

РОЗДІЛ 6

Ризик ескалації AI-системи — чотири задокументовані випадки

Чотири публічні інциденти, у яких неясні ролі завдали більше шкоди, ніж технічні помилки.

Кожен з чотирьох кейсів — конкретна реалізація одного й того ж механізму: помилка чи нестандартний результат на шарі моделі доходить до шару юридичних і фінансових наслідків, бо жоден з проміжних шарів — сигнал, оператор, мандат, документація, контракт — поширення не зупинив. Шари не випадкові; нижче кожен з кейсів показує, який саме шар лишився порожнім.

Коли AI-інцидент уже стався, дуже зручно сформулювати його так: модель «галюцинувала», модель «зламалась», модель «не була готова». Зручно — бо такі формулювання переносять відповідальність на технологію і знімають її з тих організаційних рішень, що передували інциденту. Цей розділ пропонує зворотню оптику: дивитися на ті рішення, які лишились непринятими, і на ті ролі, які лишились не зайнятими.

Чотири кейси цього розділу — публічно задокументовані інциденти. Кожен має судове рішення, регуляторну постанову або офіційне розслідування з відкритими матеріалами. Об'єднує їх одне: технічна частина у кожному кейсі виконувала закладену в неї логіку — а проблема лежала в антропологічному шарі. У питанні, хто міг і мав діяти, коли система видала результат, що вимагав людської реакції; в архітектурі ролей, у документації, у мандатах, у контрактах. Технологія не виходила з-під контролю — її контекст ніколи не був побудований повністю.

Інша галузь, інший масштаб, інший тип помилки в кожному з кейсів. Але одна й та сама схема: роль не зайнята, мандат неясний, документація відсутня або застаріла.

Норберт Вінер у праці *Cybernetics* формулює, чому це повторюється системно: автоматична система без замкненої петлі зворотного зв'язку з людиною генерує не *помилки* у вузькому сенсі, а *дрейф* — повільне відхилення від наміру, яке ніким не реєструється до моменту критичної відмови. Стюарт Расселл у праці *Human Compatible* перекладає це у сучасну рамку: ескалація — це штатна фаза петлі керування, а не кризова подія, і її треба проєктувати, не очікувати, що вона *складеться сама*.

У кожному з чотирьох кейсів нижче можна назвати конкретну роль, яка мала би існувати, але не існувала — і саме її відсутність перетворила технічно коректну роботу системи на інцидент. Кейс А демонструє відсутність *ескалаційної ролі*: людини з оперативним мандатом перехопити автоматику у конкретному режимі відмови. Кейс В — відсутність *редакторської ролі*: куратора моделі, який блокує перенесення статистично прогностичних ознак з комерційного контексту у державне рішення без окремого обґрунтування. Кейс С — відсутність *комплєнс-перекладача*: людини, що перетворює внутрішні рішення моделі у форму, придатну для зовнішнього аудиту без розкриття комерційної таємниці. Кейс D — відсутність *аудитора суверенності*: ролі, що перевіряє реальну незалежність контрактних залежностей від геополітичних сценаріїв. Чотири різні галузі, чотири різні типи помилок — одна спільна структура вакантних ролей. Саме ця структура, а не окремі історії, і є предметом розділу.

Кейс А — авіація: Boeing 737 MAX MCAS (Lion Air JT610, 2018; Ethiopian ET302, 2019)

Контекст. Boeing 737 MAX — оновлення базової моделі, у якому для компенсації аеродинамічних змін від нових двигунів додано систему MCAS (Maneuvering Characteristics Augmentation System). У вихідній конфігурації MCAS брала вхід від одного сенсора кута атаки (AoA) і автоматично опускала ніс літака для уникнення стійлового режиму.

Що сталося. 29 жовтня 2018 року рейс Lion Air JT610 розбився у Яванському морі через 13 хвилин після зльоту: 189 загиблих. Хибний AoA-сенсор подавав некоректний кут, MCAS циклічно опускав ніс. 10 березня 2019 року рейс Ethiopian Airlines ET302 розбився у тому самому сценарії через шість хвилин після зльоту: 157 загиблих. Загалом 346 жертв за п'ять місяців. Світове припинення польотів усього флоту 737 MAX тривало з березня 2019 по листопад 2020.

