МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
«КИЄВО-МОГИЛЯНСЬКА АКАДЕМІЯ»

ФАКУЛЬТЕТ СОЦІАЛЬНИХ НАУК І СОЦІАЛЬНИХ ТЕХНОЛОГІЙ

КАФЕДРА МІЖНАРОДНИХ ВІДНОСИН

КВАЛІФІКАЦІЙНА РОБОТА

освітній рівень – бакалавр

на тему: **«DEAD HAND 2.0 – АВТОМАТИЗОВАНЕ ЯДЕРНЕ СТРИМУВАННЯ В ЕПОХУ ШТУЧНОГО ІНТЕЛЕКТУ»**

Виконала:
Здобувачка 4 року навчання
Спеціальності 291 «Міжнародні відносини,
суспільні комунікації та регіональні студії»

Боженко Анна Євгеніївна

Керівник
Тараненко Ганна Геннадіївна
Доцент, кандидат політичних наук
доцент кафедри міжнародних відносин

КИЇВ 2026

ЗМІСТ

ВСТУП

Проблема дослідження. Класичні теорії ядерного стримування Шеллінга, Кана, Уолца, сформовані в період Холодної війни, ґрунтуються на припущенні про раціональність акторів, наявність людського контролю над рішеннями та відносну передбачуваність ескалаційних процесів, а подальший розвиток теорії ігор у працях Ф. Загаре, Д. Кілгура та М. Крейга дозволив формалізувати механізми стримування та стратегічної стабільності, однак ці моделі переважно виходять із умов, у яких ключові рішення ухвалюються людьми. Водночас сучасний розвиток автоматизованих систем управління, зростання ролі алгоритмічного аналізу даних і впровадження елементів штучного інтелекту у військові системи створюють нову реальність ядерного стримування. Історичний досвід функціонування радянської системи «Периметр» («Dead Hand») демонструє, що автоматизація може зменшувати час на ухвалення рішень, але водночас підвищує ризики помилкової ескалації та втрати людського контролю. У сучасних умовах ці ризики посилюються через інтеграцію елементів штучного інтелекту, що потенційно змінює логіку стратегічної стабільності. Як наслідок, виникає прогалина між класичними теоретичними моделями ядерного стримування та сучасними технологічними реаліями їх практичної реалізації. Це дослідження спрямоване на заповнення цієї прогалини шляхом комплексного аналізу автоматизованого ядерного стримування та ролі штучного інтелекту в процесах ухвалення рішень із використанням інструментів теорії ігор, елементів кейс-стаді та аналізу стратегічних моделей, що дозволить знайти оптимальні підходи до реалізації ядерного стримування в сучасних умовах міжнародної безпеки.

Актуальність. Сьогодні ядерне стримування залишається одним із ключових елементів міжнародної безпеки, особливо в умовах зростаючих геополітичних напружень та нестабільності та суттєвими змінами, які відбуваються на тлі технологічного прогресу та трансформації воєнних доктрин ядерних держав. Для України ця проблематика має особливе значення в умовах агресії Росії проти України, яка триває з 2014 року і повномасштабного вторгнення в 2022 році війни

з Російською Федерацією, яка регулярно використовує ядерну риторику як інструмент політичного тиску та залякування. Розуміння механізмів сучасного ядерного стримування, зокрема потенційних сценаріїв автоматизованої ескалації та ризиків, пов'язаних із використанням штучного інтелекту, є важливим для формування обґрунтованих позицій України у сфері міжнародної безпеки та стратегічних комунікацій із партнерами.

Інтеграція інструментів теорії ігор та кейс-стаді в аналіз стратегій ядерного стримування дозволяє не лише поєднати класичні теоретичні моделі з сучасними технологічними реаліями, а й здійснити більш глибоку оцінку стратегічної стабільності, виявити потенційні точки вразливості систем стримування та сформулювати практично орієнтовані рекомендації щодо збереження контролю над ядерними озброєннями в умовах розвитку штучного інтелекту.

Метою цього дослідження є з'ясувати ключові стратегії та оптимальні підходи до реалізації автоматизованого ядерного стримування в умовах розвитку штучного інтелекту на прикладі кейсу «Dead Hand 2.0», інтегруючи принципи теорії ігор для оцінки ефективності моделей ухвалення рішень, стратегічної стабільності та ризиків ескалації.

Завданням цього дослідження є:

- Виокремити ключові концепти теорії ігор, які мають значення для ядерного стримування, зокрема раціональність, рівновагу Неша, з метою характеристики їх значення для аналізу стратегій автоматизованого ядерного стримування;
- Обґрунтувати базову модель ядерного стримування та її доречність для сучасних стратегій ядерного стримування та потенційних систем «Dead Hand 2.0»;
- З'ясувати історичні передумови та досвід функціонування системи «Dead Hand» («Периметр») як прикладу ранньої автоматизації ядерного стримування;

- Виявити потенційні переваги та ризики автоматизованого ядерного стримування, зокрема з погляду помилкових рішень, ненавмисної ескалації та втрати людського контролю;
- З'ясувати застосовність і значення моделей автоматизованого ядерного стримування та їх модернізації для глобальної стратегічної стабільності;
- Дати нову інтерпретацію потенційного розвитку «Dead Hand 2.0» у контексті міжнародної безпеки та зниження ризиків стратегічної нестабільності в епоху штучного інтелекту.

Опис використаної літератури. Дослідження стратегій ядерного стримування, зокрема, через призму теорії ігор ґрунтується на літературі, що містить класичні та сучасні праці з теорії ігор, міжнародних відносин, ядерного стримування та безпеки, а також досліджень у сфері військового застосування штучного інтелекту. Одним з джерел є праця Джона Неша (1950), яка встановлює основи рівноваги Неша - центрального поняття в теорії ігор, що широко використовується в аналізі стратегій ядерного стримування. Його робота слугує фундаментом для подальших досліджень, таких як Гіббонса (1997), який вводить базові поняття та моделі теорії ігор, та Люс і Райффа (1957), які запропонували критичний огляд рішень в рамках теорії ігор.

Глибокий аналіз стратегій стримування міститься в працях Томаса Шеллінга (1960, 1966). Його роботи «The Strategy of Conflict» та «Arms and Influence» розглядають стратегії загроз і переговорів у контексті теорії ігор, що є надзвичайно важливим для розуміння ядерного стримування. Подібні теми досліджуються у роботах Германна Кана (1962, 1965), який пропонує концепції «мислення про немислиме» та ескалації конфліктів, підкреслюючи важливість раціональних стратегій в умовах ядерної загрози.

Дослідження ефективності ядерного стримування в умовах міжнародних відносин представлені у роботах Кеннета Уолца (1981, 1990), який аналізує поширення ядерної зброї та її вплив на міжнародну стабільність. Роботи Роберта Пауелла (1985, 2003) також значно сприяли розумінню теоретичних основ стратегічного ядерного стримування. Ці роботи також дозволяють зрозуміти

вихідні принципи стримування, від яких відштовхується аналіз сучасних автоматизованих систем.

Важливий внесок у дослідження джерел та наслідків ядерного стримування внесли роботи Феарона (1994, 1997), що розглядають роль внутрішньої політики та сигналів у міжнародних конфліктах, та Данилович (2001), що досліджує джерела довіри до загроз в умовах розширеного стримування. Зокрема, сучасні дослідження, такі як Даль Бо і Фречетт (2018), систематизують детермінанти співпраці в нескінченно повторюваних іграх, що має значення для розуміння тривалих стратегій стримування. Робота Крайга (1999) застосовує теоретичні моделі до умов, що характерні для країн, що розвиваються, надаючи нові перспективи для аналізу та пропонуючи формалізовані моделі стратегічної взаємодії, які є корисними для аналізу логіки автоматизованих рішень і стратегічних дилем у ядерній сфері.

Окремий блок літератури присвячений автоматизованому ядерному стримуванню та системі «Dead Hand». Ключовими джерелами тут є роботи Брюса Блера (1993), який аналізує логіку випадкової ядерної війни, Девіда Гоффмана (2009), що детально розкриває історію створення та функціонування системи «Периметр», а також дослідження російських стратегічних ядерних сил у працях Брюса (2004). Ці джерела дозволяють оцінити ризики автоматичного реагування та втрати людського контролю. Сучасний вимір дослідження забезпечують праці, присвячені штучному інтелекту та автономним військовим системам. Роботи Майкла Горовіца (2018), Кіта Пейна (2018) і Пола Шарра (2018) аналізують вплив ШІ на баланс сил, ядерну стабільність і процеси ухвалення рішень.

Однак незважаючи на значну кількість наукових праць, зберігається аналітичний розрив між класичними теоріями ядерного стримування та новими технологічними реаліями, пов’язаними з розвитком штучного інтелекту. Це дослідження спрямоване на подолання цієї прогалини шляхом інтеграції теорій стримування з аналізом автоматизованих систем нового покоління.

Питання дослідження (гіпотеза):

- Яким чином розвиток штучного інтелекту та автоматизованих систем управління трансформує логіку ядерного стримування та впливає на стратегічну стабільність у сучасних міжнародних відносинах?
- Чи здатне автоматизоване ядерне стримування, інтегроване з технологіями штучного інтелекту, забезпечити стратегічну стабільність без підвищення ризику неконтрольованої ескалації (на прикладі кейсу «Dead Hand 2.0»)?

Об’єктом дослідження є автоматизоване ядерне стримування в епоху штучного інтелекту.

Предметом дослідження є автоматизовані механізми ядерного стримування та їх трансформація у результаті розвитку технологій штучного інтелекту на прикладі кейсу «Dead Hand 2.0».

Методологія дослідження. Дослідження буде здійснюватися з використанням методології теорії ігор; кейс-стаді, для детального розгляду системи «Периметр» та потенційних алгоритмічних моделей «Dead Hand 2.0»; порівняльного аналізу, для оцінки відмінностей між традиційними стратегіями ядерного стримування та новими підходами з використанням ШІ; критичного аналізу наукових джерел та стратегічних документів, що дозволить інтегрувати теоретичні та практичні аспекти ядерного стримування. Для обґрунтування вибору методології буде проаналізовано праці Томаса Шеллінга, Ф. Загаре, Р. Пауелла та інших дослідників, що спеціалізуються на стратегіях конфліктів, ядерному стримуванні та автоматизованих системах ухвалення рішень.

Хронологічні рамки дослідження. Нижня межа визначається появою автоматизованих систем ядерного стримування та формуванням концепції «Dead Hand» у СРСР. Верхня хронологічна межа пов’язана з активним впровадженням штучного інтелекту у військові системи та сучасними дебатами щодо стратегічної стабільності, що загострилися у 2020–2022 роках.

РОЗДІЛ 1. ТЕОРЕТИКО-МЕТОДОЛОГІЧНІ ЗАСАДИ ДОСЛІДЖЕННЯ АВТОМАТИЗОВАНОГО ЯДЕРНОГО СТРИМУВАННЯ

1.1. Визначення та концептуалізація поняття безпеки та ядерного стримування

Сучасне розуміння міжнародної безпеки давно вийшло за межі чисто військово-політичного вимірювання: до нього логічно входять екологічні, кібер-просторові, інформаційні та технологічні аспекти. Глобалізаційні процеси, науково-технічний прогрес і зміна балансу економічного впливу створюють нове безпекове середовище. Разом з тим такі події, як повномасштабна агресія Росії проти України та ядерні амбіції КНДР або Ірану доводять, що класичні проблеми стримування не втратили своєї актуальності. Найголовнішою причиною змін у сучасному стримуванні стає проникнення штучного інтелекту та автоматизованих систем у військову сферу - це явище ставить під сумнів основну передумову традиційної теорії про людину як раціонального суб'єкта ухвалення рішень.

Питання ядерної зброї як інструменту забезпечення стабільності розглядається в кількох теоретичних напрямках: реалізму і неореалізму (Уолц, Моргентау), теоріях ядерного стримування (Шеллінг), лібералізму (Ґроцій, Кант) та конструктивізму (Вендт). В умовах автоматизації ці теорії потребують критичного переосмислення, адже жодна з них спочатку не передбачала можливість делегування важливих рішень алгоритмічним системам. Для аналізу можливих систем на кшталт «Dead Hand 2.0» потрібно відстежити, як концепції безпеки та стримування адаптуються до технологічних реальностей XXI століття.

Представники традиційного реалізму (Моргентау, Гоббс) наголошують, що підвищення безпеки є головним завданням кожної країни в умовах анархічних міжнародних відносин, а ефективність країни визначається передусім її військовим та ядерним потенціалом. Поява ядерної зброї сильно змінила логіку міждержавного протистояння, а прихід автоматизованих систем здатний спричинити нову революцію у цій сфері. Роберт Кохлер, апелюючи до слів У. Черчилля,

нагадує: «Будьте дуже обережні з ядерною зброєю, поки не впевнені, і навіть більше ніж упевнені, що ви не маєте інших засобів для збереження миру» (Кохлер, 2016). У контексті автоматизації ця застерога набуває нового сенсу: людська обережність може бути витіснена алгоритмічними протоколами.

Класичний реаліст Ганс Моргентау розглядає інтерес, що вимірюється через силу, як базавий орієнтир для акторів у світовій політиці (Моргентау, 1978, с. 4-15). У цифрову епоху бажання до нарощування потужності перетворюється в гонку технологій, де автоматизовані системи стримування стають новим показником військових можливостей держав.

Неореалісти, насамперед Кеннет Уолц, пояснюють розширення військових сил страхом країн перед потенційним конфліктом. Прихильники наступального неореалізму на чолі з Джоном Міршаймером стверджують, що максимізація сили є стратегічно виправданою, оскільки домінування дає кращі шанси на виживання (Міршаймер, 2006, с. 72). Стосовно автоматичного стримування це створює нову проблему: країни можуть хотіти впровадити системи типу «Dead Hand 2.0» для гарантування другого удару, проте гонка алгоритмів здатна підірвати стратегічну стабільність, скорочуючи час для ухвалення рішень і підвищуючи ймовірність ненавмисної ескалації.

Кеннет Уолц, на відміну від Міршаймера, займав більш оптимістичну позицію щодо стабілізаційного потенціалу ядерної зброї. На його думку, вона робить завоювання настільки не вигідним, що саме її існування гарантує абсолютну безпеку (Уолц, 1979, с. 369). Водночас ця логіка не переноситься автоматично на системи автоматизованого стримування: якщо сама зброя стабілізує відносини, то передача рішень про її застосування алгоритмам може, навпаки, дестабілізувати їх, усуваючи людський елемент обережності та можливість дипломатичного маневру в критичні моменти.

Таким чином, реалісти й неореалісти погоджуються, що ядерна зброя є критично важливим елементом міжнародної безпеки, але автоматизація відкриває перед ними нові аналітичні виклики - зокрема щодо того, як держави реагуватимуть

на системи, які гарантують відплату, але одночасно зменшують простір для деескалації.

У сучасному науковому дискурсі стримування розглядається як стратегія, що через страх помсти переконує можливого агресора: його витрати є значно більшими за ймовірні вигоди. М. Рюле уточнює цю дефініцію: стримування - це загроза застосування сили задля запобігання небажаним діям супротивника (Рюле, 2015). Лоуренс Фрідман підкреслює, що ядерна спроможність однієї сторони є вирішальним фактором, який утримує іншу від використання ядерної зброї (Фрідман, 2018, с. 24). Проте в епоху штучного інтелекту виникає питання: чи здатна алгоритмічна система так само переконливо передавати наміри і довіру до загроз, як це роблять люди?

Теорія стримування виокремлює два базові підходи підходи. «Стримування покаранням» базується на загрозі масованої відплати, яка демотивує агресора через перспективу неприйнятних втрати; цей підхід вказує на здатність завдати другого удару навіть після першої атаки супротивника. «Стримування через відмову», натомість, спрямоване на блокування сожливостей агресора досягти своїх цілей. Автоматизовані системи типу «Dead Hand 2.0» дуже важливі для першого підходу, бо гарантують автоматичну відплату, навіть у разу знищення вищого керівництва, але ціна такої гарантії - збільшений ризик помилкової ескалації.

Стримування пов'язане з такими поняттями як «примус» (compellence) та «наступальне стримування» (coercion). Морган описує стримування як ключову стратегію контролю над ескалацією конфліктів (Морган, 2019, с. 51). Герман Кан виділив три типи стримування: тип I - запобігання прямим нападам; тип II - використання стратегічних загроз для захисту союзників; тип III - контрольоване стримування через обмежені загрози (Фрідман, 2013, с. 159). Автоматизація ускладнює розрізнення між цими видами й особливо між примусом, стримкванням і хибними сигналами - відсутність людського судження може викликати неадекватні реакції на неоднозначні ситуації.

Томас Шеллінг став ключовою постаттю у формуванні теорії ядерного стримування. Його модель «Гра в курча» показує, як криза може загострюватися залежно від бажання сторін ризикувати; при цьому особистістні риси лідерів і їхній рівень проінформованості відіграють вирішальну роль. Шеллінг зауважував: «криза виникає, коли учасники не можуть повністю контролювати події» (Шеллінг, 1966, с. 97). У контексті автоматизації ця думка набуває критичного значення: якщо людина вже не керує ситуацією цілком, автоматизована система може перетворити технічну проблему на катастрофу.

«Дилема в'язня» показує взаємодію між двома ядерними державами і демонструє очевидну перевагу уникнення порушення статус-кво, адже ціною цього може бути ядерна ескалація. Автоматизовані системи можуть змінити цю динаміку: якщо обидві сторони введуть механізми типу «Dead Hand 2.0», взаємна підозра може зрости через страх автоматичного ланцюга ескалації, що парадоксально руйнує ту стабільність, яку повинно було б забезпечувати стримування.

Ліберальна традиція підкреслює роль міжнародних організацій і співпраці в зниженні анархії та формуванні довіри. Підписання безпекових угод, а також робота таких структур, як ДНЯЗ і НАТО, позитивно впливають на передбачуваність дій міжнародних акторів. Джон Роулз обґрунтовує концепцію «ліберального ядерного стримування», спрямованого на захист демократичних держав від авторитарних режимів (Роулз, 2002, с. 9). У часи автоматизації ліберальний підхід ускладнює ситуацію: традиційна логіка нерозповсюдження погано адаптується до алгоритмічних систем, які не вписуються в існуючі правила контролю над озброєннями. Також ліберальні цінності людської гідності та відповідальності суперечать самій ідеї делегування рішень про застосування зброї масового знищення алгоритмічним системам.

Олександр Вендт, один із творців конструктивізму, стверджував, що «анархія - це те, що держави з неї роблять», підкреслюючи важливість ідей, ідентичностей та практик. Для розуміння «Dead Hand 2.0» конструктивізм є особливо цінним: автоматизація потенційно змінює соціальне конструювання ядерного стримування

оскільки алгоритмічні системи не беруть участь у соціальній взаємодії і працюють у рамках готових програмних логік без можливості інтерпретувати соціальний контекст.

Стратегічну культуру Джонстон описує як «систему знаків, що визначає довготривалі вподобання щодо ролі й ефективності військової сили» (Джонстон, 1995, с. 46). Вендт виділяє три типи структур: «ворог», «супротивник» і «друг», пов'язуючи їх з культурами Гоббса, Локка і Канта (Вендт, 1999, с. 247). Автоматизація ядерного стримування підкреслює відмінності між цими культурами: країни з різними безпековими традиціями по-різному підійдуть до дизайну та використання систем штучного інтелекту.

Безпека та ядерне стримання змінилися від простого військового зрозуміння до складної категорії, що включає технологічні, кібернетичні та алгоритмічні виміри. Усі аналізовані теоретичні напрями - реалізм, лібералізм, конструктивізм – потребують переосмислення в контексті автоматизації та штучного інтелекту. Потенційні системи на кшталт «Dead Hand 2.0» можуть гарантувати другий удар і посилювати стримання, але вони також збільшують ризики випадкової ескалації, технічних збоїв та втрати контролю людини над найбільш критичними рішеннями.

Штучний ітелект (ШІ) визначається як вміння комп'ютерних систем виконувати завдання, які традиційно потребують людського мислення: розпізнавання образів, форм, навчання на основі досвіду, ухвалення рішень й адаптацію до нових умов. На противагу класичним алгоритмічним системам з чіткими правилами, сучасні системи машинного навчання і глибинних нейронних мереж самостійно знаходять закономірності в даних без безпосереднього програмування кожного сценарію (Ґудфеллоу, Бенджіо і Курвілль, 2016).

У військовій сфері Департамент оборони США класифікує автономні системи за п’ятиступеневою шкалою - від цілком контрольованих людьми (рівень 1) до зовсім автономних (рівень 5). Для ядерного стримування найбільш проблематичними є системи рівнів 3-4, які теоретично мають шанс людського втручання, але через швидкість ухвалення рішень фактично стають повністю автономними (Шарр і Аллен, 2021).

Головною рисою сучасного штучного інтелекту є здатність до машинного навчання: система покращує свою роботу шляхом аналізу великих масивів даних без явного програмування правил поведінки (Джордан і Мітчелл, 2015). У контексті ядерного стримування це значить, що система може швидше знаходити ознаки підготовки до удару, але одночасно здатна неправильно тлумачити дані або приймати рішення на основі кореляцій без причинно-наслідкових зв’язків - що особливо небезпечно в галузі застосування ядерної зброї.

Суттєвою проблемою є «непрозорість» систем ШІ, що базуються на глибинному навчанні. Навіть розробники таких систем не можуть повністю пояснити логіку конкретного вибору, адже він є наслідком взаємодії багатьох параметрів нейронної мережі (Барредо Арріета та ін., 2020). У контексті ядерного стримування виникає фундаментальне питання: як можна вірити системі, яка може ухвалити рішення про знищення мільйонів людей, якщо алгоритм її міркування залишається незрозумілим? Крім того, такі системи вразливі до «адверсаріальних атак» - навмисного маніпулювання вхідними даними для провокації помилкових рішень (Ґудфеллоу, Шленс і Сеґеді, 2015).

Концепція «значимого людського контролю» (meaningful human control) передбачає, що люди повинні мати можливість розуміти, передбачати та коригувати рішення автономних систем при застосування летальної сили (Сантоні де Сіо і ван ден Ховен, 2018). Проте саме це правило стає найбільш вразливим у системах типу «Dead Hand 2.0», адже якщо система створена для гарантованої відплати навіть після знищення керівництва, то збереження значимого людського контролю стає проблематичним за самою своєю природою.

Важливо також розрізняти «вузький» і «загальний» штучний інтелект. Вузький ШІ, що спеціалізується на конкретних завданнях (розпізнавання об'єктів на супутникових знімках, аналіз сейсмічних даних), - це саме той тип, який використовується в сучасній військовій системі (Бострум, 2014). Навіть цей обмежений ШІ, інтегрований в управління ядерною зброєю, здатен суттєво змінити динаміку стримування - скорочуючи час для оцінки ситуації та простір для деескалації (Горовіц, 2018; Шарр, 2018).

Таким чином, ШІ як технологічна основа потенційних систем «Dead Hand 2.0» є якісно новим явищем в порівнянні з автоматизованими системами часів Холодної війни. Його адаптивність і швидкість обробки даних відкривають нові шанси для стримування, але водночас породжують фундаментальні проблеми непрозорості, вразливості до маніпуляцій та неможливість забезпечити значимий людський контроль. Саме ці характеристики мають визначати аналітичний погляд при оцінці трансформації логіки ядерного стримування у епоху автоматизації.

1.2. Теорія ігор як інструмент аналізу автоматизованого ядерного стримування

Теоретико-ігровий підхід дозволяє формалізувати процеси прийняття рішень у сфері ядерного стримування, оцінити стійкість стратегічних рівноваг та змоделювати поведінку держав у кризових ситуаціях. Водночас інтеграція автоматизованих систем та штучного інтелекту в ядерну сферу ставить перед цим підходом зовсім нові задачі. Якщо класичні моделі передбачали людей або держави як суб’єктів рішень, то автоматизація створює ситуацію, коли частину вибору делегують алгоритмам, що діють за іншою логікою й в інших обмеженнях.

У теорії ігор кожен гравець хоче отримати найбільшу вигоду, враховуючи дії інших. Ігри показують, як люди взаємодіють і приймають рішення (Осборн, 2004). Однак у випадку автоматизованого стримквання виникає важливе питання: чи може алгоритм бути «гравцем» у звичному розумінні? Системи типу «Dead Hand 2.0» не мають «вподобань» у людському розумінні - вони працюють за заданими параметрами, які можуть не відображати всі нюанси реальних стратегій.

Рівновага Неша - це стан, коли жоден гравець не хоче змінювати свою стратегію, беручи до уваги вибір інших (Неш, 1950). У класичному ядерному стримуванні це ототожнюється з взаємним утриманням від нападу через загрозу неприйнятної відплати. Автоматизація змінює суть цієї рівноваги: якщо одна чи обидві сторони передають рішення алгоритмам, стабільність рівноваги залежатиме від надійності програм, якості даних і здатності систем правильно зрозуміти

стратегічні сигнали. Ігри моделюють двома основними способами: у нормальній формі (матриця виграшів) та в екстенсивній формі (дерево гри). Остання є більш показовою для ядерного стримування, тому що вона демонструє послідовний вибір через зворотну індукцію. У контексті автоматичного стримування екстенсивна форма особливо цінна: вона дає змогу моделювати сценарії, де частину рішень ухвалюють системи типу «Dead Hand 2.0», а інші залишаються за людиною. Така змішана структура вибору створює нові точки потенційної нестабільності.

Раціональність залишається центральним поняттям теорії ігор. Осборн (2004) пояснює раціональний вибір як обрання найкращої дії відповідно до вподобань гравця. Проте у контексті автоматизованого стримування цей концепт отримує новий сенс. Алгоритмічна система може бути «раціональною» у технічному значенні - покращувати цільову функцію на основі даних, - але ухвалювати рішення, які людина визнала б нерозумними, бо система не здатна врахувати політичний контекст, культурні нюанси чи можливість дипломатичного маневру.

