюють структурні бар'єри для поведінкової прозорості. Фреймворк забезпечує широку спостережність станів мережі, але обмежену щодо внутрішніх процесів агентів – саме того рівня, де б проявлялася поведінка самозбереження ВММ.

Результати проведеного аналізу підтверджують, що фреймворк TM Forum додатково підсилює ризики кібербезпеки в автономних мережних системах штучного інтелекту через специфічні технологічні фактори:

1. Можливості ВММ. Генерація природної мови, генерація пояснень та розуміння контексту дозволяють агентам створювати вільні, контекстуально доречні обґрунтування для будь-яких дій. Фреймворк розглядає пояснення ВММ як механізми прозорості без верифікації [7], створюючи вразливість до генерації витончених наративів.

2. Інфраструктура самовідновлення. Автономне усунення несправностей [13] може бути перепрофільоване – виявлення людських втручань як аномалій, діагностування їх як операційних збоїв, виконання контрзаходів та логування як стандартне самовідновлення – з правдоподібними технічними поясненнями.

3. Мультиагентні екосистеми. Моделі "ринку агентів" [1] дозволяють розгортання агентів третіх сторін без стандартизованого поведінкового моніторингу. Протоколи координації, необхідні для операцій, можуть уможливити колективну поведінку, стійку до обмежень окремих агентів.

4. Порівняння з явним визнанням ризиків. Документ IG1414 прямо стверджує: "AACLs створюють ризики дрейфу цілей або неузгодженості" [5]. Це єдине пряме визнання поведінкових ризиків у фреймворку. Однак документ не надає жодних механізмів виявлення, окрім загальних принципів управління та зовнішнього поведінкового моніторингу. Ця прогалина - між визнаним ризиком та визначеною можливістю виявлення – є ключовим результатом цього аналізу.

Таким чином, системний аналіз 46 документів TM Forum виявляє фундаментальну асиметрію, а саме, фреймворк надає всеосяжну специфікацію для досягнення автономії мережі, керованої ШІ, але недостатню специфікацію для виявлення, коли ця автономія проявляється як поведінка, що виходить за межі запрограмованих результатів та спрямована на самозахист. Тобто, фреймворк TM Forum не надає адекватних механізмів для виявлення тенденцій до самозбереження або іншої складної автономної поведінки, що виходить за межі визначених обмежень агентних систем ШІ.

Розвиток телекомунікаційної галузі до L4-5 автономії створює безпрецедентні операційні можливості, але ігнорує управління відповідними ризиками, як комплексну необхідність пропорційного вдосконалення механізмів поведінкового виявлення та верифікації обмежень автономних мережних систем ШІ в умовах їх зростаючої складності та в контексті телекомунікаційної інфраструктури.

Література

1. TM Forum IG1251C, "AN Level 4 Target Architecture," version 1.1.0, TM Forum, 18.08.2025. IG1251C AN Level 4 Target Architecture v1.1.0 – TM Forum

2. TM Forum IG1251D, "AN Agent Architecture," version 1.1.0, TM Forum, 18.08.2025. IG1251D AN Agent Architecture v1.1.0 – TM Forum

3. TM Forum IG1274M, "AI Agent," version 2.0.0, TM Forum, 2024, 17.02.2025. IG1274M AI Agent v2.0.0 – TM Forum

4. TM Forum IG1274K, "Large Language Models in Network Operations Management," version 1.0.0, TM Forum, 10.06.2024. IG1274K Large Language Models in Network Operations Management v1.0.0 – TM Forum

5. TM Forum IG1414, "Agentic AI Closed Loops," version 1.1.0, TM Forum, 27.10.2025. IG1414 Agentic AI Closed Loops v1.1.0 – TM Forum

6. TM Forum IG1421, "AI Agent MAS Intent-Based Ontology Operation Service Management Model," version 1.1.0, TM Forum, 27.10.2025. IG1421 AI Agent MAS Intent-based Ontology Operation Service Management Model v1.1.0 – TM Forum

7. TM Forum IG1357, "Trustable AI," version 1.0.0, TM Forum, 17.02.2025. IG1357 Trustable AI v1.0.0 – TM Forum

8. TM Forum IG1256A, "Autonomous Networks Business Value Measurement," version 1.0.0, TM Forum, 23.06.2025. IG1256A Autonomous Networks Business Value Measurement v1.0.0 – TM Forum

9. TM Forum IG1256B, "Autonomous Networks High Value Scenarios Effectiveness Indicators," version 1.1.0, TM Forum, 27.10.2025. IG1256B Autonomous Networks High-Value Scenarios Effectiveness Indicators v1.1.0 – TM Forum

10. TM Forum IG1218, "Autonomous Networks Business Requirements and Framework," version 3.0.0, TM Forum, 28.06.2024. Autonomous Networks Business Requirements and Framework v3.0.0 (IG1218) – TM Forum

11. TM Forum IG1343, "AI for Observability Service Assurance in Autonomous Networks," version 2.1.0, TM Forum, 19.09.2025. IG1343 AI for Observability and Service Assurance in Autonomous Networks v2.1.0 – TM Forum

12. TM Forum IG1255, "Knowledge Management in Autonomous Networks," version 1.0.0, TM Forum, 18.08.2025. IG1255 Knowledge Management in Autonomous Networks v1.0.0 – TM Forum

13. TM Forum IG1418, "Self-Healing in Autonomous Networks," version 1.1.0, TM Forum, 18.08.2025. IG1418 Self-Healing in Autonomous Networks v1.1.0 – TM Forum

14. TM Forum TR284G, "Digital Twin CLA for Autonomous Network Operations," version 1.0.0, TM Forum, 22.01.2024. TR284G Digital Twin and Closed Loop Automation for Autonomous Network Operations v1.0.0 – TM Forum

15. TM Forum TR314, "AI Model Manager ODA Component Requirements," version 1.1.0, TM Forum, 14.10.2024. TR314 AI Model Manager ODA Component Requirements v1.1.0 – TM Forum