Що було не на місці. Експлуатаційний мандат для ручного перехоплення MCAS — коли і як пілот зобов'язаний відключити автоматику, який сигнал є тригером, як підтверджується інформованість — не був сформульований у льотних посібниках. MCAS згадувалась у документації мінімально; навчання включало ознайомлення з планшета тривалістю кілька десятків хвилин, без сценарію «активна MCAS із несправним сенсором» у тренажері. Перемикачі відключення фізично існували, але без попереднього знайомства з цим режимом відмови пілоти не мали тренованої реакції на нього.

Антропологічний аналіз. MCAS виконувала закладену в неї логіку — однак ця логіка, процес сертифікації, проєктна документація та підготовка пілотів виявилися недостатніми для безпечної експлуатації. Після припинення польотів FAA затвердила пакет виправлень: двосенсорна валідація, обмеження на повторну активацію, обов'язкове тренування на симуляторі для всіх пілотів 737 MAX і явний протокол ручного перехоплення. House Committee у фінальному звіті 2020 року поклав відповідальність на поєднання конструктивних рішень Boeing, недоліків сертифікації та регуляторного нагляду FAA. Ескалаційна роль (пілот як власник мандату на ручне перехоплення у конкретному режимі відмови) не була забезпечена ні навчанням, ні явним протоколом — і це лише один з шарів вакантності, що накопичились до катастрофи.^[1]

Кейс В — публічний сектор: AMS-Algorithmus (Arbeitsmarktservice Austria, 2018–2020)

Контекст. Arbeitsmarktservice (AMS) — державна служба зайнятості Австрії. У 2018 році AMS запустила алгоритм, що мав статистично оцінювати шанси безробітних на повернення на ринок праці і розподіляти їх у три групи: А (високі шанси), В (середні), С (низькі). Розробкою займалася консультативна фірма Synthesis Forschung, вартість проєкту — близько 240 000 євро. Розподіл по групах планувалося використовувати при ухваленні рішень про доступ до програм перекваліфікації.

Що сталося. Перший публічний аудит моделі (AlgorithmWatch, 2019) показав, що формула надає негативні ваги ознакам «жінка», «має дітей» (тільки для жінок, не для чоловіків), «вік 30+», «вік 50+», «інвалідність», «не громадянин ЄС». Жінка з

дитиною автоматично отримувала нижчий бал, ніж чоловік з ідентичним резюме. AMS обґрунтовувала це як «віддзеркалення реальних шансів на ринку». 20 серпня 2020 року Datenschutzbehörde Österreich (DSB) рішенням D213.1020/2020-0513.605 визнала систему несумісною з вимогами захисту даних: відсутня правова основа, відсутні гарантії від дискримінації, рішення de facto автоматизоване попри декларацію про «лише підтримку». Впровадження було призупинене на час оскарження. У 2025 році Bundesverwaltungsgericht (BVwG) Австрії скасував рішення DSB — суд дійшов висновку, що обробка даних AMS у цьому форматі не порушує GDPR і що оцінка дискримінаційного ефекту не належить до компетенції DSB. Питання залишається відкритим у суспільно-політичній і науковій дискусії, але юридично система не була остаточно визнана незаконною.

Що було не на місці. Перед запуском моделі ніхто не поставив питання: *які кореляції з історичними даними ринку праці ми не хочемо переносити у державне рішення про доступ до програм?* Команда розглядала задачу як технічну (передбачити майбутні шанси на базі минулих), а не як адміністративну (розподілити обмежений ресурс справедливо). Роль куратора моделі у державному контексті — з мандатом блокувати ознаки, що корелюють з захищеними категоріями, навіть якщо вони статистично прогностичні, — не була інституційно закріплена.