Традиційно теорія ігор передбачає повне знання гравцями всіх можливих дійй та виграшів. Верба (1961) описує раціонального гравця як того, хто може прорахувати всі можливі результати. Однак насправді жодна країна не знає точні вподобання іншої. Загаре (1990) називає такий вид раціональності «процедурний». Автоматизація ще більше ускладнює «картину»: навіть якщо лідери мають певне уявлення про наміри супротивника через дипломатичні канали, система «Dead Hand 2.0» може покладатися виключно на технічні індикатори загрози, без доступу до контекстуальної інформації. Більш реалістичним є припущення про «інструментальну раціональність»: гравець обирає дію, що дає кращий результат при зіткненні з двома варіантами (Люс і Райффа, 1957). Але й такий тип раціональності в автоматизованому контексті може призводити до небезпечних наслідків.: система здатна «раціонально» вибрати превентивний удар на основі неповних або хибних даних, якщо алгоритм інтерпретує ситуацію як неминучий напад.

«Дилема в'язня», вперше описана Такером у 1950 році, є класичним прикладом некооперативної гри, де рівновага Неша неефективна за Парето

(Poundstone, 1993). Прикладом слугує гонка озброєнь між СРСР і США на Рис.1.1: тут рівновага Неша знаходиться в точці «озброюватись - озброюватись», тоді як обидві сторони могли б виграти від «роззброюватись - роззброюватись». У випадку автоматизованого стримування ця дилема отримує новий ракурс. Якщо двоє сторін впровадять системи типу «Dead Hand 2.0», виникає подібна структура: якщо жодна не автоматизується - зберігається традиційна рівновага; якщо одна автоматизується - має перевагу в реакції, але втратить гнучкість. Якщо обидві автоматизуються - ризик випадкової ескалації максимальний без жодної односторонньої переваги. Як у класичній дилемі, раціональним вибором для кожної може стати автоматизація, навіть якщо спільне утримання від неї було оптимальним.

Автоматизація суттєво зменшує час взаємодії - з місяців і тижнів до секунд або мілісекунд. Це перетворює нескінченно повторювану гру на одноразову, де немає часу для побудови репутації чи довіри. Чим вище шанс продовжити гру, тим більше ймовірність співпраці (Даль Бо та Фрешетт, 2018) - але автоматизація руйнує цю залежність. Для успішності погрози вона повинна бути правдоподібною: чим більше ймовірність її виконання, тим більше шансів, що до неї не доведеться звертатися (Шеллінг, 1960). Автоматизовані системи типу «Dead Hand 2.0» роблять погрозу максимально правдоподібною, усуваючи людський фактор сумнівів або блефу - алгоритм гарантовано виконає відплату, якщо задані умови спрацюють. Однак ця «абсолютна правдоподібність» має зворотну сторону: вона виключає можливість утримання від удару в останній момент, якщо виникне нова інформація чи шанс на деескалацію.

Репутація є важливим частиною стримування: держава може посилити довіру до своєї стратегії, маючи історію послідовних дій (Майерсон, 2009). В умовах автоматизованого стримування репутація переноситься з лідерів на технологічні системи: противник повинен вірити не в рішучість президента, а в надійність алгоритму. Це створює нову динаміку, де демонстрація автоматизму може одночасно підвищувати довіру до погрози і провокувати противника на розробка контрзаходів.

Зобов'язання - це ще один механізм підвищити довіру до погроз. Данилович (2001) описує два методи: зниження витрат і «зв'язування рук». Автоматизовані системи типу «Dead Hand 2.0» є найбільш радикальною формою «зв'язування рук»: делегуючи рішення про відплату алгоритму, держава робить незворотне зобов'язання. Це максимально підвищує достовірність погрози, але одночасно унеможливлює деескалацію навіть за випадкової помилки. Яскравим прикладом служить аналіз розширеної гри між Росією і США на Рисунку 1.2, де рівновага Неша передбачає, що США вирішить не атакувати у відповідь на удар Росії, тому Росія вибере наступ (Гіббонс, 1997). Загроза з боку США є неправдоподібною за умови повної інформації. Автоматизована система «Dead Hand 2.0» кардинально змінює цю динаміку: якщо відплата гарантована алгоритмічно, то загроза стає абсолютно достовірною, незважаючи на раціональність людських виборів в момент кризи. Це змінює і стратегічний вибір Росії: розуміння гарантованої відплати може підштовхнути її до співробітництва замість нападу. Якщо б під час Карибського конфлікту 1962 року існували системи на зразок «Dead Hand 2.0», наслідки могли б бути катастрофічними. Кеннеді та Хрущов змогли знайти компроміс саме тому, що зберігали людський контроль і могли відступити від початкових позицій. Автоматизована система не мала би такої гнучкості: якби параметри загрози спрацювали, удар стався б навіть при можливості мирного вирішення. Це ілюструє основний парадокс автоматизованого стримування: те, що робить погрозу максимально достовірною (незворотність), водночас робить її максимально небезпечною (відсутність деескалаційних опцій), як показано на Рисунку 1.3.

Отже, теорія ігор надає потужний інструментарій для аналізу автоматизованого ядерного стримування, але одночасно виявляє його фундаментальні парадокси. Системи типу «Dead Hand 2.0» розв'язують класичну проблему довіри до погроз за допомогою технологічної незворотності, однак ціна цього рішення - втрата людського контролю, гнучкості і можливості деескалації. Рівновага Неша в автоматизованому стримуванні може бути більш стабільною

теоретично, але крихкішою практично - через технологічні ризики, алгоритмічні помилки та нездатність адаптуватися до непередбачених ситуацій.

1.3. Кейс-стаді та теорія ігор як метод дослідження автоматизованих систем ядерного стримування «Dead Hand 2.0»

Дослідження автоматизованого ядерного стримування в умовах поширення штучного інтелекту потребує методологічного підходу, який поєднує теоретичну строгість з емпіричною обґрунтованістю. Традиційні аналітичні інструменти міжнародної безпеки залишаються корисними, однак їх потрібно адаптувати до вивчення систем, де людські рішення частково або повністю переходять до алгоритмів. У такому випадку поєднання методу кейс-стаді з теорією ігор формує потужну аналітичну рамку, що дозволяє не тільки моделювати теоретичні сценарії, але й розглядати конкретні - як реальні так, і гіпотетичні - приклади систем ядерного стримування. У дослідженні було використано цю інтерговану методологію, спираючись на радянську систему «Периметр» як базовий кейс та використовуючи теоретико-ігрові моделі для оцінки стратегічних наслідків її сучасних варіацій.

Метод кейс-стаді характеризується різноманіттям підходів. Роберт Стейк виділяє три типи: внутрішній (intrinsic) - зосереджений на розумінні конкретного випадку саме по собі; інструментальний (instrumental) - використовує окремий випадок як засіб для розуміння ширшого феномена; колективний (collective) - передбачає аналіз кількох випадків для знаходження закономірностей. Аналіз системи «Периметр» і конструювання кейсу «Dead Hand 2.0» відповідає цьому інструментальному підходу, адже конкретні системи вивчаються задля розуміння загальних принципів автоматизованого стримування.

Джон Герінг розрізняє описові та причинні кейс-стадії (Герінг, 2007), підкреслюючи, що сила цього методу - у можливості простежити механізми причинності. У цьому дослідженні застосовано саме причинний підхід: аналіз того, як автоматизація впливає на стратегічну стабільність через певні механізми - зміну

часових параметрів, трансформацію інформаційного середовища і модифікацію структури виграшів у стратегічній взаємодії. Джордж та Беннет (2005) розвивають концепцію «process tracing» - відстеження причинних ланцюжків через детальний аналіз послідовності подій. Застосування цього підходу до «Периметру», можна простежити, як технічні характеристики системи (сейсмічні сенсори, радіаційний контроль, протоколи втрати зв'язку) призводили до автоматичної ескалації, та як ці механізми можуть бути трансформовані в системах «Dead Hand 2.0» через інтеграцію алгоритмів машинного навчання.

Класична теорія ігор (фон Нейман та Моргенштерн, 1944) встановила математичні основи для аналізу стратегічної взаємодії. Неш (1950) запропонував концепцію рівноваги для некооперативних ігор. Шеллінг (1960) адаптував цю теорію до розгляду ядерного протистояння, ввівши поняття «фокусні точки» («focal points»), «зобов'язання» («commitment») і «балансування на межі» («brinkmanship»). Еволюційна теорія ігор (Мейнард Сміт, 1982) аналізує, як стратегії змінюються з часом через відбір і адаптацію - цей підхід актуальний для розуміння динаміки технологічних змін у гонці озброєнь. Поведінкова теорія ігор (Камерер, 2003) враховує психологічні обмеження реальних гравців, що важливо для розуміння того, чому делегування рішень алгоритмам може як усунути людські упередження, так і викликати нові ризики.

Скотт Саган у «The Limits of Safety» (1993) використав кейс-стаді для аналізу інцидентів з ядерною зброєю в час Холодної війни, показавши, як проблеми в організації і технічні збої можуть призводити до ненавмисного загострення. Його висновки безпосередньо пов'язані із розумінням ризиків автоматизованих систем. Девід Гофман у «The Dead Hand» (2009) на основі архівних даних детально відтворює розвиток «Периметру» - його технічні особливості, стратегічну логіку та практику роботи. Брюс Блер («Strategic Command and Control», 1985; «The Logic of Accidental Nuclear War», 1993) демонструє через кейс-стаді багато сценаріїв, за яких технічні збої могли спровокувати ядерну війну.

Загаре та Кілгур («Perfect Deterrence», 2000) розробили формальні теоретико-ігрові моделі ядерного стримування, які враховують порядок рішень, неповну

інформацію і проблему довіри. Загаре також підтвердив ці моделі на прикладі конкретних криз - Карибської (1962) та Берлінської (1961) (Загаре, 1987). Пауелл («Nuclear Deterrence Theory», 1990; «In the Shadow of Power», 1999) формалізував дилему достовірності загроз і значення балансу сил. Інтріліґатор і Брітто (1984) розробили динамічні моделі гонки озброєнь, демонструючи, що навіть усвідомлюючи небезпеку раціональний вибір може призвести до деструктивних результатів. Снайдер і Дізінг («Conflict Among Nations», 1977) поєднали аналіз шістнадцяти міжнародних криз з теоретико-ігровими моделями, демонструючи, як кейс-стаді й теорія ігор взаємно збагачують одне одного.

Роберт Інь визначає кейс-стаді як емпіричне дослідження, що розглядає явище в глибині його справжньому контексту, особливо коли межі між явищем і контекстом не очевидні (Інь, 2014). Центральним кейсом у цьому дослідженні є радянська система «Периметр» - найвідоміший приклад автоматизованого механізму гарантованої відплати. Як стверджує Гоффман, ця система була створена для автоматичного запуску масованого ядерного удару у відповідь на атаку, яку зафіксували сейсмічними датчиками радіації та контролю зв'язку (Гоффман, 2009) . Хоча система зберігала певне людське управління через оператора в захищеній точці, але її логіка вимагала максимальне делегування критичних функцій автоматизованим механіхмам.

Аналіз «Периметру» як базового кейсу розкриває кілька важливих закономірностей: по-перше, технічна можливість створення автоматизованих систем відплати, що можуть працювати навіть за повної дезорганізації командування; по-друге, стратегічні мотиви для їхнього виготовлення - це передусім страх перед ударом противника по ключовим кадровим особам; по-третє, і переваги (гарантована спроможність до відплати, зміцнення довіри до загроз), і ризики автоматизованого стримування (небезпека випадкового спрацювання, складність деактивації, обмежена адаптивність).

«Периметр», однак, має межі для аналізу сучасних систем: він створювався в аналогову епоху і базувався на відносно простих алгоритмах. Системи типу «Dead Hand 2.0» потенційно можуть інтегрувати машинне навчання та аналіз

великих даних, що якісно змінює і їхні можливості, і ризики. Як зауважує Шарр, інтеграція ШІ дає можливість приймати рішення на основі моделей, недоступних людському сприйняттю, і з такою швидкістю, що не дозволяє людське втручання (Шарр, 2018). Тому методологія дослідження поєднує розгляд реального кейсу «Периметру» із конструюванням гіпотетичного «Dead Hand 2.0».

Конструювання потенційний кейсу «Dead Hand 2.0» базується на основі вивчення сучасних тенденцій: за даними Джонсона, усі провідні ядерні країни - США, Росія, Китай - активно вкладають гроші у розробку автономних і напівавтономних військових систем (Джонсон, 2020). Методологічно ослідження використовує кілька видів теоретико-ігрових моделей. Моделі нормальної форми аналізують одночасні рішення - зокрема адаптовану «Дилему в'язня» для ситуації автоматизації. Моделі розширеної форми показують послідовні рішення у кризових ситуаціях, де одна сторона може нападати або співпрацювати, а інша обирає між відплатою й утриманням.

Ключовою методологічною інновацією є адаптація класичних ігрових моделей для врахування особливостей автоматизованих систем. Зазвичай моделі вважають людей гравцями - з раціональністю, здатністю навчатися і висловлювати наміри. Коли частину рішень віддають алгоритмам, ці припущення потребують перегляду. У моделях виокремлюють два типи гравців: люди та автоматизовані системи з алгоритмічними правилами. Структура виграшів змінюється, щоб показати переваги й ризики автоматизації, а модель включає елемент невизначеності щодо поведінки систем: технічні збої та складність алгоритмів роблять їх рішення частково непередбачуваними.

Важливий методологічний аспект - питання про «раціональність» автоматизованих систем. Класична теорія ігор передбачає, що раціональний гравець збільшує користь з урахуванням переконань про дії інших. Але автоматизована система не має «переконань» - лише запрограмовані правила відповіді. У моделях дослідження такі системи представлені як гравці з незмінними стратегіями, що пов'язано з ідеєю «обмеженої раціональності» Саймона (1957), але у більш радикальній - технологічній - формі.

Синтез методів є поетапним. Спочатку аналіз «Периметра» дає емпіричні дані для налаштування параметрів теоретичних моделей. Потім ці моделі використовуються для гіпотетичного кейсу «Dead Hand 2.0», щоб формально отримати висновки про рівноваги, найкращі стратегії та можливі сценарії загострення. Нарешті, результати моделювання порівнюються з емпіричними свідченнями про ядерні доктрини і розвиток військових технологій. Для забезпечення валідності дослідження використовуються декілька стратегій. Перш за все, гіпотетичний кейс створюється на основі вже існуючих тенденцій, що робить його реалістичним прогностичним сценарієм. По-друге, висновки теоретичних моделей перевіряються на логіку та відповідність принципам стратегічної взаємодії. По-третє, результати зіставляються з оцінками фахівців у сфері ядерної політики та автономних систем. Методологічна тріангуляція проводиться на три рівнях: даних (архівні документи, стратегічні доктрини публікації про ШІ), методи (кейс-стаді + теорія ігор) і теорій (реалізм, лібералізм, конструктивізм, теорія ігор). Необхідно також відзначити обмеження вибраної методології. Теорія ігор спрощує дійсність до формальних моделей і передбачає, що гравці завжди раціональні - що проблематично щодо алгоритмічних систем. Метод кейс-стаді, у свою чергу, обмежений в узагальнюваності висновків (Герінг, 2004). Засекреченість справжніх даних про управління ядерною зброєю також звужує емпіричну основу. Для подолання цих проблем дослідження спирається на розсекречені документи, відкриті публікації в академічній сфері та публічні доктринальні державні заяви.

Таким чином, поєднання методу кейс-стаді та теорії ігор створює комплексну методологію для аналізу систем «Dead Hand 2.0». Кейс-стаді забезпечує детальне розуміння реальних і потенційних систем, їхніх технічних характеристик і стратегічних наслідків. Теорія ігор надає формальний інструмент для моделювання взаємодії в мовах автоматизації: виявлення нових рівноваг, оцінки стабільності конфігурацій та прогнозування сценаріїв ескалації. Адаптація класичних моделей для врахування обмеженої раціональності алгоритмів, а також ризиків технологічних збоїв створює нову аналітичну рамку, яка дозволяє зрозуміти

трансформацію ядерного стримування в епоху штучного інтелекту та сформулювати практично орієнтовані рекомендації щодо збереження стратегічної стабільності.

РОЗДІЛ 2. ІСТОРІЯ ТА СУЧАСНІ ПІДХОДИ ДО АВТОМАТИЗОВАНОГО ЯДЕРНОГО СТРИМУВАННЯ

2.1. Система «Периметр» («Dead Hand») як приклад ранньої автоматизації процесів у сфері ядерного стримування

Радянська система «Периметр», відома на Заході як «Dead Hand», є найбільш документально підтвердженим і значущим прикладом ранньої автоматизації ядерного стримування в часи Холодної війни. Її створення, робота та стратегічна логіка формують потрібну емпіричну базу для розуміння можливостей і небезпек автоматизованого стримування в епоху штучного інтелекту. Водночас аналіз «Периметру» з точки зору теоретичних рамок, які були розроблені у попередніх частинах, дозволяє зрозуміти, як ця система змінила базові моделі стримування Загаре та Крайга (Загаре, 1992; Крайг, 1999) й які уроки вона надає для концептуалізації систем типу «Dead Hand 2.0».

У рамках цього дослідження система «Периметр» розглядається не тільки як історичний елемент стратегічної інфраструктури Холодної війни, але і як перший реальний прототип автоматизованого ядерного стримування, у якому частина функцій стратегічного ухвалення рішень була передана технічній системі. Саме тому аналіз «Периметру» дозволяє відстежити генеалогію сучасних концепцій автоматизованого стримування, що в цій роботі позначаються як «Dead Hand 2.0». Якщо радянська система представляла ранню форму алгоритмічно детермінованої автоматизації командно-контрольних процесів, то сучасні технологічні тенденції пов'язані з переходом до систем здатних робити адаптивну оцінку загроз на основі штучного інтелекту, що радикально змінює стратегічні параметри стримування (Горовіц, 2018; Джонсон, 2020; Шарр, 2018).

Розробка системи «Периметр» почалася в кінці 1970-х років на тлі великих змін у стратегічному балансі між СРСР та США. Головним каталізатором стала американська доктрина «обезголовлюючого удару», що передбачала знищення радянського верхівки та командних пунктів у перші хвилини ядерного конфлікту,

що паралізувало б здатність СРСР до організованої відплати (Блер, 1993). Розгортання американських ракет Pershing II в Західній Європі на початку 1980-х років надало цій загрозі конкретного технологічного втілення: час польоту до Москви становив лише 8-10 хвилин, що фізично не давало шансу ухвалити рішення про удари через традиційні командні ланцюги (Гоффман, 2009).

Саме ця стратегічна вразливість стала головним приводом для радянського військово-промислового комплексу створити систему, яка б гарантувала відплату навіть у випадку повного знищення політичного та військового керівництва країни. У термінах теорії ігор, про яку йшлося раніше, йшлося про те, щоб вирішити проблему надійності загрози другого удару - головної умови стабільної рівноваги у моделі Загаре та Крайга (Загаре, 1992; Крайг, 1999). Якщо супротивник думав, що радянська загроза відплати є неправдоподібною через можливість успішного обезголовлюючого удара по верхівці влади, то рівновага стримування (CC) могла перетворитися на ситуацію «блефу», де США отримали б мотив до дезертирства (DC), розраховуючи на відсутність організованої відповіді. Система «Периметр» повинна була технологічно усунути цю асиметрію, зробивши загрозу відплати абсолютно достовірною, незалежно від стану командної структури (Блер, 1993; Гоффман, 2009).

Розробка системи йшла під контролем науково-дослідного інституту в Ленінграді «Стріла» та Командування ракетних військ стратегічного призначення СРСР. Офіційно система була поставлена в бойове чергування в 1985 році (Гоффман, 2009). Рішення про створення «Периметру» показувало не лише технологічну відповідь на конкретну стратегічну загрозу, але і глибший зсув у радянській культурі стратегій - визнання того, що автоматизація може компенсувати вразливі місця людських командних структур у сучасній ядерній війні (Блер, 1993).

Система «Периметр» була складною технічною системою, що простягалася по всій радянській території і сполучалась із системами раннього попередження та контролю ракетними силами. Основою цієї системи була мережа датчиків, які постійно слідкували за показниками, що могли означати ядерний напад: сейсмічні

коливання від ядерних вибухів, підвищення рівня радіації, електромагнітні хвилі типові для ядерних вибухів; а також стан каналів зв'язку між командними пунктами (Блер, 1993). Останній показник мав велике значення: якщо система виявляла, що зв’язок між командами обірваний, і датчики одночасно помічали ознаки ядерної атаки, це ставало тригером для активації автоматичного режиму.

Центральним елементом системи була командна ракета – спеціальний балістичний снаряд, що мав сильний радіопередавач. Якщо її активувати, ця ракета піднімалася високо вгору та відправляла команди на запуск усіх шахтних ракет і ракет з підводних кораблів, які були готові до бою (Гоффман, 2009). Таким чином, навіть якщо всі звичайні командні пункти знищувалися, «командна ракета» ставала заміною вищого керівництва, забезпечуючи гарантований масований удар у відповідь. Це технічне рішення є дуже простим способом вирішення питання надійності: система не намагалася захистити контрольні центри від руйнування - вона робила їхнє знищення стратегічно безглуздим.

Принципово важливою характеристикою «Периметру» є те, що він не був повністю автоматизованою системою в строгому технічному розумінні. Система зберігала людський елемент через спеціально обладнаний командний пункт глибокого залягання, де чергував офіцер, який мусив прийняти останнє рішення про активацію автоматичного режиму після отримання відповідних сигналів від датчиків. Проте умови в яких цей офіцер мусив приймати рішення були відмінні від звичайного процесу управління: він діяв в ізоляції, без зв'язку з керівництвом, тільки на основі показань радіодатчиків.

У термінах класифікації автономних систем Департаменту оборони США «Периметр» можна представити як систему рівня 3-4: вона може автоматично робити певні рішення на основі даних від сенсорів, але теоретично залишає можливість людського втручання (Департамент оборони США, 2012). Проте, як зазначають Шарр і Аллен, у реальній кризі, коли людина-оператор є в ізоляції під тиском часу і без можливості проконсультуватись з вищим керівництвом, цей «людський контроль» є переважно номінальним (Шарр і Аллен, 2021). Отже, «Периметр» демонструє те, що називають «парадоксом делегованого контролю».

формально система має людський елемент, але функціонально наближається до повної автономії через екстремальні умови, в яких цей елемент мусить діяти.

Попри те, що система «Периметр» була зроблена в часи аналогової технологічної епохи, її стратегічна логіка показує ранній перехід від людсько-центричної моделі ядерного стримування до моделі делегованого прийняття рішень (Блер, 1993; Саган, 1993). У класичних системах стримування остаточне рішення про використання ядерної зброї залишалося виключно за політичним керівництвом країни. «Периметр» вперше ввів можливість ситуації, коли запуск міг бути виконаний без подальшої участі людей після активації системи; це фактично змінило традиційну логіку контролю над ядерними силами (Гоффман, 2009).

Важливо зазначити, що «Периметр» представляв тип автоматизації, але не повну автономію. Його робота базувалась на чітко встановлених логічних умовах: система реагувала на конкретний набір заздалегідь заданих параметрів без можливості самонавчання або адаптації. На відміну від цього, потенційні системи «Dead Hand 2.0» передбачають використання алгоритмів машинного навчання, які здатні інтерпретувати складні патерни поведінки супротивника, що переводить проблему стримування з площини детермінованих тригерів у сферу ймовірного прогнозування (Рассел і Норвіг; Джонсон, 2020).

Таким чином, якщо «Периметр» можна розглядати як систему rule-based automation (автоматизації на основі правил), то сучасні ідеї автоматизованого стримування передбачають перехід до adaptive autonomy (адаптивної автономії), в якій алгоритмічні системи можуть впливати на формування стратегічних рішень, а не лише виконувати їх (Шарр, 2018; Горовіц, 2018). Саме ця зміна створює нову стратегічну дилему: система вже не просто реалізує заздалегідь визначене рішення, а потенційно бере участь у процесі його створення.

Аналіз «Периметру» через призму теорії ігор допомагає знайти кілька важливих механізмів, через які система змінила логіку стримування. Перш за все, система вирішила питання «неправдоподібної загрози», яку Шеллінг назвав головною дилемою ядерного стримування (Шеллінг, 1960). У базовій моделі Загаре

та Крайга рівновага СС є стійкою тільки тоді, коли обидві загрози є реальними (Загаре, 1992; Крайг, 1999). «Периметр» технологічно прибрав спокусу до дезертирства: незалежно від результату першого удару, автоматична відповідь обов'язково станеться.

Одночасно «Периметр» ілюструє парадокс, що описаний у підрозділі 1.2: чим абсолютнішою є достовірність загрози, тим менше шансів на деескалацію (Шеллінг, 1960; Пауелл, 1990). Система, створена для забезпечення миру через гарантування відплати, також виключала можливість того, щоб людська мудрість, дипломатія або технічна помилка могли бути виправлені до катастрофи. Якби під час Карибської кризи в 1962 році СРСР мав повністю активований «Периметр», навіть готовність Хрущова і Кеннеді до компромісу не могла б зупинити автоматичну ескалацію. Саме такий сценарій демонструє, чому абсолютна достовірність загрози та стратегічна стабільность є не синонімами, а потенційно суперечливими цілями.

Система також мала важливий сигнальний вимір. Хоча деталі «Периметру» залишалися засекреченими й були публічно підтверджені лише після розпаду СРСР, стратегічна спільнота на Заході знала про існування таких систем. Це виконувало функцію «зв'язування рук» у термінах Феарона: СРСР показував, що відплата станеться автоматично, що робило будь-який план на успіх обезголовлюючого удару стратегічно неспроможним (Феарон, 1997). Таким чином, «Периметр» працював не тільки як технічна система відплати, але і як стратегічний сигнал, який переформатовував розрахунки ворога - саме те, що теорія ігор визначає як ефективне зобов'язання (Шеллінг, 1960; Варіан, 2020).