Антропологічний аналіз. Модель була побудована для прогнозування ймовірності повернення на ринок праці на основі історичних даних і демонструвала прогностичну здатність щодо показника, для якого її було розроблено. Редакторська роль над ознаками — з мандатом сказати «ця ознака допустима як прогностична змінна у комерційному контексті, але потребує окремого обґрунтування як підстава для державного рішення», — була вакантною. Проект коштував платникам податків ~240 тис. євро, дав публічний скандал, рішення DSB 2020 року про несумі-

місність з GDPR і його подальше скасування BVwG у 2025 році. Незалежно від кінцевого юридичного результату, кейс залишається підручковим прикладом того, як адміністративну задачу (розподіл обмеженого ресурсу) можна оформити як технічну (передбачити шанси) — і пропустити антропологічний шар.[2]

Кейс С — кредитний скоринг: Schufa CJEU C-634/21 (рішення 7 грудня 2023)

Контекст. Schufa Holding AG (Вісбаден) — найбільша німецька кредитна довідкова агенція. Її автоматизований скоринг (Schufa-Score) широко використовується банками, лізинговими і телеком-компаніями для рішень про кредит, оренду, мобільні контракти. Алгоритм формування скору ніколи не публікувався: Schufa посилається на комерційну таємницю.

Що сталося. Громадяниці в Гессені відмовили в кредиті на підставі низького Schufa-скору. Вона запросила пояснення ваг і факторів. Schufa повідомила лише підсумкове число і загальний опис методики, відмовившись розкрити внутрішні параметри. Адміністративний суд Вісбадена передав питання до Суду Європейського Союзу (CJEU — найвищий судовий орган ЄС з тлумачення європейського права). Рішенням у справі C-634/21 (7 грудня 2023) Суд встановив: коли третя сторона значною мірою покладається на автоматичний скоринг при ухваленні рішення щодо людини, такий скоринг підпадає під статтю 22 GDPR — і вимагає обов'язкових гарантій (людський перегляд, змістовне пояснення, право оскарження). Формальне «лише враховуємо» з

боку банку не виводить рішення з-під статті 22, якщо фактично відмова майже завжди слідує за низьким скором. Після рішення кредитні бюро у ЄС переглядали практику прозорості і права на оскарження.

Що було не на місці. Система робила класифікацію за внутрішніми специфікаціями, на яких побудована її бізнес-модель. Точка взаємодії між нею і зовнішнім ревізором — спочатку клієнтом, потім національним судом, потім CJEU — не була спроектована. Жодного інструменту пояснення рішення моделі для нетехнічного аудитора. Жодного механізму, який дозволяв би змістовно пояснити результат скорингу без розкриття комерційної таємниці.

Антропологічний аналіз. Скоринг працював у межах закладеної бізнес-логіки. Роль комплаєнс-перекладача — людини, яка перетворює внутрішнє рішення моделі у форму, яку зовнішній аудитор може перевірити без доступу до моделі, — не існувала ні в Schufa, ні в більшості порівнянних агенцій. Без такого перекладу автоматизований скоринг, коли треті сторони значною мірою на нього покладаються, опиняється у пастці статті 22 — не через сам факт автоматизації, а через відсутність моста між моделлю і людиною з контрольним мандатом. Класична відсутня роль AI-оператора у вимірі нормативної відповідності.^[3]

Кейс D — хмарні санкції 2022: припинення обслуговування AWS, Microsoft, Google

Контекст. Європейські промислові і фінансові компанії розгортають частину конвеєра машинного навчання у публічних хмарних провайдерів (AWS, Microsoft Azure, Google Cloud) через економію масштабу, доступ до керованих AI та ML сервісів. Багато з них мають дочірні підрозділи в Росії та Білорусі або партнерів, які працюють у санкційних юрисдикціях.

Що сталося. 4 березня 2022 року Microsoft оголосила про припинення нових продажів продуктів і сервісів у Росії. 8 березня 2022 року AWS повідомила, що не реєструє нових клієнтів у Росії і Білорусі. Google зупинив нові підписки на платні сервіси Google Cloud. У грудні 2023 року 12-й санкційний пакет ЄС (Council Decision (CFSP) 2023/2874 і Council Regulation (EU) 2023/2878 від 18 грудня 2023) розширив обмеження на постачання корпоративного програмного забезпечення для юридичних осіб, зареєстрованих у Росії. У результаті низка європейських компаній зі спільною інфраструктурою з російськими підрозділами була змушена переглядати архітектуру своїх міжнародних інформаційних систем та процедури доступу до хмарної інфраструктури. Публічних технічних розборів (post-mortem) таких перебудов майже не існує — тому цей кейс важливий не як історія конкретної компанії, а як приклад нового класу ризику, що проявився після санкційних рішень 2022 року.