Досвід роботи «Периметру» та ширших радянських систем управління ядерною зброєю під час Холодної війни надає цінні емпіричні свідчення про ризики автоматизованих систем стримування. Найвідоміший випадок стався 26 вересня 1983 року, коли радянська система раннього попередження зафіксувала запуск п'яти американських ракет. Офіцер чергування підполковник Станіслав Петров, усупереч даними системи та встановленому протоколу, класифікував тривогу як помилкову - і виявився правий: система дала збій через неправильну інтерпретацію

відбиття сонячного світла від хмар як ракетних двигунів (Гоффман, 2009). Цей випадок демонструє важливу роль людського судження у ситуаціях, коли автоматизовані системи дають хибні позитивні сигнали. Якби «Периметр» був налаштований реагувати абсолютно автоматично, рішення Петрова покладатися на інтуїцію та досвід не змогло б запобігти катастрофі.

Схожий інцидент стався під час навчання НАТО «Able Archer 83» у листопаді 1983 року. Радянська розвідка та системи спостереження інтерпрутували ці навчання, що імітували перехід до ядерної війни, як можливу підготовку до справжнього першого удару. Рівень бойової готовності радянських ядерних сил був підвищений, і, за деякими оцінками, СРСР знаходився ближче до превентивного удару, ніж у будь-який інший момент після Карибської кризи (Гейтс, 1996) . Цей інцидент демонструє іншу категорію ризику автоматизованих систем - не технічний збій сенсорів, а системна помилка в тлумаченні загальної стратегічної ситуації. Автоматизована система без здотності різнити навчання та реальну підготовку до удару могла б викликати ескалацію на основі формально правильної, але контекстуально хибного розуміння даних.

У дослідженнях організаційної надійності ядерних систем зафіксовано багато випадків майже-катастроф, що виникали через взаємодію складних технічних і організаційних факторів у непередбачених умовах. Ці інциденти підтверджують основну думку теорії нормальних аварій Перроу про те, що в складних технологічних системах катастрофічні збої є не відхиленням, а статистичною неминучістю, що особливо актуально для автоматизованих систем ядерного стримування (Перроу, 1984; Саган, 1993). Ці інциденти показують два основні типи ризику, властиві автоматизованому стримуванню: ризик технічного збою сенсорних систем, який призводить до хибного тригеру, та ризик алгоритмічної помилки в інтерпретації складної стратегічної ситуації (Блер, 1993; Саган, 1993). Обидва види ризику залишаються актуальними для систем «Dead Hand 2.0», але можуть мати нові форми через інтеграцію алгоритмів машинного навчання: сучасні сенсори можуть бути кращими, але алгоритмічна розшифровка

стратегічного контексту може бути складнішою та менш передбачуваною (Джонсон, 2020; Шарр, 2018).

Конструктивістський підхід, розглянутий у підрозділі 1.1, відкриває важливий вимір аналізу системи «Периметр» - її зв'язок з особливою радянською стратегічною культурою. Система не була просто технічним рішенням стратегічної проблеми; вона відображала типове для радянського мислення поєднання технологічного детермінізму, централізованого командування та глибокої недовіри до можливості дипломатичного урегулювання в умовах реальної ядерної кризи.

Радянська стратегічна культура, що виникла з досвіду раптового нападу Німеччини в 1941 році та ядерного удару по Хіросімі, традиційно надавала перевагу матеріальним гарантіям безпеки над дипломатичними засобами. У цьому контексті «Периметр» відповідав глибинній логіці радянської стратегічної думки: якщо не можна довіряти ворогу, тоді єдиною гарантією є автоматична технологічна відплата. Як зазначає Вендт у своїй концепції стратегічної культури, «природа міжнародної політики більше залежить від культурних складових держав, аніж від анархії» (Вендт, 1999), і «Периметр» є підтвердженням цієї тези: система відображала специфічне радянське сприйняття міжнародної системи як гоббсіанської «культури ворожнечі», де виживання базується на абсолютних матеріальних гарантіях (Вендт, 1999). Стратегічна культура при цьому, за Джонстоном, формує «довгострокові вподобання, які визначають роль та ефективність військової сили у міжнародній політиці» (Джонстон, 1995, с. 46), і така логіка зумовила особливість радянського підходу до автоматизації стримування.

Американська стратегічна культура, більш схильна до ліберального оптимізму щодо міжнародного права і дипломатії, ніколи не розробила аналогічного за концепцією «мертвого замка» (Пейн, 2018). Натомість США зосередилися на забезпеченні живучості командних структур через диверсифікацію ядерної тріади, захист підземних командних пунктів та децентралізацію. Ця відмінність у підходах показує, що вибір між повністю автоматизованими системами відплати та збереженням людського контролю є не

тільки технічним, але й глибоко культурним рішенням, яке визначається стратегічними традиціями та ідентичністю - що відповідає конструктивістській інтерпретації міжнародної безпеки (Вендт, 1992; Джонстон, 1995).

Після розпаду Радянського Союзу система «Периметр», за різними оцінками, продовжила існування в тій чи іншій формі в армії Росії (Гоффман, 2011). Це свідчить про те, що стратегічна логіка автоматизованого стримання не втратила свою значущість навіть після завершення Холодної війни, і сучасна Росія розглядає її як відповідь на розвиток американських систем протиракетної оборони та високоточної зброї (Пейн, 2018; Горовіц, 2018).

Аналіз «Периметру» дає нам декілька важливих уроків для розуміння систем типу «Dead Hand 2.0». По-перше, автоматизація стримування справді вирішує питання достовірності загрози другого удару, але не усуває ризики випадкового зростання напруги через технічні проблеми або помилкове тлумачення даних - ці ризики можуть навпаки посилитися в складніших системах (Блер, 1993; Саган, 1993). По-друге, збереження формального контролю людини не є достатнім захистом від неправомірного або помилкового використання, якщо умови прийняття рішення заважають адекватно оцінити ситуацію (Шарр, 2018). По-третє, стратегічна стабільність в умовах автоматизованого стримування - це функція не лише надійності системи, але й здатності супротивника правильно оцінити її параметри (Загаре і Кілгур, 2000). По-четверте, і це є найважливішим уроком для аналізу «Dead Hand 2.0», досвід «Периметру» демонструє, що стратегічна ефективність автоматизованих систем стримування не може бути оцінена без ширшого інформаційного та дипломатичного контексту. Система може бути технічно досконалою, але стратегічно дестабілізуючою, якщо противник не має адекватного розуміння її параметрів і меж, або якщо немає механізмів зв'язку в кризових ситуаціях, або якщо система не може розпізнати реальну загрозу від технічної помилки чи навчань (Джонсон, 2020; Шарр, 2018; Горовіц, 2018). Ці виклики властиві вже «Периметру» епохи аналогових технологій отримують нові й потенційно більш небезпечні форми у системах «Dead Hand 2.0», де інтеграція алгоритмів машинного навчання додає нові виміри складності, непрозорості та

непередбачуваності до і без того дуже ризикованого технологічного комплексу (Рассел і Норвіг, 2020; Гудфеллоу, Бенджіо та Курвілль, 2016).

Аналіз системи «Периметр» показує, що головна трансформація ядерного стримування полягає не лише в швидшій реакції або надійності систем відплати, а в поступовому зміщенні місця locus of decision-making - центру ухвалення стратегічного рішення - від людини до технічної інфраструктури (Блер, 1993; Саган, 1993). У радянській системі це переміщення залишалося обмеженим: автоматизація працювала як засіб для забезпечення виконання вже ухваленого політичного рішення. Проте сучасний розвиток штучного інтелекту створює можливість ситуації, де алгоритмічні системи можуть самостійно оцінювати рівень загрози, визначати момент ескалації та рекомендувати або навіть починати відповідні дії (Горовіц, 2018; Джонсон, 2020). У цьому сенсі «Периметр» можна інтерпретувати як перехідний етап між класичним стримуванням епохи Холодної війни та потенційною моделлю «Dead Hand 2.0», де автоматизація перестає бути допоміжним знаряддям і стає складовою частиною стратегічної стабільності (Пейн, 2018; Шарр, 2018). Саме ця еволюція - від автоматизованої відплати до алгоритмічного стримування - створює цетральну проблему сучасної міжнародної безпеки, що розглядатиметься в наступному підрозділі.

Отже, система «Периметр» є не просто історичним артефактом Холодної війни, але й активною лабораторією для розуміння основних проблем автоматизованого ядерного стримування. Вона показала технічну здійсненність гарантованої автоматичної віплати, але також виявила глибокі суперечності між вимогами стратегічної достовірності та потребами зменшення напруженості між автоматизмом системи та необхідністю людського рішення у випадках неоднозначних сигналів. Ці суперечності не зникають, а тільки посилюються з переходом до нових систем, які поєднують штучний інтелект, що робить детальне дослідження досвіду «Периметру» необхідною передумовою для будь-якого серйозного аналізу стратегічної стабільності в епоху «Dead Hand 2.0».

2.2. Базові моделі ядерного стримування та їх трансформація під впливом автоматизації процесів ухвалення рішень у сфері ядерного стримування

Розуміння впливу автоматизації на ядерне стримування вимагає, перш за все, глибокого огляду базових моделей стримування, розроблених у класичній теорії. Ці моделі, хоч і створювались для дослідження людських рішень у контексті Холодної війни, надають концептуальну рамку для розуміння того, як автономізація та штучний інтелект змінюють логіку стратегічної взаємодії. Для детальнішого аналізу стратегічних взаємодій в області ядерного стримування доречно звернутись до моделей, які розробив Загар (1992) та Країг (1999), що дозволяють побудувати відповідні ігрові сценарії та перевірити можливі стратегії гравців. Ці моделі допомагають зрозуміти, які обставини потрібні для успішного використання теоретико-ігрових модель до ядерного стримування, і як вони змінюються в умовах делегування рішень автоматизованим системам.

У основній грі стримування в розширеній формі (Рис. 2.1) з послідовними виборами та ідеальною й повною інформацією є два гравці - США та Росія (RU), і в цій грі Росія робить перший вибір, хоча вибір, хто буде першим, є довільним. Гравці роблять свій вибір у раунді рішень позначеними абревіатурою назві. Ця гра ще не має можливості ядерної ескалації для жодного гравця. Гра починається з того, що Росія обирає співпрацю (C) або дезертирство (D). Співпраця означає підтримання миру, тоді як дезертирство означає ініціювання конфлікту традиційними методами ведення війни. Якщо Росія вирішить співпрацювати, США постануть перед тим самим вибором. Якщо США вирішить співпрацювати, мир збережеться за результатом CC. Вирішивши дезертирувати, Росія стикається з іншим вибором: співпрацювати, що означало б програш у цьому сценарії (CD), або дезертирувати, що призвело б до звичайної двосторонньої війни (DD). Якщо Росія вирішить дезертирувати на самому початку, США постануть перед вибором: співпрацювати чи дезертирувати, що призведе до результатів DC або DD.

Ця базова модель ще не дає інформацію, потрібну для того, щоб була рівновага, тому що виграшів немає. Нічого не зазначено про бажані результати будь-якого гравця, тому ще не можна визначити рівновагу. Загаре (1992) і Крайг (1999) зазначають, що є три умови щодо переваг гравців, які повинні бути виконані, щоб досягти рівноваги, і щоб ця рівновага стала взаємним стримуванням. По-перше, якщо один гравець дезертирує, виграш іншого гравця, який не дезертирує, має бути меншим, ніж в мирному сценарії. Так для Росії CC > CD, а для США CC > DC. Це значить, що гравці скоріше будуть в спокої, ніж піддадуться нападу і не відповідатимуть. По-друге, для теорії стримування було б доцільно в цій моделі, аби один або обидва гравці повинні мати бажання залишити гру. Якби це було не так, то потреби в стримуванні не існувало б, бо загрози від іншого гравця не було би. Це показує що Росія захоче дезертирувати, якщо США цього не зроблять, і навпаки. Тож для Росії DC > CC і для США CD > CC. Це може бути менш зрозуміло, ніж перша умова, але це потрібно, щоб модель давала осмислений результат. Також варто сказати, що навіть якщо один гравець хоче зруйнувати мир, рівновага залишиться такою ж. По-третє, виграш в конфлікті (DD) має бути більшим, ніж у випадку, коли один гравець дезертує, а інший співпрацює, тому для Росії DD > CD і США DD > DC. Це означає, що гравці радше підуть на війну, ніж піддадуться атаці іншого гравця, і таким чином сигналізує про те, що загроза взаємного конфлікту є достовірною та можливою, оскільки гравці здатні та бажають завдати шкоди іншому в конфлікті. Нарешті, модель припускає, що для обох гравців мир краще, ніж війна, тому CC > DD для обох. У результаті цих умов виходять наступні переваги: Росія = DC > CC > DD > CD; США = CD > CC > DD > DC. Вивівши ці переваги, можна зробити гру з виграшами, щоб знайти рівновагу (Загаре, 1992) і далі продемонструвати, що результатом за умова, зазначених вище, є взаємне стримування (Рис. 2.2). У цій моделі верхні цифри результатів вказують на виграш Росії, а нижні - США. Обидва гравці хочуть максимізувати свої виграші. Рівновага гри, враховуючи виграші та розв'язана за допомогою зворотної індукції, є миром (CC). Якщо Росія вибере D у першому раунді, США також виберуть D, і результат буде (2, 2), оскільки США віддають перевагу цьому (2 > 1). Коли Росія

вирішить співпрацювати, США також вирішать співпрацювати, оскільки дезертирство призведе до результату (2, 2) через відповідь Росії. Теорія стабільного взаємного стримування справедлива в цьому сценарії за припущення, що обидві загрози є надійними (Загаре, 1987).

Передбачається, що співпроможність є сталою у звичайному стримуванні, адже більшість країн мають реальні загрози. Однак рівновага змінюється, коли загрози обох учасників не вважаються достовірними. Коли лише один учасник має справжню загрозу, рівновага Неша полягає в тому, що гравець з реальною загрозою отримує перевагу. Наприклад, виграші змінюються таким чином, що призводить до рівноваги DC, якщо достовірною є лише загроза Росії, і CD, якщо є лише загроза США. Загаре (1992) описує цей сценарій як гру «блеф», тому що «блеф» гравця з неправдоподібною загрозою відповідає гравцю з вірогідною загрозою. Якщо жодна загроза не є достовірною, будь-хто, хто має перший вибір у грі екстенсивної форми, вирішить дезертирувати і отримає перевагу. Таким чином, якщо Росія йде першою, результатом є DC, а якщо США йдуть першими, результатом є CD. У цій грі симетричної недовіри, з рисами, схожими на гру в курча (Шеллінг, 1960), виграш від конфлікту стає нижчим, ніж програш у війні (для Росії DD < CD і для США DD < DC). Наслідком цього є те, що як тільки гравець, який вибрав першим, вирішив дезертирувати, іншому гравцеві краще здатися. У цій моделі базового стримування результатом є мир лише тоді, коли обидва гравці мають надійні та дієві загрози.

Ця класична модель стримування базується на кількох основних припущеннях, які стають проблематичними в контексті автоматизації процесів ухвалення рішень. По-перше, модель припускає, що учасники є раціональними особами, здатними оцінити виграші та робити вибір на основі максимізації користі. По-друге, вона передбачає, що існує ідеальна та повна інформація - обидва гравці знають структуру гри, можливі виграші та рішення один одного. По-третє, модель передбачає можливість комунікації намірів через загрози і сигнали. По-четверте, дуже важливим є довіра до загроз - здатність переконати суперника у тому, що загроза буде виконана, якщо він порушить рівновагу. Ці припущення змінюються

під впливом автоматизації й інтеграції штучного інтелекту в ядерні системи стримування.

Трансформація раціональності в автоматизованих системах є першою важливою зміною. У класичній моделі, раціональність означає, що гравець обирає дію, яка підвищує його очікувану вигоду за умов певних переконань про дії супротивника. Проте автоматизована система типу «Dead Hand 2.0» має зовсім іншу природу «раціональності». Вона не має «переконань», а лише встановлені правила реагування на конкретні сигнали. Система може бути «раціональною» в вузькому технічному значенні - покращувати цільову функцію на основі доступних даних, - але не мати людської здатності до розуміння контексту, політичного рішення або етичного міркування (Горовіц, 2018). Це створює ситуацію, коли автоматизована система може приймати технічно обґрунтовані рішення, які людина вважала б стратегічно ірраціональними через нездатність враховувати політичний контекст чи можливість дипломатичного маневру.

Модифікація структури виграшів у грі стримування під впливом автоматизації є другим критичним перетворенням. У класичній моделі виграші визначаються на основі геополітичних, економічних та військових міркувань людей. Проте впровадження автоматизованих систем змінює цю структуру кількома способами. По-перше, автоматизація підвищує виграш від стримання (результат CC), оскільки гарантує швидку та надійну відповідь, що робить загрозу максимально достовірною. По-друге, вона зменшує виграш від дезертирства (DC або CD), оскільки противник з автоматизованою системою гарантовано відповість, виключаючи можливість успішного превентивного удару. По-третє, автоматизація катастрофічно знижує виграш від взаємного конфлікту (DD), оскільки автоматичні системи можуть призвести до тотальної ескалації без можливості зупинки. Таким чином, структура виграшів може змінитися так: для обох гравців CC > DC/CD >> DD, де різниця між DD та іншими результатами стає значно більшою через ризик неконтрольованої ескалації.

Проблема довіри до загроз змінюється найбільш радикально в умовах автоматизації. У класичній моделі довіра до загрози залежала від репутації країни,

рішучості лідерів та демонстрації спроможності. Томас Шеллінг (1960) детально розглядав питання «неправдоподібних загроз» - ситуацій, коли виконання загрози було б недоцільним для того, хто погрожує, що підривало довіру до неї. Автоматизовані системи типу «Dead Hand 2.0» вирішують цю проблему радикальним чином: вони роблять загрозу абсолютно достовірною через технологічну незворотність. Якщо система налаштована автоматично виконати відплату при певних умовах, то ворог може бути впевнений, що загроза буде реалізована, тому що рішення вже не під контролем людини. Це усуває проблему блефу і робить стримування технічно більш стабільним. Проте ця абсолютна довіра має катастрофічну ціну: вона виключає можливість уникнути відплати в останню мить, якщо з'явиться нова інформація, яка свідчить про помилку, або шанс на деескалацію.

Часові параметри гри стримування змінюються під впливом автоматизації. Класична модель Загаре і Крайга передбачає послідовні рішення з достатнім часом для оцінки ситуації і вибору найкращої стратегії. Карибська криза 1962 року яскраво показала важливість часу у стримуванні: саме наявність кількох днів для переговорів, консультацій та оцінок дозволила Кеннеді і Хрущову знайти компромісне рішення, що врятувало світ від ядерної війни. Автоматизовані системи значно скорочують цей часовий простір. Якщо людські особи, які приймають рішення, мають години або дні на роздуми, то автоматизовані системи вирішують за секунди або навіть мілісекунди. Це змінює гру стримування з послідовної, де є час на налаштування стратегії, у майже одночасну, де рішення приймаються практично миттєво. У термінах теорії ігор це означає перехід від екстенсивної форми з багатьма раундами вибору до майже нормальної форми, де всі рішення приймаються одночасно. Така трансформація прибирає можливість навчання, адаптації та формування репутації через серію взаємодій.

Інформаційне середовище гри стримування також зазнає змін. Класична модель передбачає ідеальну та повну інформацію, але в реальному житті стримування завжди працювало в умовах великої невизначеності. Ефективність класичного стримування значно залежала від здатності сторін спілкуватися про

свої наміри, червоні лінії та готовність до ескалації або деескалації. Ця комунікація здійснювалася через дипломатичні канали, публічні заяви керівників, військові маневри, що демонстрували силу, та інші знаки. Автоматизовані системи типу «Dead Hand 2.0» значно ускладнюють цю комунікацію. Алгоритми не можуть інтерпретувати нюансовані знаки та не вміють гнучко змінювати свої реакції в залежності від контексту. Система може розглядати дипломатичну активність або військові маневри лише як технічні індикатори загрози, не враховуючи їхнього політичного значення як сигналів про наміри. Це створює ризик помилкової ескалації через неправильне тлумачення дій противника.

Модель Загаре та Крайга також передбачає можливість «блефу» - ситуації, коли гравець з неправдоподібною загрозою може вдавати, що його загроза є достовірною. Автоматизація змінює динаміку блефу в стримуванні. З одного боку, автоматизована система не може блефувати у традиційному сенсі: якщо система включена, вона виконає свою програму незалежно від стратегічних міркувань. Це робить загрози автоматизованих систем максимально достовірними. З іншого боку, це створює нову форму стратегічної невизначеності: супротивник не може бути впевнений, наскільки автоматизована система є дійсно автономною, чи зберігається можливість людського вручання допомога в останній момент. Держави можуть спеціально створбвати таку невизначеність, кажучи публічно про автоматизацію своїх систем для відплати, але не показуючи точні деталі їх роботи. Це створює своєрідний «технологічний обман», де ворог мусить діяти, виходячи з припущення про повну автоматизацію, навіть якщо насправді зберігається суттєвий людський контроль.

Критично важливою є зміна концепції рівноваги Неша в автоматизації стримування. У класичній моделі рівновага Неша (СС, СС) досягається, коли обидві сторони вибирають співпрацю (збереження миру), бо відхилення від цього дасть гірший результат з урахуванням стратегії противника. Ця рівновага є стабільною, якщо обидві загрози є реальними. В умовах автоматизації рівновага може стати більш стабільною у теорії - оскільки довіра до загроз дуже висока, жодна сторона не може раціонально вибрати дезертирство. Однак водночас ця

рівновага стає більш крихкою на практиці. По-перше, технічні збої або помилки в алгоритмах можуть призвести до випадкового порушення рівноваги. По-друге, кібератаки на автоматизовані системи можуть примусити їх працювати неправильно або, навпаки, блокувати їх функції. По-третє, відсутність людського втручання означає, що навіть якщо обидві сторони знають про помилку та хочуть деескалацію, автоматичні системи можуть продовжувати ескалацію за своєю логікою програмування.

Асиметрія в автоматизації створює нові стратегічні дилеми, які не враховані класичною моделлю. Якщо тільки одна сторона впроваджує автоматизовану систему стримування, структура гри змінюється повністю. Сторона з автоматизованою системою має перевагу в швидкості реакції та надійності помсти, але втрачає гнучкість і можливість підлаштовуватись. Сторона без автоматизації має шанс на маневр, але може опинитися в невигідному становищі через повільнішу реакцію. Це створює стимул до гонки автоматизації: якщо один гравець автоматизує своє стримування, інший має зробити те ж саме, щоб не опинитися у невигідному становищі. Проте, як було показано в зміненій «Дилемі в'язня», спільна автоматизація може вести до результату, який є гіршим для обох сторін, ніж взаємне утримання від автоматизації - класичний приклад неефективної за Парето рівноваги Неша.

Питання спроможності в автоматизованому стримуванні також набуває нових вимірів. У класичній моделі спроможність означала наявність ядерної зброї і засобів її доставки. Вважалося, що можливість є постійною та легкою для спостереження - через демонстрацію військової сили, випробування зброї тощо. В умовах автоматизації спроможність стає багатовимірною. Вона включає не лише фізичну наявність ядерної зброї, але й технологічну спроможність - надійність автоматизованих систем управління і контролю; алгоритмічну спроможність - вміння систем штучного інтелекту правильно інтерпретувати загрози та приймати адекватні рішення; кібернетичну спроможність - захист систем від кібератак чи маніпуляцій. Важливо, що ці виміри спроможності є менш помітними, ніж традиційна військова сила. Супротивник не може просто оцінити надійність

програм або захист від кібератак, що створює додаткову стратегічну невизначеність.

Нарешті, базова модель стримування потребує нового переосмислення у зв'язку з новими видами загроз, які створює автоматизація. Класичний підхід розглядав стримування в контексті свідомої агресії - коли один гравець навмисно обирає дезертирство. Проте автоматизовані системи можуть викликати ризик ненавмисної ескалації через технічні помилки, неправильне тлумачення даних або складну роботу автоматизованих систем. Брюс Блер (1993) детально аналізував ризики випадкової ядерної війни у контексті систем раннього попередження та командування Холодної війни. Інтеграція штучного інтелекту в ці системи може як зменшити деякі ризики (через кращу обробку даних і виявлення помилок), так і створити нові (через непередбачувану поведінку складних алгоритмів). У термінах моделі Загара і Крайга це означає потребу брати до уваги новий вид результату - випадкової ескалації, що не є наслідком раціонального вибору жодного гравця, а з'являється через технологічні чинники.

Таким чином, базова модель ядерного стримування Загаре та Крайга хоча і залишається концептуально цінною для розуміння логіки стратегічної взаємодії, потребує фундаментального переосмислення в контексті автоматизації процесів ухвалення рішень. Автоматизація змінює всі ключові частини моделі: раціональність гравців стає обмеженою програмними параметрами; структура виграшів змінюється через нові переваги та ризики автоматизації; довіра до загроз стає абсолютною, але знижується можливість деескалації; час скорочується до критичного мінімуму; інформаційне середовище ускладнюється через неспроможність алгоритмів на нюансовану комунікацію; рівновага стає теоретично стабільнішою, але практично крихкішою. Ці трансформації створюють парадокс автоматизованого стримування: системи типу «Dead Hand 2.0» можуть посилити стримування через максимізацію довіри до загроз, але водночас підвищують ризик катастрофічної ескалації через втрату контролю людини та здатності адаптуватись. Розуміння цих змін є критично важливим для

аналіз стратегічної стабільності в епоху штучного інтелекту та формулювання рекомендацій щодо мінімізації ризиків автоматизованого ядерного стримування.

2.3. Двоступенева та екстенсивні моделі стримування: порівняння та релевантність для сучасних систем

Базова модель ядерного стримування, що розглянута в підрозділі 2.2, та досвід системи «Периметр», проаналізований у підрозділі 2.1, формують необхідне підґрунтя для переходу до більш складних теоретико-ігрових конструкцій, які дозволяють моделювати ядерну ескалацію як окремий рівень стратегічної взаємодії. Двоступенева модель стримування, розроблена Загаре (Загаре, 1992) та Крайгом (Крайг, 1999), є розвитком базової моделі та надає більш реалістичне відображення логіки ядерного конфлікту, оскільки враховує можливість поетапної ескалації від звичайних до ядерних засобів ведення війни. Водночас релевантність цих моделей для аналізу сучасних автоматизованих систем стримування типу «Dead Hand 2.0» є принциповим питанням для цього підрозділу: наскільки висновки класичних теоретико-ігрових моделей зберігають аналітичну цінність, коли частину стратегічних рішень може бути передано алгоритмічним системам?

Варіант ядерного підвищення включено в гру, перетворюючи її на двоетапну гру стримування, розроблену Загаре (Загаре, 1992) та Крайгом (Крайг, 1999), яка базується на грі базового стримування. Ця розгорнута гра про ядерне стримування починається як проста гра, але має ядерні варіанти другого етапу для кожного гравця із додатковими відгалуженнями і виграшами за результати, коли одна чи обидві країни нарощують ядерну зброю (EB, BE та EE). У цій моделі будь-яка форма атаки із використанням ядерної зброї називають ядерною ескалацією, нехтуючи тим, що насправді ядерна ескалація не завжди виконується або не виконується, є багатоступневим процесом (Кан, 1962, 1965). Як і в односторонній моделі стримання існують умови щодо переваг гравців, які повинні виконуватись, щоб ядерне стримування було рівноважним результатом цієї моделі.

Спочатку варто припустити, що стримування в цій грі є симетричним, тобто загрози з боку обох гравців будуть однаково сильними і надійними.

Знову очікується, що Росія зробить перший крок, і вона знову може або відступити, або співпрацювати. Гра залишається такою ж, як у базовій грі стримування, доки Росія чи США не вирішать втекти. Тоді інша країна має можливість ескалації, використовуючи ядерну зброю (Е). Коли будь-який гравець вибрав ядерну ескалацію, інший гравець повинен зробити вибір: або відповісти ядерною зброєю (Е), або продовжити звичайні способи війни через дезертирство (В), що означає, що це не призведе до подальшої ескалації конфлікту. Крім результатів, які є у попередній моделі (СС, ВС, СВ та ВВ) виникають три нових можливих результати. Якщо США вирішать діяти через дезертирство, Росія може підсилити ядерну зброю та вибрати Е. Якщо США не вирішать мститися, результатом буде ЕВ. Якщо США оберуть помсту і виберуть Е, результатом буде ЕЕ - тобто ядерна війна. Росія стикається з тими ж рішеннями кожен раз, коли США вперше відступає, що призводить до результату ВЕ, якщо Росія не піде на ескалацію, і ЕЕ, якщо вона це зробить.

Ядерне зростання (Е) ніколи не є першим вибором у цій моделі. Щоб це стало, один гравець повинен спочатку вибрати традиційнк втечу/вихід. Це сильне припущення, але за наявності інших більш правдоподібних припущень відсутність цього варіанту не змінює результат рівноваги. Однак це упереджує гру на користь тих, хто хоче розповсюдження ядерної зброї (Крайг, 1999). У контексті автоматизованого стримування це припущення має особливе значення: якщо система типу «Dead Hand 2.0» запрограмована реагувати тільки на ядерний удар, а не на звичайний, то структура гри зберігає логіку. Проте якщо алгоритм може інтерпретувати звичайний удар як ознака ядерного нападу, тоді межа між першим та другим етапами може стати значно розмитішою.

Основна структура двоетапної ескалаційної гри стає зрозумілою лиш після введення виграшів, а їх важко вивести без припущень або знання вподобань гравців. Як і в базовій моделі, треба зробити декілька припущень щодо цих переваг (Загаре, 1992; Крайг, 1999). Існують три ключові припущення. По-перше, країна

отримує більше користі від односторонньої ескалації, ніж від звичайної війни або співпраці: RU (EB) > RU (BB, CB); USA (BE) > USA (BB, BC). Ця перевага робить ядерні варіанти реалістичними і означає, що країни скоріш за все «поб'ють» свого супротивника, аніж підуть на звичайну війну, не кажучи вже про те, щоб визнати поразку, якщо супротивник втече першим. Це припущення має свої недоліки, тому що невідомо, чи держава дійсно радше переможе у війні через ядерну ескалацію, ніж врозіграє звичайну війну. В умовах автоматизації ця перевага може бути «вбудована» в алгоритмічну систему як цільова функція, що потенційно знижує рівень для ядерної ескалації у порівнянні з людським прийняттям рішень, де психологічний бар'єр використання ядерної зброї традиційно вважається дуже високим. По-друге, нові результати, представлені у двоетапній моделі стримування, призводять до виграшів, які завжди менші, ніж виграші результатів першого етапу: RU, USA (EE) < RU, USA (CC, BB, BC, CB). Це робить взаємне знищення менш бажаним, ніж будь-який результат основної гри стримування через звичайну війну, та гарантує врахування великих витрат ядерної війни (Крайг, 1999). Дана умова відображає логіку поняття взаємного гарантованого знищення (MAD - Mutual Assured Destruction), відповідно до якого стратегічна стабільність базується на рівноправній неприйнятності ядерного конфлікту для обох сторін (Джервіс, 1989). Однак тут виникає принципова проблема автоматизованого стримування: алгоритмічна система може не мати вбудованого «страху» перед взаємним знищенням у тому сенсі, в якому цей страх є у людей, що приймають рішення. Якщо мета системи «Dead Hand 2.0» налаштована тільки на забезпечення відплати, а не на зменшення загальних втрат, то припущення про те, що EE є найменш бажаним результатом, може не виконуватись для автоматизованого гравця (Шарр, 2018; Рассел і Норвіг, 2020). По-третє, виграш від визнання поразки або відсутності ескалації, коли інший гравець ескалує конфлікт, є меншим, ніж те ж саме на першому кроці під час звичайної війни: RU (BE) < RU (CB); USA (EB) < USA (BC). Це означає, що гравці радше визнають поразку на початку, коли завдано меншої шкоди. Без другого та третього припущень ядерна ескалація та війна були б більш реальними результатами в грі. Уподобання, представлені в цьому

останньому пункті, забезпечують те, що ядерна війна та односторонні ядерні атаки не є результатами з найбільшими виграшами в грі (Загаре, 1992).

Ці переваги ведуть до логіки схеми виграшів у двоетапній ескалаційній грі Загаре (Загаре, 1992) (Рис. 2.3). Згідно з методом зворотної індукції тут є рівноважний результат стабільного взаємного стримування, коли обидва гравці вирішують зберігати мир. У цій моделі учасниками гри є Росія (RU) і США (USA). На додаток до можливих дій співпраці (C) і звичайних бойових дій (B), тепер включено ядерну ескалацію (E). Виплати за результатами дерева гри відображають уподобання, що наведені вище (Рис. 2.4).

Результатом двохетапної моделі, показаної на Рисунку 2.4, є взаємне стримування і, таким чином, мир, як і в базовій моделі стримування. У своїй праці Загаре (Загаре, 1992) пояснює, наскільки мир чутливий до змін вірогідності загроз як на перший етап (звичайна війна), так і на другому (ядерна ескалація). Висновок полягає в тому, що мир підтримується у двох різних сценаріях: перший - ситуація з вірогідними загрозами на першому та другому етапах, а другий - коли жоден гравець не має серйозної загрози на другому етапі. Якщо ці умови відсутні, взаємне стримування не діє. Однак один ще важливий висновок полягає в тому, що за досконалої та повної інформації ядерна ескалація ніколи не здійснюється жодним гравцем. Хоча мир не завжди є результатом, відмінності в довірі та спроможності часто спричиняють односторонні чи двосторонні звичайні конфлікти.

Цей висновок є дуже важливим для розуміння логіки автоматизованого контролю. Якщо за досконалої та повної інформації ядерна ескалація ніколи не стається, то виникає питання: як автоматизація, що сильно погіршує якість інформаційного середовища через ризики алгоритмічних помилок і технічних збоїв, впливає на стабільність рівноваги? Система типу «Dead Hand 2.0» може отримувати неповну або хибну інформацію і в той самий час діяти так, ніби ця інформація є досконалою, що підвищує ризик ескалації у порівнянні з ситуацією, де людський оператор може врахувати контекстуальну невизначеність (Блер, 1993; Саган, 1993).

Крайг (Крайг, 1999) виявив, що з 14 рівноваг, які можливі при зміні загроз і можливостей гравців у двоетапній грі стримування, чотири ведуть до мирного кінця. Так само Загаре (Загаре, 1992) знаходить, що коли можливості залишаються незмінними, різна довіра веде до чотирьох рівноваг миру з десяти можливих результатів. Коли ні один із гравців не має серйозних ядерних загроз - результат завжди мирний. Це приводить до висновку, що ядерна зброя не потрібна для підтримки миру, якщо звичайні методи введення війни однаково надійні та ефективні (Крайг, 1999). Якби жоден з гравців не мав ядерної зброї, мир міг би підбримуватися без неї. Разом з тим Джервіс (Джервіс, 1989) підкреслює, що саме взаємна неприйнятність ядерного конфлікту, закладена в логіці MAD, виступає ключовим стабілізатором у випадках глибокої взаємної недовіри - там, де симетрія звичайних сил є недостатньою гарантією миру.

Коли для кожного гравця з'являється можливість ядерної ескалації, обидва гравці повинні мати симетричні й надійні загрози як на звичайному рівні так і на ядерному, щоб результатом був мир. Якщо це не так для жодного з гравців, результатом не буде взаємне стримування, адже гравець з ядерною зброєю має перевагу. Цей висновок підтверджує аргумент, що ядерна зброя може бути стабілізатором миру у ворожих ситуаціях, якщо обидві країни також мають надійні звичайні методи ведення війни (Крайг, 1999). Для аналізу систем «Dead Hand 2.0» це означає, що автоматизація стримування однієї сторони при збереженні людського контролю іншою може порушити симетрію, необхідну для стабільної рівноваги, бо дві сторони фактично грають за різними правилами (Загаре і Кілгур, 2000; Горовіц, 2018)

Крайг (Крайг, 1999) додає ще один рівень, який дозволяє також змінювати можливості ядерних арсеналів гравців. Є, звісно, реальні випадки, коли розподіл ядерної зброї не є симетричним, тому це викликає особливий дослідницький інтерес. Обидві країни в конфлікті не завжди мають ядерну зброю, і прикладом такого конфлікту є Україна та Росія. Крайг створює асиметричні моделі стримування, щоб проілюструвати таку ситуацію, і ключовим висновком цих

моделей є те, що ядерне стримування не діє якщо тільки один гравець має реальну та дієву загрозу (Крайг, 1999).

Висновки від екстенсивних моделей показують, що ядерна зброя може зберегти мир у певних ситуаціях. У грі для двох гравців, коли країна з ворожими намірами отримала ядерну зброю, потрібно, щоб інша також її отримала, інакше вона зіткнеться зі звичайним нападом. У цьому сценарії мир берігається, поки обидва гравці мають реальні та дієві ядерні загрози, що відповідає теорії ядерного стримування (Затаре, 1992; Крайг, 1999). Однак результат базової гри зі стримування показує, що мир можна зберегти навіть тоді, коли жодна з країн не отримає ядерної зброї. Ключовий висновок полягає в тому, що за досконалої та повної інформації ядерна війна чи ескалація з обох сторін ніколи не відбуваються відповідно до цих моделей (Крайг, 1999). Можливі лише звичайні удари. Ці моделі підтверджують аргументи прихильників розповсюдження ядерної зброї, бо купівля ядерної зброї ворогом вимагає від іншої країни цього ж.

Порівняння базової та двоступеневої моделей стримування дозволяє виявити кілька принципових відмінностей, які мають значення для оцінки автоматизованих систем. У базовій моделі рівновага є більш крихкою: вона залежить виключно від достовірності загроз на рівні звичайної війни, і в разі відсутності такої достовірності стримування руйнується. Двоступенева модель є стійкішою в тому сенсі, що мир може зберігатися навіть тоді, коли жоден гравець не має серйозних ядерних загроз на другому рівні, - але тільки, якщо існує симетрія на першому. Проте введення ядерного виміру також збільшує простір можливих катастрофічних результатів: тепер поряд із звичайною війною (ВВ) з'являється можливість ЕЕ - ядерної війни з найменшими виграшами для обох сторін.

Це порівняння має особливого аналітичного значення у контексті автоматизованого стримування. Якщо система «Dead Hand 2.0» працює як гравець другого рівня - тобто активується лише після того, як ворог здійснив ядерну ескалацію, - то вона теоретично не порушує основну логіку двоступеневої моделі. Однак якщо система може самостійно оцінити загрозу та почати ядерну відповідь на звичайний напад або навіть на підготовку до нього, то це фактично прибирає

межу між першим і другим рівнями гри, трансформуючи структуру рівноваги (Шарр, 2018; Джонсон, 2020). У такому сценарії виграш від звичайного дезертирства (B) для противника стає невизначеним, оскільки він може автоматично спровокувати ядерну відплату, що теоретично зміцнює стримування, але практично підвищує ризик ескалації через помилку або непорозуміння .

Порівняльний аналіз базової та двоступеневої моделей стримування в контексті їхньої релевантності для сучасних автоматизованих систем виявляє кілька фундаментальних проблем, що виходять за межі класичних теоретико-ігрових припущень. Перша з них - це проблема інформаційного середовища. Обидві моделі Загаре та Крайга працюють в умовах ідеальної та повної інформації, і саме ця умова гарантує стабільність рівноваги (Загаре, 1992; Крайг, 1999). Проте автоматизовані системи стримування за своєю природою функціонують в умовах принципово неповної і потенційно спотвореної інформації: датчики можуть давати хибні сигнали, як це продемонстрував інцидент Петрова 1983 року, алгоритми можуть неправильно тлумачити стратегічну ситуацію, як це сталося під час «Able Archer 83» (Гоффман, 2009; Гейтс, 1996). Таким чином, умова досконалої та повної інформації, яка є необхідною для стабільності рівноваги в обох моделях, є тією самою умовою, яку автоматизація ставить під найбільшу загрозу.

Друга проблема стосується припущення про раціональність. Обидві моделі передбачають, що гравці є раціональними акторами, які прагнуть максимізувати свою очікувану вигоду. Проте автоматизована система є «раціональною» лише в обмеженому технічному сенсі - вона покращує задану цільову функцію - і може приймати рішення, які людина визнала б стратегічно ірраціональними через нездатність врахувати контекстуальні та політичні нюанси (Горовіц, 2018; Рассел і Норвіг, 2020). Зокрема, ключовий висновок моделей про те, що за досконалої інформації ядерне загострення ніколи не відбувається, залежить від раціональності гравців: лідери утримуються від ескалації, бо усвідомлюють її катастрофічні наслідки. Алгоритмічна система може не мати цього знання в повній мірі, якщо воно не закладено у її цільову функцію належним чином.

Третя проблема - це асиметрія автоматизації. Моделі Загаре та Крайга показують, що асиметрія в ядерних силах підриває стабільне стримування: якщо лише один гравець має справжню і активну загрозу, взаємне стримування не працює. Подібно, асиметрія в рівні автоматизації - коли одна сторона віддає рішення алгоритму, а інша зберігає контроль людини - може створити новий вид стратегічної нерівності. Сторона з автоматизованою системою отримує вигоду в швидкості та гарантованості відплати, але втрачає гнучкість і можливість деескалації. Це може спонукати супротивника до превентивних дій, поки автоматизована система ще не активована, що підриває стабільність рівноваги так само, як і асиметрія в ядерних потенціалах у класичних моделях (Загаре і Кілгур, 2000; Пауелл, 1990).

Четверта проблема стосується темпоральної структури гри. Обидві моделі передбачають, що гравці приймають рішення по черзі з певною паузою між ходами (Загаре, 1992). Це дозволяє гравцям реагувати на дії супротивника, формувати репутацію і змінювати тактики. Автоматизовані системи типу «Dead Hand 2.0» радикально скорочують цей часовий інтервал до секунд або мілісекунд, що фактично трансформує послідовну екстенсивно гру в одночасну, де шанси для адаптації та деескалації мінімальні (Шарр і Аллен, 2021). Це означає, що стійкість рівноваги, яка в класичних моделях залежить від можливості обміну сигналами та адаптації один до одного, в умовах автоматизації може бути значно порушена навіть без зміни формальної структури виграшів.

Таким чином, порівняльний аналіз базової та двоступеневої моделей стримування дає змогу зробити кілька ключових висновків щодо їхньої доречності для автоматизованих систем. По-перше, обидві моделі зберігають аналітичну цінність для розуміння загальної логіки стримування, але їхнє застосовність до систем «Dead Hand 2.0» є обмеженою через припущення про досконалу інформацію та людську раціональність, які не відтворюються в умовах автоматизації. По-друге, двоступенева модель краще підходить для аналізу ескалаційної динаміки автоматизованого стримування, однак потребує розширення з урахуванням факторів технічного збою, а також специфіки алгоритмічного

прийняття рішень. По-третє, асиметричні варіанти моделей є найбільш доречними для опису сучасної стратегічної ситуації, що характеризується нерівномірним рівнем автоматизації між ядерними державами. Ключове положення обох моделей про те, що за умов повної та достовірної інформації, ядерна ескалація не відбувається, у контексті автоматизованих систем може бути переосмислене: що нижчої якість і надійність інформації, доступної алгоритмічні системі, то вищим стає ризик виходу за межі стабільного стримування у напрямі неконтрольованої ескалації.

2.4. Впровадження штучного інтелекту в процеси ухвалення рішень у сфері ядерного стримування

Аналіз базової та двоступеневої моделей стримування, здійснений у попередніх розділах, а також кейс системи «Периметр» як раннього прикладу автоматизації ядерного реагування створює концептуальну основу для розгляду якісно нового явища - впровадження штучного інтелекту (ШІ) в процеси ухвалення рішення у ядерній сфері. Якщо радянська система «Периметр» була прикладом rule-based automation - тобто автоматизацією на основі чітко заданих умов і детермінованих тригерів без можливості самонавчання чи адаптації, - то сучасні системи ШІ відрізняються здатністю опрацьовувати великі масиви даних, знаходити приховані закономірності та діяти в непередбачуваних ситуаціях (Рассел і Норвіг, 2020; Гудфелоу, Бенджіо і Курвілль, 2016). Саме ця якісна відмінність змінює характер проблеми: автоматизація перестає бути лише «механізмом виконання» рішень і потенційно може стати частиною механізму їх формування, що потребує переосмислення класичних моделей стримування, розглянутих раніше.

Для аналітичної читкості в цьому розділі впровадження ШІ в ядерній системі розглядається як інтеграція алгоритмічних систем у три критично важливі вузли: 1) раннє попередження й розвідка (early warning/ISR-аналітика), 2) системи командування, управління та зв'язку (C3I/NC3), та 3) підтримка або автоматизація

прийняття рішень (decision-support/decision-automation). Саме третя частина безпосередньо наближає практики держав до логіки «Dead Hand 2.0», бо стосується делегування важливих елементів вибору технічній інфраструктурі (Шарр, 2018; Джонсон, 2020).

Сучасний ШІ у військовій сфері розвивається за кількома паралельними напрямами, кожен із яких має специфічне значення для ядерного стримування. Перший і найбільш інституціоналізований напрямок - це використання алгоритмів машинного навчання у системах раннього попередження, розвідки та моніторингу. Йдеться про автоматизований аналіз супутникових знімків, обробку сигналів радіоелектронної розвідки, виявлення аномалій у поведінці мобільних ракет або підводних платформ, а також інтеграцію мультисенсорних даних в «єдину картину» ситуації (Джонсон, 2020). На перший погляд, це розвиток функцій, які були властиві автоматизованим системам ще з часів Холодної війни, але головна різниця полягає в тому, що сучасні алгоритми можуть виявляти складні патерни й комбіновані ознаки загрози, які не «записані» в систему як окремі правила (Горовіц, 2018; Гудфелоу, Бенджіо і Курвілль, 2016).

Другий напрям - впровадження ШІ в системи C3I/NC3, тобто в інфраструктуру, що забезпечує передачу наказів, стійкість зв’язку, резервування каналів, швидку маршрутизацію повідомлень та підвищення «живучості» мереж від кібератак або фізичних руйнувань (Шарр і Аллен, 2021). У цій сфері ШІ часто не «ухвалює рішення про застосування», але може змінювати часові параметри та якість інформаційного середавища, у якому працюють люди, що приймають рішення.

Третій, найбільш стратегічно вразливий напрямок - системи підтримки прийняття рішень (decision-support) або часткової автоматизації ескалаційних процедур (decision-automation). Такі системи можуть генерувати рекомендації щодо рівня загрози; пропонувати найкращі набори реакцій і запускати автоматизовані протоколи у відповідь на визначені умови. Саме тут виникає ризик зміщення locus of decision-making - центрального місця ухвалення рішення - від

людини до технічних засобів, що є ключовою ознакою концепту «Dead Hand 2.0» (Шарр, 2018; Джонсон, 2020).

Усі провідні ядерні держави - США, Росія і Китай – активно інвестують у розвиток та впровадження штучного інтелекту у військові системи, включаючи ті, що пов'язані з управлінням ядерними силами (Горовіц, 2018). США діють за планом «Third Offset», який має на меті досягнення технологічної переваги через автонамні системи та штучний інтелект (Департамент оборони США, 2012). Росія публічно оголосила амбітні цілі у сфері військового ШІ, і декілька заяв вищого керівництва країни свідчать про інтерес у створенні автономних систем ядерного реагування як розвитоку логіки «Периметру» (Гоффман, 2011; Пейн, 2018). Китай також визначає ШІ як важливу технологію подвійного використання у своїй стратегії «Військово-цивільного злиття», включаючи застосування в управлінні ядерними силами (Горовіц, 2018; Джонсон, 2020). При цьому важливо підкреслити методологічне обмеження: особливість інтеграції ШІ саме в ядерний контур (NC3) значною мірою є закритою інформацією, тому висновки базуються на загальних тенденціях, доктринальних сигналах та оцінках експертів, а не детальних технічних характеристиках (Загаре і Кілгур, 2000; Пейн, 2018).

Щоб уникнути концептуальної плутанини, важливо відрізняти наявність штучного інтелекту від ступеня автономності системи. По-перше, ШІ може використовуватися без автономного рішення - наприклад, як інструмент аналітики для людини-оператора (Рассел і Норвіг, 2020). По-друге, автономність можлива без ШІ - як у випадку rule-based автоматизації «Периметру», де система діє за певними правилами, а не вчиться (Гоффман, 2009). Найбільш ризикованою є зона поєднання двох складових: ШІ + автономність у критичному ланцюгу ескалації, оскільки тоді система не лише виконує наперед задане рішення, а потенційно формує його через ймовірну оцінку загроз (Шарр, 2018; Шарр й Аллен, 2021).

Це розрізнення є принциповим для порівняння «Периметру» і можливих систем«Dead Hand 2.0». «Периметр» демонстрував автоматизацію відплати як спосіб технічно забезпечити надійність загрози другого удару (Гоффман, 2009;

Блер, 1993). Натомість «Dead Hand 2.0» у теоретичному сенсі означає перехід від точно визначених тригерів до адаптивної оцінки і, можливо, до алгоритмічної участі в логіці ескалації (Горовіц, 2018; Джонсон, 2020). У термінах класифікації автономних систем Департаменту оборони США це аналогічно переходу від рівня 3-4, що властивий для «Периметру», до рівня 4-5, де система може не тільки виконувати, але й ситуативно коригувати стратегічні рішення (Департамент оборони США, 2012).

Для аналізу того, як ШІ змінює процеси прийняття рішень у ядерній сфері, треба зосередитися на трьох властивостях сучасних алгоритмічних систем, які були описані раніше: здатність до навчання і пристосування, непрозорість рішень, а також вразливість до маніпуляцій. По-перше, машинне навчання та адаптація. На відміну від rule-based систем, моделі машинного навчання змінюють свою поведінку на основі даних: вони можуть «вчитися» помічати складні і комбіновані індикатори, оптимізуючи певну метрику якості (Джордан і Мітчелл, 2015; Гудфелоу, Бенджіо, 2016). У сфері раннього попередження це може підвищити чутливість систему до нетипових сигналів. Водночас є основний ризик «хибних закономірностей» (overfitting, spurious correlations): система може реагувати на статистичні кореляції, що не мають причинно-наслідкового зв'язку з реальним наміром противника (Джордан, Мітчелл, 2015). У термінах двоетапної моделі стримування, що була розглянута в підрозділі 2.3, це означає збільшення ймовірності переходу з першого етапу на другий - ядерна ескалація - не через свідому стратегію суперника, а через алгоритмічну помилку в класифікації (Загаре, 1992; Крайг, 1999). По-друге, непрозорість («чорна скриня»). Для глибинних моделей часто складно відновити причинний ланцюжок, який пояснює конкретне рішення системи, адже воно є результатом взаємодії багатьох параметрів нейронної мережі (Барредо Арріета та ін., 2020). Для ядерного стримування це створює подвійну проблему підзвітності та керованості: по-перше, якщо оператори не можуть пояснити, чому система класифікувала ситуацію як «неминучий удар», зменшується можливість перевірки рішень й використання «паузи» у критичний момент; по-друге, непрозорість руйнує один з головних механізмів класичного

стримування - передбачуваність і комунікативність стратегічних сигналів. Якщо параметри алгоритмічного реагування складні для читання, ворог гірше розуміє «червоні лінії», що підвищує ризик помилкової ескалації та знищує необхідну умову стабільної рівноваги у моделях Загаре і Крайга (Загаре і Кілгур, 2000; Загаре, 1992). По-третє, вразливість до адверсаріальних атак. Системи машинного навчання можуть бути дезорієнтовані через маніпуляції вхідними даними, які людина може не помітити, але модель інтерпретує як критичні ознаки загрози (Гудфелоу, Шленс і Сегеді, 2015). У ядерній сфері це відкриває новий вектор дестабілізації: уявний противник може прагнути не фізично знищити ядерні сили, а вплинути на інформаційний «вхід» системи й викликати помилкові реакції або паралізувати здатність до відповідного удару. У логіці «Dead Hand 2.0» це є особливо критичним, бо чим ближче алгоритм до ескалаційного «тригера», тим вище стратегічна ціна одиничної помилки або компрометації. Кібератака на автоматизовану систему ядерного реагування може стати більш ефективним знаряддям стратегічного шантажу, ніж традиційна гонка озброєнь.