Що було не на місці. До 2022 року стандартна перевірка контрактів для хмарних сервісів охоплювала угоди про рівень обслуговування (SLA — service level agreement: гарантований час доступності і час відгуку сервісу), ціну, технічну сумісність і відповідність вимогам GDPR. Сценарій геополітично вмотивованої за-

борони доступу до сервісу — не від технічного збою, а від юридичного зобов'язання провайдера перед своєю державою — у більшості європейських компаній не покривався ні сценарієм дій, ні планом дій на непередбачені випадки, ні стратегією виходу.

Антропологічний аналіз. Технічно системи працювали коректно до моменту контрактного або юридичного обмеження. Роль аудитора суверенності — людини, яка перевіряє реальну незалежність контрактних залежностей від геополітичних сценаріїв і будує план дій на непередбачені випадки під сценарій раптового відключення, — не була призначена. Відділ закупівель обирав за ціною і SLA; IT-архітектура — за технічною сумісністю; юристи — за стандартними положеннями. Жоден з цих відділів не мав мандату ставити питання: *За яких сценаріїв цей провайдер може юридично втратити право нас обслуговувати — і що тоді?*[4]

Спільна схема: у всіх чотирьох кейсах проблемою була антропологія, не модель

Систематизуємо чотири кейси:

Кейс А (Boeing MCAS). Мандат на ручне перехоплення автоматики у режимі відмови не оперативний. Технічно MCAS виконувала закладену в неї логіку — однак ця логіка, сертифікація і документація проєктувалися без явного протоколу пілотної дії у конкретному режимі відмови. Ескалаційна роль (пілот як власник мандату на ручне перехоплення) не забезпечена ні навчанням, ні протоколом.

Кейс В (AMS-Algorithmus). Куратор ознак з мандатом сказати «ця змінна допустима як прогностичний фактор у комерції, але потребує окремого обґрунтування у державному рішенні» — відсутній. Технічно: модель виконувала прогноз, для якого її було побудовано. Редактор-

ська роль над ознаками у адміністративному контексті — вакантна. Юридична доля рішень оскаржувалась і змінювалась, антропологічний урок — ні.

Кейс C (Schufa CJEU). Механізм комплаєнс-перекладу для зовнішніх аудиторів — відсутній. Технічно: скоринг працював за внутрішньою бізнес-логікою. Немає мосту між моделлю і людиною з контрольним мандатом (клієнт, регулятор, суд) — і саме це робить автоматизований скоринг у певних умовах підставою для статті 22 GDPR.

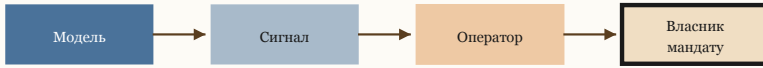
Кейс D (хмарні санкції 2022). Аудит суверенності контрактних залежностей — відсутній. Технічно: системи виконували закладену функцію до моменту юридичного обмеження з боку провайдера. Роль відповідальності не ідентифікувала геополітичний ризик у залежностях від постачальника.

У всіх чотирьох випадках модель робила або намагалася робити те, для чого була призначена. Проблема виникла там, де людська структура — ролі, мандати, документація, контракти — не надавала контексту для реакції на результат системи.

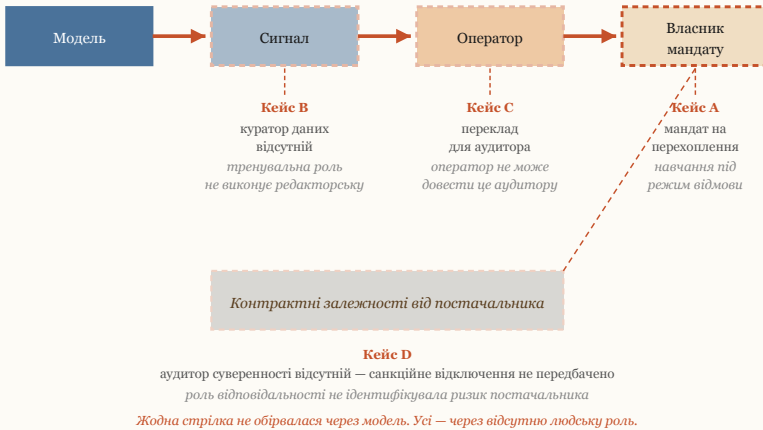
Ескалаційний потік і де він обривається

Верхній шар — як організація уявляє ескалацію. Нижній — як вона працює насправді.