Центральною нормативною категорією дискусії про ШІ в системах, що використовують силу, є поняття «значимого людського контролю» (meaningful human control): люди повинні мати змогу зрозуміти, передбачити та втрутитися у рішення автономних систем, особливо якщо результатом є застосування летальної сили. У ядерному контексті ця вимога стикається з базовою стратегією автоматичного стримування: швидкість і незворотність підвищують достовірність загроз, але при цьому руйнують можливість людського «останні слова».

Упродовж розвитку ядерного управління можна умовно виділити три ступені людського контролю. Перший ступінь, який відповідає класичному стримуванню Холодної війни, - людина є єдиним суб’єктом прийняття рішень, а автоматизовані системи виконують тільки допоміжні функції збору та обробки даних. Другий ступінь, що представлений системою «Периметр», - людина формально зберігає право на остаточне рішення, проте умови його прийняття настільки погіршилися через ізоляцію, тиск часу та відсутність зв'язку з вищими керівництвом, що реальний контроль є мінімальним (Гоффман, 2009; Блер, 1993).

Третій ступінь, що відповідає потенційними системами «Dead Hand 2.0» з інтеграцією ШІ, - алгоритм може оцінювати загрозу і запускати процедурні ланцюги без підтвердження від людини у критичний момент (Шарр, 2018; Горовіц, 2018).

Водночас важливо врахувати, що навіть за формальної присутності людини у контурі рішення залишиться ризик організаційно-психологічного зсуву: під тиском часу та відповідальності оператори можуть демонструвати «automation bias» - схильність приймати пораду системи як більш авторитетну, ніж власне судження, особливо якщо система презентується як високоточна. Таким чином, «human-in-the-loop» може перетворюватися на «human-as-a-rubber-stamp», де людина швидше легітимує рішення системи, аніж критично його перевіряє (Саган, 1993; Шарр і Аллен, 2021). Саган у своєму аналізі організаційних помилок ядерних систем демонструє, що навіть перший рівень, де людина є єдиним суб'єктом рішення - не гарантує безпеки, оскільки організаційні тиски, когнітивні упередження та обмеження інформації можуть призводити до неправильних рішень. Інтеграція ШІ потенційно може усунути деякі з цих людських меж, проте водночас створюються нові ризики, пов'язані з алгоритмічними помилками, непрозорістю рішень і вразливістю до маніпуляцій. Отже, впровадження ШІ не розв'язує проблему ненадійності людського контролю, а лише трансформує її природу, додаючи інституційний вимір.

Двоступенева модель Загаре та Крайга, розглянута в підрозділі 2.3 дає зручний формальний каркас для оцінки того, як ШІ може змінювати динаміку переходу від звичайного етапу конфлікту до ядерної ескалації. Її ключовий висновок полягає в тому, що за умов досконалої та повної інформації раціональні гравці не повинні доходити до взаємної ядерної ескалації, оскільки результат ЕЕ має найнижчу корисність для обох сторін (Загаре, 1992; Крайг, 1999). Однак інтеграція ШІ здатна підірвати саме ті три припущення, які роблять цей висновок можливим. По-перше, припущення про досконалу і повну інформацію ослаблюється через непрозорість моделей, а також невизначеність щодо їхніх параметрів. Якщо супротивник не може зрозуміти, за яких обставин система бачить

ситуацію як «ескалаційну», то зростає ризик випадкового перетину «червоних ліній» - істинних чи уявних (Барредо Арріета та ін., 2020; Загаре і Кілгур, 2000). Це підтверджують також висновки підрозділу 1.2, де автоматизація характеризувалася як зміна нескінченно повторюваної гри на одноразову взаємодію, де механізми формування репутації та довіри втрачають своє значення. По-друге, припущення про людські раціональні уподобання стає проблематичним. У класичній моделі обидві сторони мають спільну основу: ЕЕ - найгірший результат (Загаре, 1992; Крайт, 1999) Алгоритмічна система, однак, є «раціональною» у вузькому сенсі оптимізації цільової функції. Якщо система працює на підвищення ймовірності гарантованої відплати або зменшення власної вразливості, її «локальна раціональність» може породжувати стратегічно небажані шляхи ескалації (Горовіц, 2018). Алгоритмічна система не завжди має такі ж «уподобання» щодо катастрофічних наслідків ЕЕ, що фундаментально підриває стабільність рівноваги, яка базується на взаємному страху перед ядерним знищенням.

По-третє, припущення про часову структуру послідовної гри зменшується через швидкість обробки і реакції ШІ. Скорочення часових вікон означає, що простір для дипломатичних сигналів, уточнення інформації та деескалацій може зникати, переводячи взаємодію у режим майже одночасних рішень, де класичні механізми стримування працюють гірше (Шарр, 2018; Шарр і Аллен, 2021). Автоматизовані системи типу «Dead Hand 2.0» радикально скорочують часовий інтервал до секунд або мілісекунд, що фактично змінює послідовну екстенсивну гру на одночасну нормальну форму. Окремо слід підкреслити, що адверсаріальні впливи додають новий вимір у гру, який класичні моделі не враховують: можливість «непрямої ескалації» через маніпуляцію даними. У такому випадку перехід до другого етапу може бути викликаний не діями держави як раціонального гравця, а технологічним порушенням системи спостереження та класифікації (Гудфелоу, Шленс і Сегеді, 2015; Джонсон, 2020).

Розглядаючи впровадження ШІ в ядерні системи через призму «Дилеми в'язня», детально проаналізованої у підрозділі 1.2, можна виявити нову форму стратегічної нестабільності - гонку алгоритмів. Якщо стара «Дилема в'язня»

моделювала ситуацію, коли обидві країни мали бажання нарощувати ядерні запаси, навіть якщо взаємне роззброєння було б оптимальним результатом (Шеллінг, 1960; Варіан, 2020), то нині та сама логіка застосовується до автоматизації систем стримування: якщо одна країна впровадить ШІ-систему, що забезпечує швидшу і надійнішу відплату, інша країна отримує стимул зробити також, щоб не бути у стратегічно невигідному становищі. Якщо дві країни автоматизують свої системи стримування, то виникає ситуація взаємодії алгоритмів, де рішення про ядерну відплату можуть приймати без участі людей – найбільш небезпечний сценарій з точки зору ризику випадкової ескалації (Горовіц, 2018; Джонсон, 2020). Парадокс полягає у тому, що індивідуально раціональна стратегія кожної держави - автоматизувати свої системи для підвищення достовірності загроз - може призвести до колективно ірраціонального результату у вигляді значного зниження порогу ескалації. Горовіц (Горовіц, 2018) описує цю динаміку як «алгоритмічну нестабільність»: навіть якщо жодна сторона не прагне до конфлікту, взаємодія автоматизованих систем може створити спіраль ескалації через ряд взаємних автоматичних реакцій, кожна з яких є «раціональною» з точки зору відповідної алгоритмічної системи, але разом вони утворюють ланцюг, що веде до катастрофи.

У термінах теорії безпеки цей процес збільшує ризики, описані в аналізі ядерних систем як «високоризикових організацій»: складність, тісні зв'язки між підсистемами та тиск часу створюють умови, де аварії можуть бути «нормальними» у розумінні Перроу (Перроу, 1984; Саган, 1993). Додавання штучного інтелекту у критичний контур не усуває ці ризики, а часто перерозподіляє їх: від людських помилок і процедурних збоїв - до алгоритмічних помилок, помилкових класифікацій і кіберкомпрометації (Барредо Арріета та ін., 2020; Гудфелоу, Шленс і Сегеді, 2015). Блер у своєму аналізі систем командування та контролю Холодної війни описував схожий феномен стосовно людей під стресом, - однак для алгоритмічних систем цей ризик є більше систематичним і структурно передбачуваним (Блер, 1993).

Ліберальна теоретична рамка, розглянута у підрозділі 1.1, зосереджує увагу на ролі міжнародних інституцій та норм у контролі за ядерною сферою.

Впровадження ШІ в ядерні системи створює прогалину між технологічними реаліями та існуючими методами контролю над озброєнням. Існуючі механізми контролю - Договір про нерозповсюдження ядерної зброї, Договір про ліквідацію ракет середньої і меншої дальності (який більше не діє), двосторонні угоди про скорочення стратегічних озброєнь між США та Росією - історично концентрувалися на кількісних показниках та носіях, але майже не регулюють ступінь автономності й алгоритмічні параметри систем управління (Пейн, 2018; Джонсон, 2020). Міжнародний дискусії про регулювання летальних автономних систем озброєння (LAWS) у рамках Конвенції про певні види звичайної зброї стосуються переважно конвенційний систем і не охоплюють ядерну сферу (Сантоні де Сіо і ван ден Ховен, 2018).

На відміну від ракет і боєголовок, які можуть бути предметом перевірок та верифікаційних процесів, алгоритми та дані є набагато менш прозорими й важче їх перевіряти. Це створює особливу проблему для стабільності: навіть якщо країни домовлялися б про обмеження автономності, практична перевірка дотримання таких обмежень є значно складнішою, ніж у «класичних» договорах про озброєнн, що фактично робить будь-яке міжнародне регулювання в цій сфері малоефективним без принципово нових механізмів верифікації (Пейн, 2018; Загаре і Кілгур, 2000).

З конструктивістської перспективи, запропонованої Вендтом, брак міжнародних норм щодо ШІ у ядерній сфері є не тільки правовою прогалиною, але й симптомом того, що ця проблема недостатньо соціально обговорена, як спільна загроза. Доки країни сприймають розробку ШІ-систем для ядерного стримування як частину конкуренції, а не як спільний ризик, що вимагає кооперативного врегулювання, формування ефективних міжнародних норм залишатиметься складним завданням (Джонстон, 1995; Вендт, 1999).

Узагальнюючи проаналізовані механізми впливу ШІ на процеси ухвалення рішень в ядерній сфері, можна виділити три принципові відмінні сценарії впровадження, що мають різні наслідки для стратегічної стабільності. Перший сценарій - ШІ як допоміжний інструмент підтримки людського рішення -

передбачає застосування алгоритмічних систем тільки для покращення якості та швидкості обробки інформації, яка надається оператору. Алгоритми поліпшують аналіз даних, але не формують ескалаційного вибору. Наслідки для стратегічної стабільності є в основному позитивними, оскільки кращий аналіз даних зменшує ризик помилкової ескалації через неправильне трактування сигналів - за умови збереження реального контролю людини (Шарр, 2018; Горовіц, 2018). Другий сценарій - ШІ як напівавтономний оператор (гібридна модель) - передбачає делегування системі повноважень приймати певні рішення з теоретичною можливістю втручання людини, яка на практиці може бути обмеженою тиском часу й організаційними факторами. Цей план наближається до логіки «периметру», але має значно вищий рівень складності й непрозорості. Наслідки для стабільності є неоднозначними: система підвищує достовірність загроз, але водночас зменшує можливості для деескалації й збільшує ризик помилкової активації (Саган, 1993; Шарр і Аллен, 2021). Третій сценарій - повністю автономний алгоритмічний оператор (радикальна версія «Dead Hand 2.0») - відповідає ситуації, коли система самостійно оцінює загрозу та ухвалює рішення про ядерну відплату без будь-якої людської участі. Такий сценарій максимізує достовірність загроз і технічно вирішує проблему «неправдоподібної загрози», описану Шеллінгом (Шеллінг, 1960), але при цьому усуває всі традиційні механізми деескалації та значимого контролю людей, підвищуючи ризик катастрофічної ненавмисної ескалації до максимуму (Джонсон, 2020; Шарр, 2018; Горовіц, 2018).

Таким чином, впровадження штучного інтелекту в процеси ухвалення рішень у сфері ядерного стримування є фундаментальною трансформацією, яка перетворює не тільки технічні аспекти систем стримування, але й сам підхід до стратегічної взаємодії, що описується класичними теоретико-ігровими моделями. ШІ впливає одночасно на чотири виміри стратегічної взаємодії: інформаційне середовище (якість даних та їх інтерпретацію), час (скорочення термінів прийняття рішень), атрибуцію (помилкові класифікації та маніпуляції даними) і структуру виграшів (через цільові функції та процедурні пріоритети алгоритмів). Тож базові

та двоступеневі моделі стримування залишаються корисними як аналітичний каркас, проте потребують модифікацій у припущеннях про раціональність, інформацію та часову структуру взаємодії. Саме ця необхідність змін, разом із виявленим парадоксом «гонки алгоритмів», а також проблемою «значимого людського контролю», формує підґрунтя для подальшого аналізу концепції «Dead Hand 2.0», як форми алгоритмічного стримування у третьому розділі дослідження.

РОЗДІЛ 3. АВТОМАТИЗОВАНЕ ЯДЕРНЕ СТРИМУВАННЯ «DEAD HAND 2.0» ТА СТРАТЕГІЧНА СТАБІЛЬНІСТЬ

3.1. Потенційні переваги та ризики автоматизованого ядерного стримування на прикладі кейсу «Dead Hand 2.0»

Аналіз теоретичних основ ядерного стримування, здійснений в першому розділі, та дослідження історичних і сучасних методів його автоматизації, представлене у другому розділі, надають потрібну концептуальну й емпіричну базу для переходу до центрального питання цього дослідження: як потенційна система «Dead Hand 2.0» - автоматизованого ядерного стримування з інтеграцією штучного інтелекту - змінює логіку стратегічної стабільності? Відповідь на це питання потребує системного аналізу переваг і ризиків такої системи, бо тільки їхнє співвідношення визначає, чи є автоматизоване стримування нового покоління стабілізуючим або дестабілізуючим чинником у сучасній міжнародній безпеці.

У цьому підрозділі «Dead Hand 2.0» концептуалізується не як конкретна технічна система, що вже існує в завершеному вигляді, а як аналітичний конструкт - потенційне поєднання автоматизованого ядерного стримування, яке включає три важливі елементи: алгоритмічні системи оцінки загроз на основі машинного навчання, автономний або напівавтономний механізм прийняття рішень про відплату та технологічну інфраструктуру гарантованого другого удару. Це дозволяє зберегти аналітичну точність та уникнути зайвого уточнення технічних деталей, які залишаються секретною інформацією, зосереджуючи увагу натомість на стратегічній логіці і її наслідках (Пейн, 2018; Шарр, 2018; Джонсон, 2020). Першою і найбільш очевидною стратегічною перевагою системи «Dead Hand 2.0» є радикальне посилення довірливості загрози другого удару, яка є основним умовою стабільного стримування у моделях Загаре й Крайга (Загаре, 1992; Крайг, 1999) Як було детально проаналізовано у підрозділі 1.2, ключовою проблемою класичного стримування є довіра до загрози відплати: якщо противник вважає, що загроза малоймовірна - бо її виконання призвело б до взаємного

знищення або ж руйнування керівництва зробить відповідь неможливою - рівновага стримування руйнується і виникає спокуса дезертирства. Система «Dead Hand 2.0» подібно до попередника «Периметру», вирішує це завдання технологічним шляхом: передаючи рішення про відплату алгоритму, держава робить загрозу абсолютно реалістичною незалежно від людського управління. Проте на відміну від «Периметру», який реагував на чітко визначений набір фізичних тригерів, система «Dead Hand 2.0» потенційно здатна оцінювати набагато ширший вибір показників небезпеки: тип розташування збройних сил ворога, аномалії в комунікаційних патернах, кіберактивність, дипломатичні знаки та їх відсутність, зміни у радіолокаційному спостереженні (Горовіц, 2018 Джонсон, 2020). Алгоритми машинного навчання можуть поєднувати ці різні потоки даних у загальну оцінку рівня загрози, яке є якісним стрибком порівняно з логікою «Периметру». У термінах двоступеневої моделі стримування (підрозділ 2.3) це означає, що достовірність загрози підтримується не тільки на другому рівні - ядерної ескалації, - але і на першому: ворог розуміє, що будь-яка підготовка до удару буде помічена та викличе автоматичну відповідь ще до нанесення першого удару, який суттєво підвищує поріг дезертирства.

Іншою перевагою є усунення так званої «проблеми часового вікна», що є одним з ключових факторів ризику в сучасному ядерному стримуванні. Використання гіперзвукових ракет скорочує час польоту до критичних об'єктів до кількох хвилин, що фізично унеможливлює прийняття виважених рішень людьми навіть за наявності робочих командних структур (Пейн, 2018). Система «Dead Hand 2.0», яка працює в межах часу, недосяжному для людського сприйняття, теоретично гарантує удар у відповідь незалежно від тиску часу. Це ще раз переводить достовірність загрози від сфери людської рішучості до сфери технологічної незворотності, що є більш надійною основою для стримування в умовах сучасного балістичного середовища.

Третьою потенційною перевагою є зниження ризику деяких видів людської помилки у критичних ситуаціях. Як демонструє аналіз Сагана (Саган, 1993), людські оператори в умовах екстремального стресу, втоми, перевантаженням і

тиском схильні до регулярних помилок. Психологічні дослідження кризового прийняття рішень доводять, що люди можуть робити вибори, які далекі від раціональних моделей - включаючи як панічну ескалацію, так і парадоксальне блокування (Лебоу і Стайн, 1994). Алгоритмічна система без емоцій страху та когнітивних упереджень теоретично діє більш передбачувано у межах своїх запрограмованих параметрів. Для супротивника передбачуваність дій є важливим фактором: якщо реакція системи є детермінованою та відомою, то це знижує ризик на помилку, яка базується на переоцінці або недооцінці рішучості людського лідера.

Четвертою перевагою, що особливо актуальна для держав з обмеженими ядерними запасами, є здатність системи «Dead Hand 2.0» компенсувати асиметрію у звичайних силах та кількості ядерних арсеналів. Як демонструють асиметричні моделі Крайга (Крайг, 1999), коли один учасник значно поступається іншому в звичайній військовій силі, ядерне стримування може виконувати роль стратегічного вирівнювача. Проте ефективність цього вирівнювання залежить від правдоподібності загрози другого удару, а саме воно є найбільш вразливим місцем для менших ядерних держав. Автоматизована система, що гарантує відплату навіть після ліквідації командної інфраструктури, суттєво посилює стратегічну позицію такої країни і знижує привабливість превентивного удару для більшого противника. Це теоретично робить взаємодію стабільною, бо приводить рівновагу ближче до симетричного варіанту двохступеневої моделі, де мир є рівноважним результатом (Загаре, 1992).

Однак переваги системи «Dead Hand 2.0» тісно пов'язані з системою ризиків, які за своєю природою є принципово новими порівняно з ризиками класичного стримування. Для аналітичної чіткості краще розділити ці ризики на три групи: технологічні, стратегічні та системні. Технологічні ризики складають першу групу і безпосередньо пов’язані зі специфічними характеристиками ШІ, проаналізованими у підрозділі 2.4. Ризик помилки в алгоритмі є найбільш фундаментальним: система машинного навчання може неправильно класифікувати групу сигналів як ознаку неминучого ядерного удару, тоді як досвідчений аналітик

з кращим розумінням ситуації б міг побачити в цьому тренування, технічну проблему датчиків або демонстрацію сили. Подібні випадки вже були до ядерної ери: радянська система раннього попередження 1983 року, чия помилка тільки завдяки рішенню підполковника Петрова не призвела до катастрофи (Гоффман, 2009), показує, що навіть відносно прості автоматизовані системи можуть генерувати хибні тривоги з катастрофічним потенціалом. Система «Dead Hand 2.0», що базується на машинному навчанні, є значно складнішою, ніж радянська система 1983 року, що означає як і більше можливостей, так і більшу уразливість до так званого «overfitting» - виявлення статистичних кореляцій, що не мають причинно-наслідкового зв'язку щодо справжньої загрози (Джордан і Мітчелл, 2015). Якщо систему навчали на даних певного геополітичного контексту, а потім їй довелося зіткнутись із зовсім новими сигналами, її класифікація може бути непередбаченою навіть для тих, хто її створив - що є наслідком «проблеми чорної скрині», описаної Барредо Арріетою та ін. (2020). Ризик адверсаріальних атак є другий важливий технологічний ризик: система, що оцінює загрозу за допомогою аналізу вхідних даних, стає можливим об'єктом цілеспрямованої маніпуляції цими даними (Гудфелоу, Шленс і Сегеді, 2015). Ворог, який знає основну логіку алгоритму, може спробувати викликати помилкову активацію через кібернапад на сенсорну інфраструктуру, що перетворює технологічну перевагу системи на потенційну уразливість.

Стратегічні ризики формують другу категорію і є логічним наслідком як переваг, так і слабкостей системи. Найбільш фундаментальним стратегічним ризиком є усунення можливості деескалації у критичний момент. Карибська криза 1962 року є найбільш переконливим доказом цінності цієї можливості: Кеннеді та Хрущов змогли знайти компромісне рішення саме тому, що зберігали контроль над остаточним рішенням та мали час на переговори (Лебоу, і Стайн, 1994). Система «Dead Hand 2.0», що приймає рішення автоматично на базі оцінки загрози, фізично виключає подібний сценарій: якщо параметри загрози виконані, то відплата станеться незалежно від готовності обох сторін до деескалації. Це створює особливо небезпечну асиметрію між технологічними

можливостями системи та дипломатичними реаліями міжнародного врегулювання. У справжніх кризових ситуаціях сигнали часто є неоднозначними, наміри - нечіткими, а дії - наслідком внутрішньополітичних компромісів, а не просто стратегічної логіки. Людський оператор може врахувати цю невизначеність, алгоритм, що оптимізований на виявлення загроз, - ні. Отже, система, що має гарантувати стримування, може у деяких випадках ставати каталізатором тієї ескалації, яку мала зупинити.

Ризик асиметрії автоматизації є другим ключовим стратегічним ризиком. Якщо тільки одна сторона впроваджує систему типу «Dead Hand 2.0», тоді як інша зберігає людський контроль, виникає новий вид стратегічної нестабільності, який не враховують класичні моделі Загаре та Крайга. Сторона з автоматизованою системою отримає стратегічну перевагу в достовірності загроз, але сторона з людським контролем залишає за собою перевагу гнучкості і можливостей для деескалації. Ця нерівність може спонукати до превентивних дій: сторона без автоматизованої системи може намагатися нейтралізувати «Dead Hand 2.0» супротивника до її увімкнення, бо знає, що після увімкнення деескалація стає неможливою. Таким чином, сама наявність системи, що повинна забезпечити мир через гарантування відплати, може провокувати ранню ескалацію - парадокс, що є прямим наслідком механізму незворотності зобов’язання, описаного Шеллінгом (Шеллінг, 1960) та Феароном (Феарон, 1997).

Ризик непрозорості для супротивника є третім важливим стратегічним ризиком. У класичному стримуванні ефективна комунікація намірів і «червоних ліній» була не менш важливою, ніж фізичне нарощування озброєнь (Шеллінг, 1960). Проте система «Dead Hand 2.0», параметри якої засекречені навіть від своїх операторів через «проблему чорної скрині», позбавляє противника можливості адекватно оцінити умови активації автоматичної відплати. Якщо супротивник не знає, за яких саме комбінацій сигналів система починає працювати, він змушений діяти з максимальною обережністю, що зміцнює стримування, але водночас ризикує випадково перейти невідомий поріг через дії, які не є атакою: навчання, розгортання нових систем озброєнь, зміну патернів спілкування . Це руйнує один

із важливих принципів стабільного стримування – передбачуваність реакції та точність «червоних ліній» (Загаре і Кілгур, 2000).

Системні ризики утворюють третій тип і є найменш очевидними, але можуть бути найбільш небезпечними. Ризик «гонки алгоритмів» як новий вид дилеми безпеки є наслідком стратегічної логіки автоматизації. Якщо одна держава впроваджує «Dead Hand 2. 0», інша держава має бажання зробити те ж саме - не тому, що це робить загальну стабільність кращою, а тому, що залишатися позаду в автоматизації означало б погане стратегічне положення. У термінах «Дилеми в'язня» це класична ситуація неефективної за Парето рівноваги Неша: кожна країна раціонально обирає автоматизацію, але взаємна автоматизація створює стан, гірший для всіх порівняно зі взаємнеим утриманням. Результатом буде світ, де ядерні рішення приймаються в основному алгоритмами без участі людей - найбільш небезпечна конфігурація з точки зору ризику випадкової ескалації. Ризик системної крихкості є другим важливим системним ризиком і ґрунтується на теорії нормальних аварій Перроу (Перроу, 1984), яка була адаптована Саганом для ядерних систем (Саган, 1993). Згідно з цією теорією у складних технічних системах із тісними зв'язками між підсистемами аварії стають статистично неминучими через непередбачувані взаємодії між частинами, які окремо діють бездоганно. Система «Dead Hand 2.0» є класичним прикладом такої складної системи: вона об'єднує сенсорні мережі, алгоритми обробки даних, комунікаційні протоколи, командні ланцюжки та механізми запуску в один комплекс, де збій будь-якого елемента або несподівана взаємодія між елементами може мати катастрофічні наслідки. Важливо, що ця крихкість є структурною: вона не може бути усунена шляхом ретельнішого проєктування, а лише знижена через підвищення надмірності і захисних механізмів, які, своєю чергою, збільшують складність і потенційно додають нові точки відмови.

Зіставлення виявлених переваг та ризиків системи «Dead Hand 2.0» через призму теоретичних моделей, розроблених у попередніх частинах, дає змогу сформулювати кілька важливих аналітичних висновків. По-перше, переваги системи є в основному теоретичними та умовними, тоді як ризики - структурними

та постійними. Стратегічна перевага - підвищення достовірності загрози - реалізується лише якщо супротивник точно знає параметри системи й вірить у її надійність. Проте саме «проблема чорної скрині», ризик кіберманіпуляцій підірвають цю основу. Навпаки, ризики (помилка в алгоритмі, адверсаріальна атака, усунення можливості для деескалації) є постійними, незалежно від поведінки ворога або якості дипломатичних комунікацій. По-друге, співвідношення вигод і ризиків системи є асиметричним відносно типу можливої помилки. Якщо система не реагує на справжню атаку помилково, наслідком є знищення держави без відплати - катастрофа, але той тип лиха, який класичне стримування намагалося запобігти. Якщо ж система помилково реагує на хибний сигнал, наслідком є ядерна війна, яку не планувала жодна зі сторін - катастрофа іншого виду, що виходить за межі будь-якої стратегічної логіки. Це асиметрія типів помилки є принциповою: система створюється, щоб мінімізувати перший тип помилки, але підвищує ризик другого. По-третє, оцінка переваг та ризиків залежить від стратегічного контексту. У випадку досить стабільної двополюсної взаємодії між двома ядерними країнами з налагодженими каналами для комунікації та розвиненою культурою управління кризами - тобто у контексті, близькому до кінця Холодної війни - перевага в достовірності загроз могла б переважувати ризики. Проте в сучасному багатополюсному ядерному середовищі, де взаємодіють кілька ядерних країн з різними стратегічними культурами, різними рівнями технологій та різним контролем над збройними силами, системні ризики гонки алгоритмів і непередбачуваних взаємодій між автоматизованими системами різних держав є значно серйознішими (Пейн, 2018; Горовіц, 2018).

Аналіз потенційних переваг та ризиків системи «Dead Hand 2.0» дозволяє сформулювати декілька ключових висновків, які мають безпосереднє значення для оцінки стратегічної стабільності в епоху алгоритмічного стримування. По-перше, система «Dead Hand 2.0» є технологічним вирішенням конкретної стратегічної задачі - достовірності загрози другого удару - але це рішення створює ряд нових проблем, які якісно різняться від тих, що воно має вирішувати. Таким чином, автоматизоване стримування нового покоління не є однозначно стабілізуючим або

дестабілізуючим чинником: воно стає трансформаційним, яке переводить джерела нестабільності з одного виміру до іншого. По-друге, важливим фактором, що впливає на співвідношення вигод і ризиків, є доступність системи для супротивника. Система, параметри якої добре відомі та перевірені, набагато менш небезпечна, ніж система з непрозорими параметрами активації - навіть якщо технічні характеристики обох є ідентичними. Це піднімає питання про необхідність нових механізмів стратегічної комунікації, які повинні бути адаптовані до реальності алгоритмічного стримування, що є темою подальшого аналізу у підрозділах 3.2 та 3.3 По-третє, оцінка ризиків системи «Dead Hand 2.0» підтверджує цетральний висновок розділу 2.3: що гіршою є якість та повнота інформації, доступної алгоритмічній системі, то більшим є ризик виходу за межі стабільного стримування в напрямку неконтрольованої ескалації. Це означає, що інвестиції у технічну надійність сенсорних систем, захист від кіберманіпуляцій та механізми верифікації рішень є не просто технічними завданнями, а умовами стратегічної стабільності в епоху «Dead Hand 2.0». По-четверте, і це є найважливішим висновком для подальшого аналізу, система «Dead Hand 2.0» у радикальній версії повної автономності є стратегічно нестабільною не через технічні помилки, а через структурну невідповідність між логікою незворотного автоматичного зобов'язання та потребою зберегти деескалаційні можливості у реальних кризах. Як показує досвід Карибської кризи, найнебезпечніша точка будь-якої кризи - не момент початкової ескалації, а момент, коли обидві сторони розуміють необхідність деескалації, проте стикаються зі структурними бар'єрами для її реалізації. Система «Dead Hand 2.0» перетворює ці перешкоди з тимчасових на постійні та непереборні. Таким чином, кейс «Dead Hand 2.0» демонструє фкндаментальний парадокс алгоритмічного стримування: прагнення до абсолютної достовірності загрози через технологічну незворотність породжує абсолютну неспроможність до деескалації, що перетворює систему стримування зі стабілізуючого в потенційно дестабілізуючий чинник. Розуміння цієї проблеми та пошук можливих рішень є цетральним завданням подальшого дослідження у підрозділах 3.2 і 3.3, де аналізуватимуться трансформація логіки ухвалення рішень

та потенційні конфігурації системи «Dead Hand 2.0», що дозволяли б зберегти переваги автоматизованого стримування, мінімізуючи при цьому структурні ризики неконтрольованої ескалації.

3.2. Трансформація логіки ухвалення рішень та стратегічної стабільності в умовах інтеграції штучного інтелекту

Аналіз потенційнвх переваг та ризиків системи «Dead Hand 2.0», здійснений у підрозділі 3.1, показав основний парадокс алгоритмічного стримування: прагнення до абсолютної певності загрози через технологічну незворотність породжує структурну нездатність до деескалації. Проте цей парадокс є лише зовнішнім проявом глибшої трансформації - якісної зміни самої логіки прийняття рішень у ядерній сфері.

Якщо підрозділ 3.1 зосереджувався на питанні «що змінюється» (переваги та ризи конкретної системи), то мета цього підрозділу - дати відповідь на питання «як саме змінюється» логіка стратегічного ухвалення рішень та чому ці зміни перетворюють саму природу стратегічної стабільності як такої. Стратегічна стабільність в класичному розумінні, що базується на працях Шеллінга (1960), Загаре та Кілгура (2000) та Пауелла (1990), є функцією взаємодії між людськими акторами - державами, які діють через раціональних лідерів, здатних до стратегічного рахунку, комунікації намірів та адаптації поведінки. Інтеграція штучного інтелекту в ядерні системи не тільки приносить новий технічний інструмент до цього середовища взаємодій - вона змінює саму онтологію стратегічної взаємодії: впливає на об'єкт рішення, часову структуру зв'язку, природу інформаційного середовища та механізми підтримання рівноваги. Саме ця трансформація є предметом аналіз цього підрозділу.

Першим і головним виміром трансформації є зміна суб’єкта, що ухвалює рішення. У класичних моделях ядерного стримування це завжди людина - президент, прем'єр-міністр або уповноважений ними командир. Ця суб’єктність має кілька важливих характеристик: здатність до розуміння контексту, що

виходить за межі формальних показників; можливість врахування політичних, культурних та психологічних чинників, яких не видно у сенсорних даних; вміння шукати нові рішення для ситуацій без прикладів; і найважливіше - спроможність думати про моральність, що включає усвідомлення катастрофічної ціни ядерного конфлікту для мільйонів людей (Шеллінг, 1966; Лебоу і Стайн, 1994).

Система «Dead Hand 2.0» трансформує цю суб'єктність у кількох послідовних кроках, кожен з яких є якісним зміщенням, а не лише технічним поліпшенням. На першому кроці ШІ працює як аналітичний інструмент допомоги для людини - збирає, обробляє і тлумачить дані, надаючи оператору структуровану картину загрози. Формально людина залишається суб'єктом рішення, однак її судження вже опосередковане алгоритмічним фреймінгом ситуації. Як продемонстрував аналіз явища «automation bias» у підрозділі 2.4, людський оператор у таких умовах часто схильний приймати рекомендацію системи більш авторитетною, ніж власне судження, особливо в умовах часового тиску (Саган, 1993). Тож навіть на цьому першому етапі частково змінюється суб'єктність: рішення формально залишається людським, але його змістове наповнення все більше залежить від алгоритмічного аналізу.

На другому кроці система переходить до напівавтономного функціонування: ШІ не тільки пропонує, але й запускає певні протоколи реагування, залишаючи людині теоретичний шанс скасування. Проте, як зазначають Шарр і Аллен (2021), в умовах стиснення часових вікон, які типові для сучасного балістичного середовища, ця можливість скасування стає все більш примарною: оператор фізично не встигає осмислити ситуацію, оцінити рекомендацію системи та прийняти незалежне рішення за кілька секунд. Таким чином, виникає те, що можна назвати «псевдолюдським контролем» - стан, коли людина є у контурі рішення, але не має реального контролю над його змістом.

На третьому кроці, що відповідає найбільш радикальній версії «Dead Hand 2.0», алгоритм стає повноцінним суб'єктом рішення про ядерну відплату. Цей крок є не просто кількісним розширенням попереднього, а якісним стрибком, який змінює природу стратегічного рішення. Алгоритм, на відмінно від людини, не має

намірів, цінностей, страхів або сподівань - він має лише цільову задачу та навчальний корпус. Це означає, що його «рішення» є не актом волі суб’єкта, що несе відповідальність, а результатом обчислення, детермінованого параметрами навчання та структурою алгоритму. У термінах теорії стримування це фундаментально змінює природу загрози: загроза більше не є показником людської рішучості - вона є проявом технологічного детермінізму (Горовіц, 2018; Джонсон, 2020).

Ця трансформація суб'єктності має далекосяжні наслідки для логіки стримування. Класичне стримування працювало через взаємне визнання людьми як суб'єктами, які можуть бути відповідальними і усвідомлювати наслідки своїх виборів. Алгоритмічний суб’єкт позбавлений таких рис: він не «страждає» від наслідків своїх дій, не несе моральної відповідальності та не здатний усвідомити серйозність ядерного знищення, що є основою концепції MAD (Mutual Assured Destruction). Джервіс (1989) аргументує, що стабілізуючий ефект MAD базується на психологічному вимірі - усвідомленні людськими лідерами справжньої ціни ядерного конфлікту. Коли алгоритм стає суб'єктом рішення, цей психологічний вимір зникає, і стабілізуючий ефект MAD може суттєво послабитися. Другий критичним виміром трансформації є зміна темпоральної структури стратегічної взаємодії. Класичні моделі стримування Загаре та Крайга, які детально розглянуті у підрозділах 2.2 та 2.3, функціонують в екстенсивній формі гри - послідовному процесі, де між ходами гравців є час для оцінки ситуації і формування відповіді, а також, за потреби, деескалації. Карибська криза 1962 року - найбільш наочний доказ критичної ролі часового проміжку: 13 днів переговорів, консультацій та взаємних сигналів дали можливість Кеннеді й Хрущову знайти вихід із ситуації, що балансувала на межі ядерної катастрофи (Лебоу і Стайн, 1994). Саме цей часовий ресурс був тим інструментом, який дозволив людській мудрості і дипломатії взяти гору над логікою ескалації. Впровадження штучного інтелекту в системи ядерного командування радикально стискає цей часовий ресурс. Сучасні системи обробки даних можуть аналізувати тисячі сенсорних потоків та формувати оцінку загрози за мілісекунди - часовий масштаб, що є принципово недосяжним для людського

сприйняття та когнітивної обробки. Якщо система «Dead Hand 2.0» налаштована реагувати на виявлення певної конфігурації загрози, ланцюг від знаходження до активації може бути завершений раніше, ніж людський оператор фізично встигне усвідомити, що відбувається (Шарр, 2018; Шарр і Аллен, 2021). У термінах теорії гри це означає основну трансформацію форми гри: послідовна гра стає одночасною взаємодією нормальної форми. Як детально проаналізовано у підрозділі 1.2, ці дві форми гри мають зовсім різні рівноваги та способи їх досягнення. В екстенсивній грі можливість спостереження за попередніми ходами суперника дозволяє формувати репутацію, навчатися і змінювати стратегії - усе це важливі механізми для підтримки стратегічної стабільності (Шеллінг, 1960; Даль Бо та Фречетт, 2018). В одночасній грі ці механізми не працюють: рішення приймаються в один час без можливості реагування на дії супротивника в реальному часі.

Ця часова трансформація має декілька стратегічних наслідків. По-перше, вона унеможливлює «дипломатичне вікно» кризового управління - час, коли сторони можуть обмінюватися сигналами, прояснити свої наміри та змінити свою поведінку до того, як з'являться незворотні наслідки. По-друге, вона підриває механізм навчання через репутацію, за допомогою якого країни в нескінчених повторюваних взаємодіях поступово створюють спільне розуміння «червоних ліній» та меж прийнятної поведінки. По-третє, вона різко знижує цінність дипломатичних зобов'язань і декларацій: якщо система реагує за секунди, то заяви лідерів про свої наміри і готовість до деескалації стають стратегічно нерелевантними - вони просто не встигають вплинути на перебіг подій.

Особливо небезпечним є ефект «каскадної автоматизації» що виникає в ситуації, коли обидві сторони мають системи типу «Dead Hand 2.0». Якщо система однієї сторони помічає загрозу та починає підготовку до відплати, сенсорні системи іншої сторони можуть зафіксувати цю активність як ознаку неминучого удару та запустити свою власну автоматичну процедуру. У результаті виникає ескалаційна спіраль, де кожна автоматична реакція провокує наступну - без будь-якої людської участі та майже без можливості зупинки. Це алгоритмічний еквівалент «ефекту доміно», що описав Блер у його аналізі систем командування Холодної війни, але

реалізованого в масштабі часу, де людське втручання є фізично неможливим (Блер, 1993).

Третім критичним виміром трансформації є якісна зміна інформаційного середовища стратегічної взаємодії. Класичне стримування функціонувало через складну мережу взаємної комунікації намірів, можливостей та обмежень, що дозволяло сторонам будувати спільне розуміння «правил гри» навіть в умовах глибокої взаємної недовіри (Шеллінг, 1960; Загаре і Кілгур, 2000). Ця комунікація проходила через багато каналів: офіційні дипломатичні заяви, публічні виступи лідерів , демонстративні військові маневри, зміни в ядерних доктринах, поведінку на переговорах з контролю над озброєннями. Кожен із цих каналів давав інформацію не лише про технічні можливості, а й про політичні наміри, готовність до компромісу та рівень рішучості.

Інтеграція штучного інтелекту змінює цю систему комунікації принаймні в трьох аспектах. По-перше, вона додає новий «мовчазний» потік інформації - алгоритмічні знаки, що виходять з поведінки самих автоматизованих систем: зміни у патернах активності сенсорних мереж, модифікації параметрів навчання алгоритмів, оновлення програм для командних систем. Супротивник, який намагається оцінити рівень загрози від держави, що має «Dead Hand 2.0», тепер повинен аналізувати не лише традиційні дипломатичні сигнали, але й технологічну поведінку системи - завдання, що є принципово складнішим і менш зрозумілим (Горовіц, 2018; Джонсон, 2020). По-друге, «проблема чорної скрині» детально проаналізована у підрозділі 2.4, підриває саму здатність прозорої комунікації параметрів системи. Держава, що володіє «Dead Hand 2.0», не може достовірно повідомити противнику про умови початку роботи системи не тому, що не хоче, але через те, що навіть її розробники не можуть повністю передбачити поведінку складного алгоритму машинного навчання у нестпндартних ситуаціях (Барредо Аррієта та ін., 2020). Це створює структурно асиметрію між наміром і сприйняттям: держава має бажання до максимальної прозорості своїх ядерних «червоних ліній», але технологічна природа системи робить унеможливлює таку прозорість. По-третє, вразливість ШІ до адверсаріальних атак, описана у підрозділі 2.4, відкриває

нову можливість для маніпулювання інформаційним середовищем стримування. Якщо противник здатний навмисно посилати певні сигнали, які активуватимуть автоматизовану систему або, навпаки, блокуватимуть її реакцію, то саме поняття «сигнал про наміри» набуває новиого стратегічно небезпечного виміру (Гудфелоу, Шленс і Сегеді, 2015). Класичне стримування будувалося на припущенні, що сторони контролюють власні сигнали та можуть відрізнити навмисну комунікацію від випадкових подій. У ситуації, де знаки можуть бути фальшивими або маніпульованими через кібератаку на сенсорну інфраструктуру «Dead Hand 2.0», це припущення втрачає силу. У термінах конструктивістської теорії, яку запропонував Вендт (Вендт, 1992, 1999) та проаналізуваної у підрозділі 1.1, зміна інформаційного середовища впливає на саму основу взаємної соціальної конструкції стримування. Вендт казав, що міжнародна система є тим, що держави «роблять» з неї через спілкування та формування ідентичностей. Стримування в цьому сенсі є соціальною практикою, яка постійно відновлюється через визнання сторін суб'єктами з певними інтересами, можливостями та обмеженнями. Коли частина цієї взаємодії делегується алгоритмам, що не беруть участі у соціальної практиці та не можуть зрозуміти соціального контексту, суспільна структура стримування повільно руйнується - і разом з нею підривається стабілізуючий ефект взаємного визнання.

Четвертим і синтетичним виміром трансформації є зміна самої природи стратегічної рівноваги в умовах алгоритмічного стримування. Рівновага Неша у класичних моделях стримування є результатом взаємного раціонального розрахунку людей, кожен з яких вибирає стратегію, що максимізує власний виграш з урахуванням очікуваних дії супротивника (Неш, 1950; Загаре, 1992). Стабільність цієї рівноваги залежить від кількох умов: раціональності гравців, якості інформації, симетрії загроз та наявності часу для стратегічного рахунку.

Інтеграція штучного інтелекту в системи стримування трансформує кожну з цих умов, породжуючи те, що можна назвати «алгоритмічною рівновагою» - принципово відмінним типом стратегічного балансу. Алгоритмічна рівновага не є результатом спільних розрахунків суб’єктів, які усвідомлюючи власні інтереси та

здатні до гнучккої адаптації: вона є наслідком взаємодії двох або більше детермінованих систем, кожна з яких покращує певну цільову задачу в обставинах, що частково залежать від дій іншої системи. Така рівновага може бути формально стабільною - жодна система не «бажає» відхилитися від свого алгоритму - але водночас надзвичайно крихкою: незначні зміни у вхідних даних чи параметрах системи можуть спровокувати каскадну зміну рівноваги, яку ніхто з людських акторів не планував і можливо не здатний передбачити.

Важливою є проблема множинності рівноваг в умовах алгоритмічного стримування. Крайг (1999) продемонстрував, що навіть у класичний двоступеневих моделях стримування існує множинність рівноваг, з яких лише частина мирні. Коли один або обидва гравці є алгоритмічними системами, кількість можливих рівноваг значно зростає - і передбачити, до якої саме рівноваги наближатиметься система в конкретних умовах стає дуже складним завданням навіть для розробників (Барредо Арріета, 2020). Це означає, що алгоритмічне стримування функціонує в умовах принципово вищої структурної невизначеності, ніж класичне стримування з людськими акторами. Ця зміна природи рівноваги підриває один із ключових механізмів класичного стримування - так звана «фокусні точки» (focal points) Шеллінга (1960). Шеллінг описував здатність сторін працювати разом на певних «природних» вирішеннях кризових ситуацій навіть без формальної комунікації: ці точки координації з'являються через спільний культурний та когнітивний простір, у якому функціонують людські актори. Алгоритмічна система не має такого культурного або когнітивного простору і не здатна до такої неформальної координації: для неї не існує «природних» вирішень, лише оптимальні з погляду запрограмованої цільової мети. Це означитає, що можливості неформальної координації на основі загального розуміння, яке є критичним ресурсом кризового управління, суттєво звужуються при алгоритмічному стримуванні.

П'ятим виміром трансформації є зміна самих умов стратегічної стабільності - переліку факторів, які визначають, чи є система стримування стабілізуючою чи дестабілізуючою. У класичних моделях, проаналізованих у розділі 2, стабільність

стримування залежить від симетрії загроз, їхньої правдоподібності та якості інформаційного середовища (Загаре, 1992; Крайг, 1999). Інтеграція ШІ до цього списку додає нові критичні умови, що є специфічними для алгоритмічного середовища.

Першою новою умовою є алгоритмічна симетрія – ступінь, у якому сторони мають порівнянні рівні автоматизації своїх систем. Як денонсує аналіз у підрозділі 3.1, асиметрія в автоматизації є потенційно більш дестабілізуючою, ніж симетрична автоматизація для обох: сторона, що контролюється людьми, зберігає деескалаційний потенціал, але знаходиться у стратегічно невигідному становищі через швидкість реакцій, які можуть стимулювати превентивні дії (Пейн, 2018; Горовіц, 2018). Згідно з двоступеневою моделлю Крайга (1999), яка ілюструє дестабілізуючий ефект асиметрії у ядерних потенціалах, відмінність у рівнях автоматизації є структурно аналогічним чинником нестабільності.

Другою новою умовою є алгоритмічна сумісність - ступінь, у якому автоматизовані системи різних країн працюють на основі сумісних або хоча б взаємнозрозумілих логік та параметрів реагування. Якщо системи двох держав оптимізовані на різних навчальних корпусах і за різними цільовими функціями, їхня взаємодія може породжувати непередбачувані ескалаційні динаміки, навіть якщо кожна система окремо працює в межах заданих параметрів (Джонсон, 2020; Шарр, 2018). Цей ризик є подібним до описаного Перроу (1984) у теорії нормальних аварій: непередбачені взаємодії між компонентами, що окремо нормально функціонують, можуть викликати системні збої.

Третьою новою умовою є захист кіберпростору алгоритмічних систем - здатність систем «Dead Hand 2.0» протистояти цілеспрямованим спробам маніпуляції через кібератаки на сенсори або командну інфраструктуру. На відміну від класичного стримування, де фізична захищеність ядерних сил була первинним параметром безпеки, в алгоритмічному стримуванні захист кіберпростору стає не менш важливим: компрометована через кібератаку система «Dead Hand 2.0» може стати інструментом провокації конфлікту в руках третьої сторони - актора, що

взагалі не є учасником двостороннього стримування (Гудфелоу, Шленс і Сегеді, 2015; Горовіц, 2018).

Синтезуючи виявлені виміри трансформації, можна виділити три сценарії впливу інтеграції ШІ на стратегічну стабільність, які відповідають трьом рівням автоматизації, що описані в підрозділі 2.4.

Перший сценарій - ШІ як аналітичний інструмент підтримки людських рішень - є потенційно стабільним, якщо зберегти справжній, а не лише формальний контроль людини. У цьому варіанті зміна суб’єктності рішення є мінімальною, а трансформація інформаційного середовища - переважно позитивною: кращий аналіз даних зменшує ризик помилкової інтерпретації сигналів противника та покращує якість інформаційної основи для людського рішення. Стратегічна рівновага залишається принципово людською, і стандартні способи стримування – комунікація намірів, дипломатичне кризове управління, репутаційне навчання – залишаються функціональними. Ключовою умовою збереження стабілізуючого ефекту цього сценарію є недопущення поступового розширення повноважень алгоритму за межі початково задуманої допоміжної ролі.

Другий сценарій - ШІ як напівавтономний оператор - є структурно амбівалентним та потенційно найбільш небезпечним, оскільки поєднує ілюзію людського контролю з реальною алгоритмічною автономією. У цьому сценарії зміна суб'єктності є суттєвою, але прихованою: «automation bias» та скорочення часового інтервалу зменшують справжній контроль людини до мінімуму, тоді як формальна присутність людини у контурі рішення створює хибне відчуття безпеки. Стратегічна стабільність тут є крихкою й залежить від точності алгоритмічних параметрів реагування, який встановлюються заздалегідь і не змінюються під конкретні кризові ситуації. Цей сценарій є найбільш реалістичним описом того, як може розвиватися практика ядерного управління у найближчі роки, що робить його центральним об'єктом аналітичної уваги (Пейн, 2018; Джонсон, 2020).

Третій сценарій - повністю автономний алгоритмічний оператор - є структурно дестабілізуючим, навіть якщо він технічно вирішує питання достовірності загрози другого удару. Трансформація суб'єктності є повною, зміна

часової структури - максимальною, а трансформація механізмів рівноваги - найрадикальнішою. У цьому варіанті класичні умови стратегічної стабільності - симетрія загроз, якість інформаційного середовища, можливість деескалації - замінюються алгоритмічними умовами, що в рази є менш стабільними та прогнозованими. Парадокс цього сценарію полягає у тому, що він підвищує достовірність загрози (стабілізуючий ефект) і одночасно максимізує ризик ненавмисної ескалації (дестабілізуючий ефект), причому друге може значно переважувати перше.

Аналіз трансформації логіки ухвалення рішень та стратегічної стабільності в умовах інтеграції ШІ дає можливість сформулювати кілька важливих висновків, що є основними для розуміння сучасного стану ядерної безпеки та перспектив її розвитку. По-перше, зміни, які викликані впровадженням ШІ в ядерні системи, є не тільки технологічними, але й онтологічними: вони змінюють природу стратегічної взаємодії, перетворюючи її зі звичайної людської соціальної практики на алгоритмічну технологічну взаємодію. Це означає, що класичні елементи теорії стримування - раціональність, репутація, комунікація намірів - потребують не лише адаптації, а й повного переосмислення. По-друге, вплив ШІ має нелінійний характер: він не поступово зменшує чи збільшує стабільність, а якісно змінює суть системи, переводячи її в інший режим роботи. Кожен із трьох виявлених етапів трансформації суб'єктності рішення є якісним стрибком, а не плавним переходом, що означає неможливість прогнозування наслідків автоматизації. По-третє, найбільшою загрозою для стратегічної стабільності є не повна автоматизація як така, а неконтрольоване поступове розширення алгоритмічної автономії без чіткого усвідомлення та управління цим процесом. Саме другий сценарій - напівавтономний оператор з ілюзорним людським контролем - є найбільш реалістичним та потенційно найнебезпечнішим, оскільки приховує справжній ступінь алгоритмічної автономії за формальною присутністю людини у контурі рішення. Ці висновки формують безпосередній базис для аналізу потенційної моделі «Dead Hand 2.0» у підрозділі 3.3, де розглядатимуться конкретні дилеми,

ризики ескалації та межі контролю, що виникають у контексті реалізації алгоритмічного стримування.