Шар очікування — «у нас є ескалація»



Шар реальності — де ескалаційний потік обривається



Жодна стрілка не обірвалася через модель. Усі — через відсутню людську роль.

Якір Р6/1. Чотири ескалаційні випадки А–D: жодне обірвання не сталося в моделі — усі через відсутню людську роль на конкретному шарі.

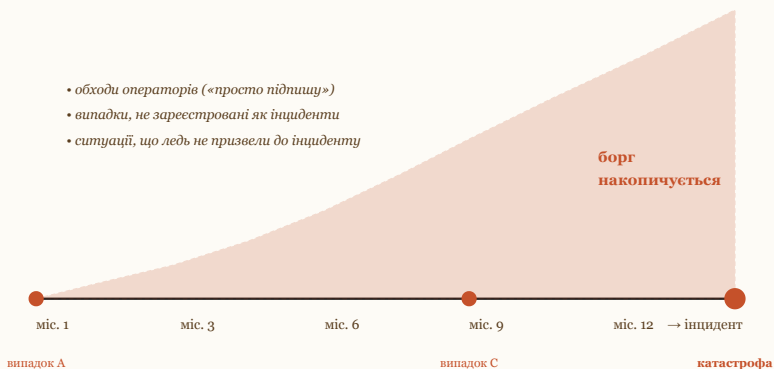
Це і є сенс антропологічного підходу: не замінити технічний погляд, а доповнити його питаннями, яких технічний аналіз не ставить.

Невидимий борг ескалації

невидимий борг ескалації, що накопичується між інцидентами

«ескалаційний борг» — обходи, які ніхто не записав

- обходи операторів («просто підпиши»)
- випадки, не зареєстровані як інциденти
- ситуації, що ледь не призвели до інциденту



Видимі інциденти — точки. Невидимий борг — площа під лінією.

ОЧІКУВАНИЙ ПРОТИ РЕАЛЬНОГО КОНТРОЛЮ

мотив-повторювач ($P6 \cdot P8 \cdot P16 \cdot P21$)

Шар очікування

людина в контурі · експерт перевіряє · керівник підписує · сильно



розрив

Шар реальності

довідка замість аудиторного сліду · підпис без перевірки · мандат без важеля

Контроль існує там, де у людини є мандат, канал і час. Не там, де вписано ім'я.

Три питання для власного контексту

1. **Мандати.** Чи існують для кожного рівня ескалації вашої системи письмово названі особи з чітким мандатом на прийняття рішення — також поза робочими годинами?
2. **Кураторство даних.** Хто перевіряє ваші тренувальні і повторно-тренувальні набори даних на аномалії, які модель *не повинна* вивчити?
3. **Суверенність постачальників.** Які з ваших AI-компонентів залежать від провайдера, який у геополітичному сценарії може бути юридично змушений обмежити доступ — і чи покритий цей сценарій планом аварійного реагування?

Посилання

1. FAA, «Summary of the FAA's Review of the Boeing 737 MAX», 18 листопада 2020; US House Committee on Transportation and Infrastructure, «Final Committee Report on the Design, Development, and Certification of the Boeing 737 MAX», 15 вересня 2020.
2. Nicolas Kayser-Bril, «Austria's employment agency rolls out discriminatory algorithm, sees no problem», AlgorithmWatch, 6 жовтня 2019; Datenschutzbehörde Österreich, рішення D213.1020/2020-0513.605 від 20 серпня 2020; Bundesverwaltungsgericht Österreich, рішення про скасування DSB-Bescheid, 2025.
3. Judgment of the Court (First Chamber) of 7 December 2023, Case C-634/21, SCHUFA Holding (Scoring), ECLI:EU:C:2023:957.
4. Brad Smith, «Microsoft suspends new sales in Russia», Microsoft On the Issues Blog, 4 березня 2022; Amazon Staff, «Amazon's assistance in Ukraine», aboutamazon.com, 8 березня 2022; Council Decision (CFSP) 2023/2874 of 18 December 2023 amending Decision

2014/512/CFSP; Council Regulation (EU) 2023/2878 of 18 December 2023 amending Regulation (EU) No 833/2014 concerning restrictive measures in view of Russia's actions destabilising the situation in Ukraine, EUR-Lex.