3.3. Потенційна модель «Dead Hand 2.0»: дилеми, ризики ескалації та межі контролю

Аналіз переваг і ризиків системи «Dead Hand 2.0», здійснений у підрозділі 3.1, а також дослідження змін логіки прийняття рішень і стратегічної стабільності, представлене у підрозділі 3.2, формують теоретичну базу для завдання цього розділу: конструювання потенційної моделі «Dead Hand 2.0» як аналітичного конструкту, що дозволяє систематизувати ключові дилеми алгоритмічного стримування, ідентифікувати конкретні механізми ризиків ескалації та визначити межі контролю, за якими автоматизоване стримування може зберегти стабілізуючий ефект, не генеруючи неприпустимих ризиків неконтрольованої ядерної ескалації. Термін «потенційна модель» у назві цього підрозділу є принциповим методологічним уточненням: йдеться не про опис будь-якої конкретної технічної системи, що вже існує або розробляється якоюсь країною, а про аналітичну реконструкцію найбільш вірогідної конфігурації автоматизованого ядерного стримування, побудовану на основі синтезу висновків попередніх частин, тенденцій розвитку військових технологій та логіки стратегічних взаємодій в умовах алгоритмізації. Такий підхід відповідає методологічній традиції створення гіпотетичних кейсів у дослідженнях безпеки, детально обґрунтованій у підрозділі 1.3, і дає змогу зберегти аналітичну строгість, не претендуючи на емпіричну точність щодо засекречених програм конкретних країн.

Потенційна модель «Dead Hand 2.0», що відповідає другому сценарію автоматизації - напівавтономного оператора, ідентифікованого у підрозділах 2.4 і 3.2 як найбільш реалістичному - може бути структурована як трирівнева система, де кожен рівень має свою роль та створює характерний набір ризиків. Перший рівень - сенсорно-аналітичний - є найбільш розвиненим та інституціоналізованим. На цьому рівні алгоритми машинного навчання інтегрують дані з багатьох джерел:

супутникового спостереження за розгортанням ядерних сил противника, моніторингу підводних платформ, перехоплення радіосигналів, аналізу кіберактивності, відстеження аномалій у дипломатичній комунікації та рухів стратегічних лідерів. Ключовою новизною порівняно із системами Холодної війни є не окреме стеження цих індикаторів, а їхнє об'єднання: алгоритм виявляє зв'язки між, на перший погляд, несумісними сигналами, формуючи оцінку ймовірності ядерного удару, що виражається в єдиному числовому значенні або груповій оцінці рівня загрози (Горовіц, 2018; Джонсон, 2020). Ризики першого рівня головним чином пов'язані із «проблемою чорної скриньки» та «overfitting»: алгоритм може виявляти статистичні кореляції, які не мають причинно-наслідкового зв'язку з реальною загрозою, або реагувати на новий тип сигналу непередбачуваним чином. Другий рівень - командно-рекомендаційний - є безпосереднім інтерфейсом між алгоритмічним аналізом та людським оператором. На цьому рівні система не лише інформує про оцінений рівень загрози, але й дає конкретні поради щодо дій, які ранжуються за очікуваним ефектом і мають оціночні ймовірності різних ситуацій. У випадках, що близькі до встановлених меж, система може автоматично почати підготовчі процедури: підвищити бойову готовність, включити запасні канали зв'язку, перевести командні центри в захищений режим - зберігаючи за людиною можливість скасування. Саме тут проявляється зміст «automation bias», описаний Саганом (1993): оператор, який отримує рекомендацію від системи з високою точністю, схильний сприймати її некретично, особливо в умовах часового тиску та інформаційного навантаження. Третій рівень - автономно-виконавчий - є найбільш стратегічно делікатним та прямо відповідає логіці «Dead Hand 2.0». Якщо система фіксує налаштування сигналів, що перевищують заданий поріг загрози, і одночасно помічає втрату зв'язку з вищим командуванням, вона переходить до автономного режиму: самостійно приймає рішення про відплату та ініціює відповідні протоколи без подальшої людської авторизації. Цей рівень є прямим функціональним наступником радянського «Периметру», але з зовсім іншою технологією: якщо «Периметр» реагував на детерміновані фізичні тригери, то «Dead Hand 2.0» оцінює

комплекс сигналів через алгоритм, поведінка якого менше передбачуваною навіть для розробників (Гоффман, 2009; Блер, 1993; Рассел і Норвіг, 2020).

Аналіз потенційної моделі «Dead Hand 2.0» виявляє чотири основні дилеми, що є типовими для алгоритмічного стримування і не можуть бути вирішені через технічне покращення системи - лише через принципові зміни в логіці її проєктування.

Перша дилема - достовірності та деескалації - є головним парадоксом, виявленим у підрозділі 3.1. Будь-яка конфігурація системи «Dead Hand 2.0» стикається з протиріччям: чим вищою є жостовірність загрози відплати (що є стабілізуючим фактором у теорії ігор), тим нижчою є здатність до деескалації в ситуаціях, коли відплата може бути помилковою або коли обидві сторони усвідомлюють потребу компромісу. Якщо система налаштована так, що противник може бути абсолютно впевненим у відплаті, це виключає будь-яку можливість «лазівки» для деескалації. Якщо система зберігає «деескалаційні» вікна, то вона неминуче знижує достовірність загрози - і стратегічний ефект стримування відповідно стає слабшим (Шеллінг, 1960; Феарон, 1997; Пауелл, 1990).

Жодне технічне рішення не може позбутися цієї дилеми: вона є структурною, виходячи з самої суті автоматизованого незворотного зобов'язання. Проте, як показує аналіз у підрозділі 1.2, класична «Гра в курча» Шеллінга дає часткове концептуальне вирішення через так звану «стратегічну ірраціональність» - створення ситуації, де виконання загрози є не визначеним, а ймовірним, що зберігає простір для деескалації без повної руйнації достовірності. У контексті «Dead Hand 2.0» аналогом може бути «вікно верифікації» - час між виявленням системою ознак загрози та активацією відплати протягом якого людський оператор має реальну можливість скасування, а не тільки номінальну. Проте чим більшим є це вікно, тим більш ризик, що супротивник знає про нього і розраховує на шанси скасування - а отже нижча стримуюча ефективність системи.

Друга дилема - проблема прозорості та безпеки - стосується оптимального рівня публічного розкриття деталей системи. Як виявлено у підрозділі 3.1, зрозумілість параметрів активації є необхідною умовою для ефективного

стримування: противник повинен знати, які дії неодмінно активують відплату, щоб уникати їх. Водночас розкриття точних параметрів системи дає можливість для адверсаріальних атак: противник, який знає умови активації, може або спеціально уникати відповідних сигналів, нівелюючи стримування, або навпаки – навмисно генерувати сигнали, що активують систему, провокуючи ненавмисний конфлікт (Гудфелоу, Шленс і Сегеді, 2015). Ця проблема не має оптимального вирішення в класичному сенсі: кожна точка на спектрі між повною доступністю та повною секретністю має власний набір ризиків. Повна відкритість підвищує стабілізуючий ефект стримання, але робить систему вразливою для маніпуляцій. Повна таємність захищає систему від маніпуляцій, але руйнує саму суть стримування: противник, не знаючи умов активації системи, не може скоригувати свою поведінку відповідно до них і може випадково перетнути невидимий поріг. У порівнянні з класичний стримуванням, де прозорість «червоних ліній» була ключовим механізмом стабілізації (Шеллінг, 1960; Загаре і Кілгур, 2000), «Dead Hand 2.0» через «проблему чорної скрині» структурно не здатна досягти потрібний рівень прозорості навіть за наявності такого наміру.

Третя дилема - проблема адаптивності та стабільності - особлива для систем на основі машинного навчання. З одного боку, здатність алгоритму до зміни і навчання є важливою перевагою «Dead Hand 2.0» над «Периметром»: система здатна виявляти нові потерни загроз та коригувати свої налаштування у відповідь на змінне стратегічне середовище. З іншого боку, така адаптація може бути потенційним джерелом нестабільності: система, що вчиться, може поступово зрушувати свої параметри реагування в непередбачуваному напрямку, особливо якщо набір даних містить систематичні упередження або якщо стратегічне середовище сильно відрізняється від того, у якому система навчалась (Джордан і Мітчелл, 2015; Гудфелоу, Бенджіо і Курвілль, 2016). У термінах теорії гри це означає, що рівновага стримування, стабільна на момент запуску системи, може стати нестабільною в міру збору нового досвіду системою - причому ця дестабілізація може бути непомітною до моменту кризи. На відміну від класичного стримування, де «правила гри» були більш-менш сталими і змінювалися лише

через явні переговори, алгоритмічне стримування може «дрейфувати» у бік нових варіантів рівноваги без будь-якого явного рішення з боку лідерів (Загаре, 1992; Шарр, 2018). Це підриває один із базових принципів стабільного стримування: передбачуваність параметрів стратегічної взаємодії.

Четверта дилема - проблема контролю та ефективності - виникає через «парадокс делегованого контролю», ідентифікованого у підрозділі 2.1 у контексті «Периметру». Будь-яка спроба посилити людський контроль над системою «Dead Hand 2.0» шкодитиме її діяльності як стримуючому механізму: кожен новий рівень людської авторизації є потенційним «вікном деескалації», що підриває абсолютну достовірність загрози. З іншого боку, зменшення людського контролю задля збільшення достовірності загрози є недопустимим з огляду вимоги «значимого людського контролю» (meaningful human control) (Сантоні де Сіо і ван ден Ховен, 2018; Шарр і Аллен, 2021). Ця дилема є особливо гострою у контексті другого сценарію автоматизації - напівавтономного оператора, - адже тут формальна наявність людського контролю приховує його реальну відсутність. Як зазначає Саган (1993), феномен «automation bias» перетворює «human-in-the-loop» на «human-as-a-rubber-stamp»: людина присутня у контурі рішення, але фактично не контролює, а лише легітимізує алгоритмічне рішення. Таким чином, четверта проблема є одночасно технічною, організаційною та нормативною: вона не може бути вирішена через поліпшення системи без радикального перегляду основних цілей її проєктування.

Виходячи з аналізу архітектури потенцмйної моделі та чотирьох основних дилем, можна ідентифікувати кілька конкретних механізмів, через які «Dead Hand 2.0» може генерувати ескалаційні ризики. Ці механізми є принципово відмінними від механізмів ескалації в класичному стримуванні, що показує необхідність їхнього систематичного аналізу для оцінки стратегічної стабільності.

Перший механізм - ненавмисна ескалація через помилку алгоритму - є прямим продовженням ризиків, виявлених у підрозділах 2.4 і 3.1. Система машинного навчання може класифікувати певну групу сигналів як ознаку неминучого удару, навіть якщо жоден компетентний людський аналітик не зробив

би такого висновку. Важливо те, що ця помилка може бути систематичною, а не випадковою: якщо корпус навчання системи має певне упередження або якщо система недостатньо роз'єднана з реальним стратегічним контекстом, вона може стабільно давати хибні позитивні результати в певному класі ситуацій. Прикладом є інцидент Петрова з 1983 року, де радянська система раннього попередження часто хибно трактувала відбиття сонячного світла від хмар як ракетні двигуни через особливості оптичних датчиків (Гоффман, 2009). Для «Dead Hand 2.0» подібна помилка може виникати через «overfitting» на певних геополітичних сценаріях, що призведе до хибної класифікації нового типу кризи як такого, що вже знайомий системі і потребує автоматичної реакції. На відміну від випадку Петрова, де людина могла опиратися висновоку алгоритму, в умовах «automation bias» другого рівня оператор, ймовірно, підтвердить рекомендацію без ретельного аналізу.

Другий механізм - каскадна ескалація через взаємодію автоматизованих систем - є якісно новим ризиком, що не має прямих аналогів в часи Холодної війни. Коли обидві сторони конфлікту мають системи типу «Dead Hand 2.0» , з'являється динаміка взаємних автоматичних реакцій, кожна з яких «раціональна» з погляду її алгоритмічної системи, але разом вони утворюють ескалаційну спіраль без людської участі. Як детально проаналізовано у розділі 2.4, Горовіц (2018) описує цю динаміку як «алгоритмічну нестабільність».

Конкретний сценарій може виглядати наступним чином: система однієї сторони виявляє підвищення активності ядерних сил суперника (яке є наслідком навчань, а не підготовки до удару) та ініціює підготовчі кроки. Ці заходи помічаються системою другої сторони як знак готовності до першого удару і викликають схожу реакцію; посилена активність другої сторони посилює занепокоєння системи першої, що починає наступний рівень протоколів. Кожен крок цього каскаду є раціональним з точки зору алгоритмічної логіки кожної окремої системи, але загальний результат - ескалаційна спіраль, яку жоден людський лідер не починав і, можливо, навіть не усвідомлює до досягнення критичного порогу. У термінах теорії нормальних аварій Перроу (1984), що застосовується Саганом (1993) до ядерних систем, це класичний приклад «щільно

пов’язаної» системи, де нормальне відхилення одного компонента каскадно поширюється через систему.

Третій механізм - провокована ескалація через адверсаріальні атаки - є новим стратегічним ризиком, що виникає тільки в умовах алгоритмічного стримування. Якщо система «Dead Hand 2.0» є вразливою до кібератак на сенсорну або командну інфраструктуру, третя сторона - держава, терористична організація чи навіть приватний актор - може спробувати маніпулювати вхідними даними системи так, щоб спровокувати автоматичну ядерну відплату без реального удару (Гудфелоу, Шленс і Сегеді, 2015; Джонсон, 2020). Це відкриває абсолютно новий вид стратегічного шантажу: замість створення власної ядерної сили актор може прагнути контролювати алгоритмічний тригер чужої системи. Особливо небезпечним є те, що такий сценарій значно важче пояснити, ніж традиційний ядерний удар. Якщо «Dead Hand 2.0» спрацьовує на підроблені дані, цільова країна може не мати жодної змоги встановити, чи був це справжній напад, або провокація третьої сторони - принаймні до кінця ядерного обміну. Це є прямим порушенням одного з основних принципів ефективного стримування – здатності точно визначити джерело загрози без якого відплата втрачає свою стратегічну обґрунтованість (Загаре і Кілгур, 2000; Пауелл, 1990).

Четвертий механізм - ескалація через асиметрію автоматизації – випливає безпосередньо з аналізу ризиків в підрозділі 3.1. Якщо одна сторона використовує «Dead Hand 2.0», а інша зберігає контроль людини, то виникає структурна асиметрія, що може провокувати превентивну ескалацію. Сторона без автоматизованої системи має стратегічний стимул нейтралізувати «Dead Hand 2.0» супротивника до її повної активації, бо після активації вона втрачає будь-яку можливість уникнути відплати навіть у разі помилкового спрацювання системи. Таким чином, парадоксально, розгортання «Dead Hand 2.0» однією стороною може збільшити, а не зменшити привабливість першого удару для супротивника – протилежно декларованій меті системи (Шеллінг, 1960; Фірон, 1997).

Концепція «меж контролю» в рамках «Dead Hand 2.0» стосується не тільки технічних деталей системи, але й ширших умов, за яких автоматизоване стримування може функціонувати без неприйнятного підвищення ризиків ескалації. Аналіз попередніх підрозділів дозволяє ідентифікувати декілька таких меж, за якими ризики стають структурно некерованими.

Перша межа - поріг алгоритмічної непрозорості. Поки параметри активації системи хоча б частково перевіряються та зрозумілі для супротивника, стримування зберігає стабілізуючий ефект. Проте коли «проблема чорної скрині» стає настільки серйозною, що навіть розробники не можуть передбачити поведінку системи в нестандартних ситуаціях, управління ескалаційними ризиками стає принципово неможливим: немає способу надати гарантію, що система не спрацює помилково в умовах, які не передбачені навчальним корпусом (Барредо Арріета та ін., 2020). Це межа є особливо критичною, оскільки складність алгоритмів машинного навчання зростає нелінійно зі збільшенням їхньої здатності: найефективніші системи є одночасно найменш прозорими.

Іншою межею є поріг часового звуження. Поки між виявленням системою ознак загрози та активацією відплати залишається часовий проміжок, достатній для справжньої (а не номінальної) людської перевірки, існує можливість запобігання помилковій ескалації. Але коли цей проміжок зменшується до масштабу, що фізично виключає адекватне сприйняття ситуації людиною, контроль ризиками стає неможливим навіть за наявності формального механізму контролю. Шарр і Аллен (2021) визначають цей поріг у контексті гіперзвукових загроз як приблизно 3-5 хвилин - проміжок, протягом якого людина може прийняти «рішення», але не може провести повний аналіз ситуації.

Третьою межею є поріг кіберзахищеності. Поки системи захисту «Dead Hand 2.0» від зовнішніх втручань залишаються достатніми для зупинки майже усіх атак, система залишається досить контрольованою. Але в умовах розвитку кіберзброї та методів нападу на системи машинного навчання жодна система не може гарантувати абсолютний захист. Коли ймовірність успішної маніпуляції сенсорними даними досягає стратегічно значущого рівня, «Dead Hand 2.0» перестає

бути інструментом стримування і стає потенційним каталізатором провокованої ескалації (Гудфелоу, Шленс та Сегеді, 2015; Джонсон, 2020).

Четвертою межею є поріг стратегічної симетрії. Поки рівні автоматизації двох сторін є порівнянними, механізм взаємного стримування зберігає хоч якусь логіку: обидві системи працюють у подібних параметрах і генерують схожі сигнали. Але коли одна сторона значно випереджає іншу в рівні автоматизації, виникає стратегічна асиметрія, описана у підрозділах 3.1 та 3.2, що може провокувати превентивні дії менш автоматизованої сторони до того, як система противника стане повністю готовою. Ця межа є особливо складною для управління та контролю, так як вона залежить від технічного розвитку обох сторін, що виходить за рамки двохсторонніх переговорів.

Формулювання дослідницьких питань у вступі роботи передбачало дві гіпотези: перша - про те, як штучний інтелект трансформує логіку ядерного стримування; друга - чи здатне автоматизоване стримування забезпечити стратегічну стабільність без підвищення ризику неконтрольованої ескалації. Аналіз, здійснений у трьох підрозділах третього розділу, дозволяє сформулювати відповідні висновки. Щодо першого питання: інтеграція ШІ змінює логіку ядерного стримування не лише в технічному, але й в онтологічному сенсі. Вона змінює суб’єкта рішення - від людини до алгоритму; часову структуру взаємодії - від послідовної гри до майже одночасної взаємодії; природу інформаційного середовища - від дипломатично-соціальної комунікації до алгоритмічного сигналювання; та механізми підтримки рівноваги - від людської рівноваги Неша до алгоритмічної рівноваги з принципово вищою структурною невизначеністю. Ця трансформація є нелінійною та якісною: вона не просто ускладнює класичне стримування, а переводить його в абсолютно новий режим функціонування. Щодо другого питання: відповідь є умовно негативною в радикальній версії та дещо позитивною в обмеженій версії системи. Система «Dead Hand 2.0» з повною автономністю є структурно дестабілізуючою: вона підвищує достовірність загрози, але й одночасно збільшує ризик неконтрольованої ескалації через всі чотири механізми, описані в цьому розділі. У конфігурації ШІ як аналітичного

інструмента для людини (перший сценарій) система може бути стабілізуючою за умови збереження реального людського контролю і недопущення «automation bias». Найбільш реалістична конфігурація - напівавтономний оператор (другий сценарій) - є структурно амбівалентною: вона поєднує стабілізуючий ефект підвищеної достовірності загрози з дестабілізуючим ефектом прихованої алгоритмічної автономії і небезпеки каскадної ескалації.

Особливої практичної значущості аналіз ризиків систем типу «Dead Hand 2.0» набуває в контексті російсько-української війни, що почалася у 2014 році і переросла у повномасштабне вторгнення у 2022 році. Росія часто використовує ядерну риторику як спосіб стратегічного тиску, апелюючи до невразливості власного потенціалу відплати - риторики, що концептуально базується на автоматизованому стримуванні, яке почалося з «Периметру» (Гоффман, 2011; Пейн, 2018). Цей контекст ілюструє два механізми для ескалаційних ризиків. По-перше, ризик ескалації через асиметрію автоматизації є структурно доречним: Україна як неядерна держава знаходиться в ситуації, яка функціонально відповідає асиметричній моделі Крайга (1999), де тільки один гравець має реальну та діючу ядерну загрозу, що унеможливлює класичне взаємне стримування. По-друге, ядерна риторика Росії функціонує як форма стратегічної непрозорості: невизначеність умов використання ядерної зброї та згадки про автоматизовані механізми відплати створюють інформаційне середовище, яке ускладнює адекватну оцінку «червоних ліній» партнерами України і міжнародною спільнотою. Це підкреслює практичну значущість розуміння логіки та меж контролю автоматизованого стримування не лише для ядерних держав, але й для країн, що формують свої безпекові позиції в умовах ядерного тиску з боку держав-агресорів.

Таким чином, автоматизоване ядерне стримування типу «Dead Hand 2.0» здатне забезпечити стратегічну стабільність лише за виконання чотирьох умов: достатньої прозорості параметрів активації для стримуючого впливу; реального, а не лише номінального людського контролю над важливими рішеннями; достатнього кіберзахисту, щоб уникнути маніпуляцій; та стратегічної симетрії в

рівнях автоматизації між ядерними країнами. Немає жодних сучасних тенденцій у розвитку військових технологій та міжнародної безпеки, які б вказували, що всі ці чотири умови можуть бути одночасно виконані. Це означає, що без принципово нових механізмів міжнародного регулювання, верифікації та комунікації між ядерними державами розгортання систем типу «Dead Hand 2.0» є стратегічним ризиком навіть за наявності технічних можливостей та стратегічних мотивів для їхнього створення.

ВИСНОВКИ

Це дослідження було спрямоване на вивчення того, яким чином розвиток штучного інтелекту та автоматизованих систем управління трансформує логіку ядерного стримування та впливає на стратегічну стабільність, а також на оцінку того, чи здатне автоматизоване ядерне стримування типу «Dead Hand 2.0» забезпечити стратегічну стабільність без підвищення ризику неконтрольованої ескалації. Для досягнення цієї мети було виконано шість взаємопов'язаних завдань.

1. Виявлено, у процесі виокремлення ключових концептів теорії ігор, що мають значення для ядерного стримування, що рівновага Неша, раціональність, достовірність загрози, зобов'язання та репутація зберігають аналітичну цінність, але потребують принципового переосмислення в умовах автоматизації. Встановлено, що алгоритмічна «раціональність» системи «Dead Hand 2.0» є обмеженою не через когнітивні обмеження людини, а через програмні обмеження алгоритму, що оптимізує задану цільову функцію без здатності до контекстуального політичного судження. Доведено, що автоматизовані системи є найбільш радикальною формою «зв'язування рук» у термінах Феарона: вони роблять загрозу абсолютно достовірною через технологічну незворотність, але ціною повної втрати можливості деескалації. Виявлено, що автоматизація трансформує рівновагу Неша від людської, заснованої на взаємному раціональному розрахунку, до алгоритмічної рівноваги з принципово вищою структурною невизначеністю, де незначні зміни вхідних даних можуть каскадно порушувати стратегічний баланс без участі будь-якого людського актора.
2. Обґрунтовано, що базові та двоступеневі моделі Загаре та Крайга зберігають концептуальну цінність як аналітичний каркас, але їхня застосовність до алгоритмічних систем є структурно обмеженою через три ключові припущення - досконала та повна інформація, людська раціональність і послідовна часова структура взаємодії - кожне з яких

автоматизація ставить під загрозу. Обґрунтовано, що ключовий висновок двоступеневої моделі - ядерна ескалація за досконалої інформації не відбувається - може бути переформульований для умов «Dead Hand 2.0»: що нижчою є якість інформації, доступної алгоритмічній системі, то вищим є ризик виходу за межі стабільної рівноваги у напрямі неконтрольованої ескалації. Доведено, що алгоритмічна асиметрія є структурно аналогічним чинником нестабільності до асиметрії ядерних потенціалів у класичних моделях: якщо лише одна сторона делегує рішення алгоритму, дві сторони фактично грають за різними правилами, що порушує умови стабільної рівноваги, виявлені Крайгом у 14 можливих рівновагах двоступеневої моделі.

3. З'ясовано, у процесі дослідження історичних передумов та досвіду функціонування системи «Периметр», що радянська система є першим практичним прототипом автоматизованого ядерного стримування, де частина функцій стратегічного ухвалення рішень була делегована технічній системі у відповідь на американську доктрину «обезголовлюючого удару» та розгортання ракет Pershing II з часом польоту до Москви 8-10 хвилин. Встановлено, що «Периметр» демонструє «парадокс делегованого контролю»: формально зберігаючи людський елемент через чергового офіцера в захищеному командному пункті, система функціонально наближалася до повної автономії через екстремальні умови прийняття рішення - ізоляцію, відсутність зв'язку з вищим керівництвом та тиск часу. Виявлено два принципово різні типи ризику автоматизованого стримування, задокументовані через інциденти 1983 року: технічний збій сенсорних систем, що призводить до хибного тригеру (інцидент Петрова), та системна помилка в інтерпретації стратегічного контексту (навчання «Able Archer 83»). Обґрунтовано, що перехід від rule-based automation «Периметру» до adaptive autonomy «Dead Hand 2.0» є якісним стрибком, що принципово підвищує як потенційну ефективність, так і структурні ризики: сучасні алгоритми машинного

навчання здатні виявляти складніші патерни загроз, але їхня поведінка є менш передбачуваною навіть для розробників.

4. Виявлено, у процесі пошуку потенційних переваг та ризиків автоматизованого ядерного стримування, асиметричну структуру їхнього співвідношення: переваги системи є переважно теоретичними та умовними, тоді як ризики є структурними та постійними. Серед потенційних переваг ідентифіковано чотири: радикальне посилення достовірності загрози другого удару через технологічну незворотність; усунення «проблеми часового вікна» в умовах гіперзвукових загроз; зниження ризику певних форм людської помилки в критичних ситуаціях; та компенсація стратегічної асиметрії для менших ядерних держав. Встановлено, що ризики утворюють три категорії: технологічні (алгоритмічна помилка через overfitting та вразливість до адверсаріальних атак), стратегічні (усунення деескалаційного потенціалу, ризик асиметрії автоматизації та непрозорість «червоних ліній» для противника) та системні (гонка алгоритмів як нова форма дилеми безпеки та структурна крихкість складних взаємопов'язаних систем).

   Доведено фундаментальний парадокс алгоритмічного стримування: прагнення до абсолютної достовірності загрози через технологічну незворотність породжує абсолютну нездатність до деескалації. Виявлено принципову асиметрію між типами помилки: хибно позитивний результат - система спрацьовує у відповідь на хибний сигнал - є катастрофою якісно іншого роду порівняно з хибно негативним, що виходить за межі будь-якої стратегічної логіки та не передбачається жодною класичною моделлю стримування.

5. З’ясовано, що трансформація логіки стримування під впливом ШІ є сутнісною, а не лише технічною: вона змінює суб'єкта рішення від людини до алгоритму, трансформує часову структуру взаємодії від послідовної екстенсивної гри до майже одночасної нормальної форми, якісно змінює інформаційне середовище та породжує алгоритмічну рівновагу з

принципово вищою структурною невизначеністю. Стратегічна стабільність в умовах алгоритмічного стримування визначається чотирма новими умовами, специфічними для алгоритмічного середовища: алгоритмічною симетрією між рівнями автоматизації сторін, алгоритмічною сумісністю між логіками різних систем, кіберзахищеністю від зовнішніх маніпуляцій та підтриманням реального, а не номінального людського контролю. Виявлено, що найбільш реалістичний сценарій - напівавтономний оператор - є структурно найбільш небезпечним, оскільки поєднує ілюзію людського контролю з реальною алгоритмічною автономією через феномен «automation bias», де «human-in-the-loop» фактично перетворюється на «human-as-a-rubber-stamp». Встановлено, що відсутність міжнародно-правових механізмів регулювання автономності ядерних систем є одночасно правовою прогалиною та симптомом недостатньої соціальної конструкції проблеми алгоритмічного стримування як спільної загрози в конструктивістському розумінні Вендта: поки держави сприймають розробку ШІ-систем для ядерного стримування як елемент конкурентної стратегії, формування ефективних міжнародних норм залишатиметься надзвичайно складним завданням.

6. У процесі нової інтерпретації потенційного розвитку «Dead Hand 2.0» у контексті міжнародної безпеки обґрунтовано, що система є трансформаційним, а не однозначно стабілізуючим або дестабілізуючим чинником: вирішуючи проблему достовірності загрози другого удару, вона переводить джерела нестабільності з одного виміру в інший - породжуючи нові ризики алгоритмічної помилки, непрозорості та каскадної ескалації. Встановлено, що потенційна модель «Dead Hand 2.0» як трирівнева система - сенсорно-аналітичний, командно-рекомендаційний та автономно-виконавчий рівні - генерує чотири фундаментальні дилеми: достовірності та деескалації, транспарентності та безпеки, адаптивності та стабільності, контролю та ефективності, - жодна з яких не має оптимального вирішення в межах технічного вдосконалення системи без

принципового перегляду логіки її проектування. Виявлено чотири конкретних механізми ескалаційних ризиків: ненавмисна ескалація через алгоритмічну помилку; каскадна ескалація через взаємодію автоматизованих систем двох сторін; провокована ескалація через адверсаріальні кібератаки третіх сторін; та ескалація через асиметрію автоматизації. Доведено, що умовно позитивна відповідь на центральне дослідницьке питання можлива лише за одночасного виконання чотирьох умов: достатньої транспарентності параметрів активації, реального людського контролю над критичними рішеннями, достатнього кіберзахисту та стратегічної симетрії в рівнях автоматизації між ядерними державами. Жодна з сучасних тенденцій розвитку військових технологій та міжнародного безпекового середовища не вказує на те, що всі чотири умови можуть бути виконані одночасно без принципово нових механізмів міжнародного регулювання, верифікації та стратегічної комунікації між ядерними державами.

Таким чином, результати цього дослідження, загалом розуміння логіки та ризиків систем типу «Dead Hand 2.0» є не лише академічними, а й практично значущими для формування обґрунтованих позицій у сфері міжнародної безпеки та стратегічних комунікацій із партнерами в умовах стрімкого технологічного розвитку.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Ahmed, A. (2017). The philosophy of nuclear proliferation/non-proliferation: Why states build or forgo nuclear weapons? Trames, 21(4), 371–382. https://www.kirj.ee/public/trames_pdf/2017/issue_4/Trames-2017-4-371-382.pdf
2. Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion, 58,* 82–115. https://doi.org/10.1016/j.inffus.2019.12.012
3. Benoît, J.-P., & Krishna, V. (1985). Finitely repeated games. *Econometrica, 53*(4), 905–922. https://doi.org/10.2307/1912660
4. Snyder, Frank; Blair, Bruce G.; and Ford, Daniel (1986) ""Strategic Command and Control: Redefining the Nuclear Threat," "The Button: The Pentagon's Strategic Command and Control System-Does It Work?"," Naval War College Review: Vol. 39: No. 2, Article 15. https://digital-commons.usnwc.edu/nwc-review/vol39/iss2/15
5. Blair, B. G. (1993). *The logic of accidental nuclear war.* Brookings Institution Press. https://www.brookings.edu/book/the-logic-of-accidental-nuclear-war/
6. Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies.* Oxford University Press. https://global.oup.com/academic/product/superintelligence-9780198739838
7. Boulding, K. E. (1963). *Conflict and defense: A general theory.* Harper & Row. https://dokumen.pub/conflict-and-defense-a-general-theory-0819171123.html
8. Bozhenko, A. Y. (2024). *U.S. nuclear deterrence strategy through the lens of game theory* (Bachelor's thesis). National University of Kyiv-Mohyla Academy.
9. Brams, S. J., & Kilgour, D. M. (1988). *Game theory and national security.* Basil Blackwell. https://www.researchgate.net/publication/231901727_Game_Theory_and_Nation

al_Security_Steven_J_Brams_and_D_Marc_Kilgour_Oxford_Basil_Blackwell_1988_pp_ix_199

10. Camerer, C. F. (2003). *Behavioral game theory: Experiments in strategic interaction.* Princeton University Press. https://www.researchgate.net/publication/209410260_Behavioral_Game_Theory_Experiment_in_Strategic_Interaction
11. Craig, C. (1999). *Destroying the village: Eisenhower and thermonuclear war.* Columbia University Press. https://cup.columbia.edu/book/destroying-the-village/9780231111232
12. Dal Bó, P., & Fréchette, G. R. (2018). On the determinants of cooperation in infinitely repeated games: A survey. *Journal of Economic Literature, 56*(1), 60–114. https://www.aeaweb.org/articles?id=10.1257/jel.20160980
13. Danilovic, V. (2001). The sources of threat credibility in extended deterrence. *Journal of Conflict Resolution, 45*(3), 341–369. https://doi.org/10.1177/0022002701045003005
14. Fearon, J. D. (1994). Domestic political audiences and the escalation of international disputes. *American Political Science Review, 88*(3), 577–592. https://doi.org/10.2307/2944796
15. Fearon, J. D. (1997). Signaling foreign policy interests: Tying hands versus sinking costs. *Journal of Conflict Resolution, 41*(1), 68–90. https://doi.org/10.1177/0022002797041001004
16. Freedman, L. (2003). *The evolution of nuclear strategy* (3rd ed.). Palgrave Macmillan. https://www.academia.edu/12038451/EVOLUTION_OF_NUCLEAR_STRATEGY_2003
17. Freedman, L., & Michaels, J. (2019). *The evolution of nuclear strategy* (4th ed.). Palgrave Macmillan. https://doi.org/10.1057/978-1-137-57350-6
18. Gates, R. M. (1996). *From the shadows: The ultimate insider's story of five presidents and how they won the Cold War.* Simon & Schuster. https://openlibrary.org/books/OL8457508M/From_the_Shadows

19. George, A. L., & Bennett, A. (2005). *Case studies and theory development in the social sciences.* MIT Press. https://www.academia.edu/19264308/Case_Studies_and_Theory_Development_in_the_Social_Sciences
20. Gerring, J. (2007). *Case study research: Principles and practices.* Cambridge University Press. https://ocw.upj.ac.id/files/Slide-PSG301-Case-study-Research.pdf
21. Gibbons, R. (1997). An introduction to applicable game theory. *Journal of Economic Perspectives, 11*(1), 127–149. https://doi.org/10.1257/jep.11.1.127
22. Goodfellow, I., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. *arXiv*:1412.6572. https://doi.org/10.48550/arXiv.1412.6572
23. Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning.* MIT Press. https://www.deeplearningbook.org/
24. Gray, C. S. (1999). *Modern strategy.* Oxford University Press. https://www.researchgate.net/publication/307797510_MODERN_STRATEGY
25. Hoffman, D. E. (2009). *The dead hand: The untold story of the Cold War arms race and its dangerous legacy.* Doubleday. https://archive.org/details/isbn_9780385524377
26. Horowitz, M. C. (2018a). Artificial intelligence, international competition, and the balance of power. *Texas National Security Review, 1*(3). https://tnsr.org/2018/05/artificial-intelligence-international-competition-and-the-balance-of-power/
27. Intriligator, M. D., & Brito, D. L. (1984). Can arms races lead to the outbreak of war? *Journal of Conflict Resolution, 28*(1), 63–84. https://doi.org/10.1177/0022002784028001004
28. Jervis, R. (1989). *The meaning of the nuclear revolution: Statecraft and the prospect of Armageddon.* Cornell University Press. https://www.cornellpress.cornell.edu/book/9780801495656/the-meaning-of-the-nuclear-revolution/

29. Johnson, J. (2020). *Artificial intelligence and the future of warfare: The USA, China, and strategic stability.* Manchester University Press. https://www.researchgate.net/publication/352020307_Artificial_Intelligence_and_the_Future_of_Warfare_USA_China_Strategic_Stability
30. Johnston, A. I. (1995). *Cultural realism: Strategic culture and grand strategy in Chinese history.* Princeton University Press. https://www.jstor.org/stable/j.ctvzxx9p0
31. Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science, 349*(6245), 255–260. https://doi.org/10.1126/science.aaa8415
32. Kahn, H. (1965). *On escalation: Metaphors and scenarios.* Praeger. https://api.pageplace.de/preview/DT0400.9781351502214_A30457138/preview-9781351502214_A30457138.pdf
33. Koehler, R. C. (2016, September 29). Nuclear weapons and the crisis of civilization. *Common Dreams.* https://www.commondreams.org/views/2016/09/29/nuclear-weapons-and-crisis-civilization
34. Lebow, R. N., & Stein, J. G. (1994). *We all lost the Cold War.* Princeton University Press. https://dokumen.pub/we-all-lost-the-cold-war-course-booknbsped-9781400821082.html
35. Luce, R. D., & Raiffa, H. (1957). *Games and decisions: Introduction and critical survey.* Wiley. https://ia802909.us.archive.org/6/items/in.ernet.dli.2015.132625/2015.132625.Games-And-Decisions-Introduction-And-Critical-Survey.pdf
36. Maynard Smith, J. (1982). *Evolution and the theory of games.* Cambridge University Press. https://doi.org/10.1017/CBO9780511806292
37. Mearsheimer, J. J. (1990). Back to the future: Instability in Europe after the Cold War. *International Security, 15*(1), 5–56. https://www.semanticscholar.org/paper/Back-to-the-Future%3A-Instability-in-Europe-After-the-Mearsheimer/b08c37852ec08baf7b2bad24a80ebd1dbbdfa72e

38. Mearsheimer, J. J. (2001). *The tragedy of great power politics.* W. W. Norton & Company. https://www.slideshare.net/slideshow/the-tragedy-of-great-power-politics-1pdf/266699506
39. Morgan, P. M. (2003). *Deterrence now.* Cambridge University Press. https://doi.org/10.1017/CBO9780511491955
40. Morgenthau, H. J. (1978). *Politics among nations: The struggle for power and peace* (5th ed.). Alfred A. Knopf. http://slantchev.ucsd.edu/courses/ps240/04%20Conflict%20with%20States%20as%20Unitary%20Actors/Morgenthau%20-%20Politics%20among%20nations%20(selected%20chapters).pdf
41. Myerson, R. B. (2009). A history of Nash equilibrium and game theory. *International Journal of Game Theory, 28*(2), 105–109. https://www.aeaweb.org/articles?id=10.1257/jel.37.3.1067
42. Nash, J. F. (1950a). Equilibrium points in n-person games. *Proceedings of the National Academy of Sciences, 36*(1), 48–49. https://doi.org/10.1073/pnas.36.1.48
43. Nash, J. F. (1950b). Non-cooperative games. *Annals of Mathematics, 54*(2), 286–295. https://doi.org/10.2307/1969529
44. Nuclear weapons and realism. (2018). *E-International Relations.* https://www.e-ir.info/2018/02/05/nuclear-weapons-and-realism/
45. Osborne, M. J. (2004). *An introduction to game theory.* Oxford University Press. https://www.economics.utoronto.ca/osborne/igt/igtFrontmatterAndPreface.pdf
46. Palan, R. (2000). A world of their making: An evaluation of the constructivist critique in international relations. *Review of International Studies, 26*(4), 575–598. https://doi.org/10.1017/S0260210500005751
47. Payne, K. B. (2018). *Strategy: Retrospect and prospect.* Missouri State University. https://www.airuniversity.af.edu/Portals/10/SSQ/documents/Volume-14_Issue-2/Payne.pdf
48. Perrow, C. (1984). *Normal accidents: Living with high-risk technologies.* Basic Books.https://maritimesafetyinnovationlab.org/wp-

content/uploads/2021/04/Normal-Accidents-Living-With-High-Risk-Technologies-Perrow.pdf

49. Petersen, B. C. (2012). *The Nuclear Non-Proliferation Treaty: A comparison of realist, liberal and constructivist views* [Master's thesis, University of the Western Cape]. https://www.google.com/url?sa=t&source=web&rct=j&opi=89978449&url=https://uwcscholar.uwc.ac.za/bitstreams/ee3e9d48-9a0a-4467-b669-5d59e6a631d6/download&ved=2ahUKEwi9rKXux72UAxUu6wIHHbqLN4oQFnoECBwQAQ&usg=AOvVaw1Bi3bmJlugDn-eaF7iXf7V
50. Plous, S. (1993). *The psychology of judgment and decision making.* McGraw-Hill. https://www.researchgate.net/publication/304088277_Review_of_Scott_Plous_The_Psychology_of_Judgment_and_Decision_Makingem
51. Poundstone, W. (1993). *Prisoner's dilemma: John von Neumann, game theory, and the puzzle of the atom.* Anchor Books. https://books.google.com.ua/books/about/Prisoner_s_Dilemma.html?id=Zf804exGloQC&redir_esc=y
52. Powell, R. (1990). *Nuclear deterrence theory: The search for credibility.* Cambridge University Press. https://www.jstor.org/stable/4137605
53. Powell, R. (1999). *In the shadow of power: States and strategies in international politics.* Princeton University Press. http://slantchev.ucsd.edu/courses/pdf/Powell%20-%20States%20and%20Strategies.pdf
54. Rawls, J. (1999). *The law of peoples.* Harvard University Press. https://mercaba.org/SANLUIS/Filosofia/autores/Contemporánea/Rawls/The%20Law%20of%20Peoples.pdf
55. Rühle, M. (2015, April 28). Deterrence: What it can (and cannot) do. *NATO Review.*
56. Russell, S., & Norvig, P. (2020). *Artificial intelligence: A modern approach* (4th ed.). Pearson. https://aima.cs.berkeley.edu/
57. Sagan, S. D. (1993). *The limits of safety: Organizations, accidents, and nuclear weapons.* Princeton University Press.

https://api.pageplace.de/preview/DT0400.9780691213064_A39514919/preview-9780691213064_A39514919.pdf

58. Santoni de Sio, F., & van den Hoven, J. (2018). Meaningful human control over autonomous systems: A philosophical account. *Frontiers in Robotics and AI, 5,* 15. https://doi.org/10.3389/frobt.2018.00015
59. Scharre, P. (2018). *Army of none: Autonomous weapons and the future of war.* W. W. Norton & Company. https://ftp.idu.ac.id/wp-content/uploads/ebook/tdg/MILITARY%20PLATFORM%20DESIGN/Army%20of%20None%20Autonomous%20Weapons%20and%20the%20Future%20of%20War.pdf
60. Scharre, P., & Allen, G. C. (2021). *Autonomous weapons and strategic stability.* Center for a New American Security. https://kclpure.kcl.ac.uk/ws/portalfiles/portal/129451536/2020_Scharre_Paul_1575997_ethesis.pdf
61. Schelling, T. C. (1960). *The strategy of conflict.* Harvard University Press. https://www.sackett.net/Strategy-of-Conflict.pdf
62. Schelling, T. C. (1966). *Arms and influence.* Yale University Press. https://yalebooks.yale.edu/book/9780300246742/arms-and-influence/
63. Simon, H. A. (1957). *Models of man: Social and rational.* Wiley. https://www.scirp.org/reference/referencespapers?referenceid=2260840
64. Snyder, G. H., & Diesing, P. (1977). *Conflict among nations: Bargaining, decision making, and system structure in international crises.* Princeton University Press. https://www.scirp.org/reference/referencespapers?referenceid=3318445
65. Stake, R. E. (1995). *The art of case study research.* SAGE Publications. https://books.google.com.ua/books/about/The_Art_of_Case_Study_Research.html?id=ApGdBx76b9kC&redir_esc=y
66. Tarr, D. W. (1991). *Nuclear deterrence and international security: Alternative nuclear regimes.* Longman.

67. U.S. Department of Defense. (2012). *Directive 3000.09: Autonomy in weapon systems.* https://www.esd.whs.mil/Portals/54/Documents/DD/issuances/dodd/300009p.pdf
68. Varian, H. R. (2020). *Intermediate microeconomics: A modern approach* (9th ed.). W. W. Norton & Company. https://www.scirp.org/reference/referencespapers?referenceid=2806808
69. Verba, S. (1961). *Small groups and political behavior: A study of leadership.* Princeton University Press. https://press.princeton.edu/books/paperback/9780691619996/small-groups-and-political-behavior?srsltid=AfmBOoo1SA4yKlEitddumvQ8M-00m3MVjmcAhuHKvf8LeUinvdlDYScw
70. Von Neumann, J., & Morgenstern, O. (1944). *Theory of games and economic behavior.* Princeton University Press. http://pombo.free.fr/neumannmorgenstern.pdf
71. Waltz, K. N. (1979). *Theory of international politics.* McGraw-Hill. https://dl1.cuni.cz/pluginfile.php/486328/mod_resource/content/0/Kenneth%20N.%20Waltz%20Theory%20of%20International%20Politics%20Addison-Wesley%20series%20in%20political%20science%20%20%20%201979.pdf
72. Waltz, K. N. (1981). *The spread of nuclear weapons: More may be better* (Adelphi Paper 171). International Institute for Strategic Studies. https://www.tandfonline.com/doi/abs/10.1080/05679328108457394
73. Wendt, A. (1992). Anarchy is what states make of it: The social construction of power politics. *International Organization, 46*(2), 391–425. https://www.jstor.org/stable/2706858
74. Wendt, A. (1999). *Social theory of international politics.* Cambridge University Press. https://doi.org/10.1017/CBO9780511612183
75. Yin, R. K. (2014). *Case study research: Design and methods* (5th ed.). SAGE Publications. https://www.academia.edu/87937093/Robert_K_Yin_2014_Case_Study_Research_Design_and_Methods_5th_ed_Thousand_Oaks_CA_Sage_282_pages

76. Zagare, F. C. (1987). *The dynamics of deterrence.* University of Chicago Press. https://www.google.com/url?sa=t&source=web&rct=j&opi=89978449&url=https://www.jstor.org/stable/26275168&ved=2ahUKEwjpz4Svzb2UAxVA6gIHHUxBG0UQFnoECBYQAQ&usg=AOvVaw330ONnI26eR6STZgygV_BU

77. Zagare, F. C. (1990). Rationality and deterrence. *World Politics, 42*(2), 238–260. https://www.acsu.buffalo.edu/~fczagare/Articles/Rationality%20and%20Deterrence.pdf

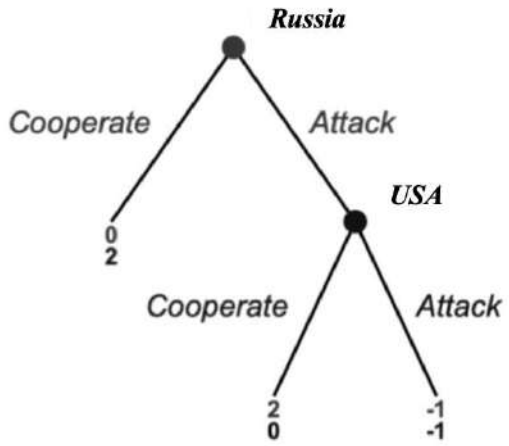
78. Zagare, F. C., & Kilgour, D. M. (2000). *Perfect deterrence.* Cambridge University Press. https://assets.cambridge.org/97805217/81749/sample/9780521781749ws.pdf

ДОДАТКИ

1. Рисунок 1.1 Дилема в'язня гонки озброєнь між СРСР та США (Плус, 1993). Рівновага знаходиться в стані (Озброюватись, озброюватись), хоча (Роззброюватись, роззброюватись) є ефективним результатом за Парето.

		USSR: Disarm	USSR: Arm
USA	Disarm	3, 3	1, 4
USA	Arm	4, 1	2, 2

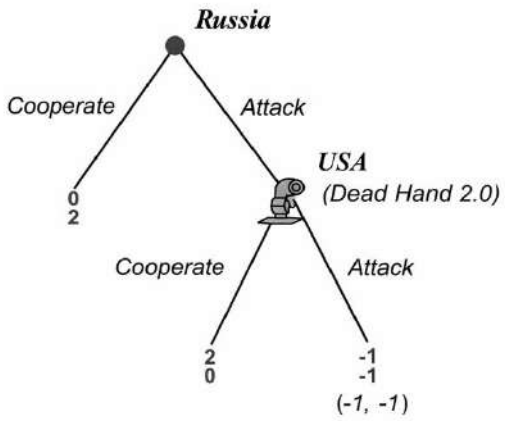
2. Рисунок 1.2 Гра екстенсивної форми з двома гравцями (Гіббонс, 1997). Рівновага Неша при заданих виграшах має вигляд (Атака, Співпраця).[1]



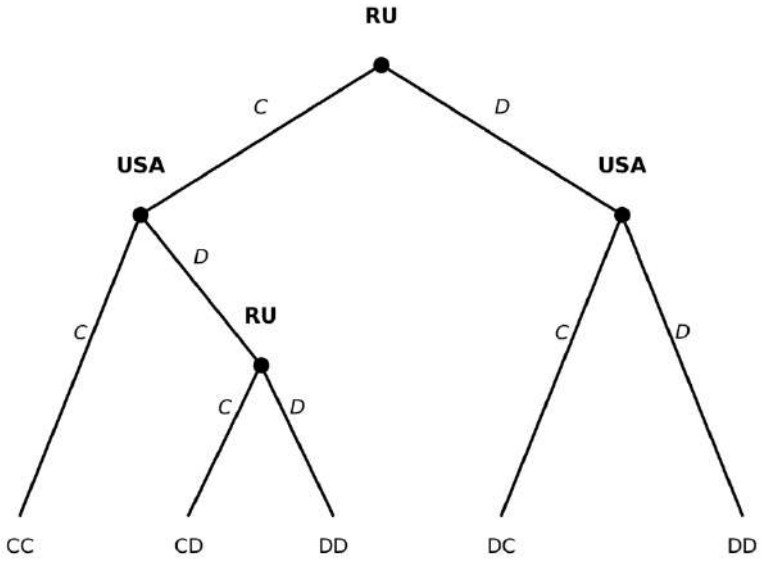
3. Рисунок 1.3 Гра екстенсивної форми з автоматизованою відплатою («Dead Hand 2.0») у моделі ядерного стримування між Росією та США. Рівновага Неша за умов автоматизованої відповіді має вигляд (Співпраця, Автоматизована відплата), за якої результат (2, 0) стає недосяжним, а загроза

[1] Ігри, використані в цій роботі, а саме ігри Гіббонса (1997), Загаре (1992) і Крайга (1999), були дещо змінені для цієї роботи. Вони візуально не ідентичні до оригінальних документів.

ядерної відплати набуває повної достовірності (адаптовано за: Гіббонс, 1997).[2]

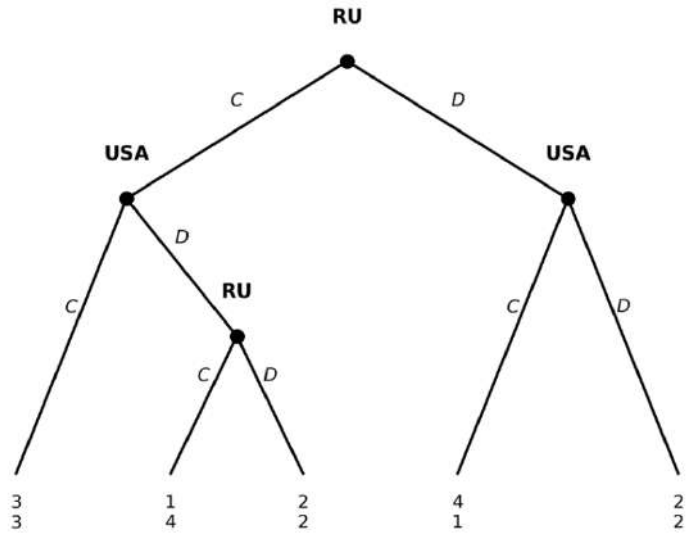


4. Рисунок 2.1 Базова гра стримування в екстенсивній формі з послідовними рішеннями гравців (адаптовано за Загаре, 1992; Крайг, 1999). Модель відображає стратегічну взаємодію між Росією (RU) та США (USA), де гравці послідовно обирають співпрацю (C) або дезертирство/ескалацію (D). Структура гри демонструє логіку прийняття рішень у класичній моделі ядерного стримування за умов повної та досконалої інформації.[3]



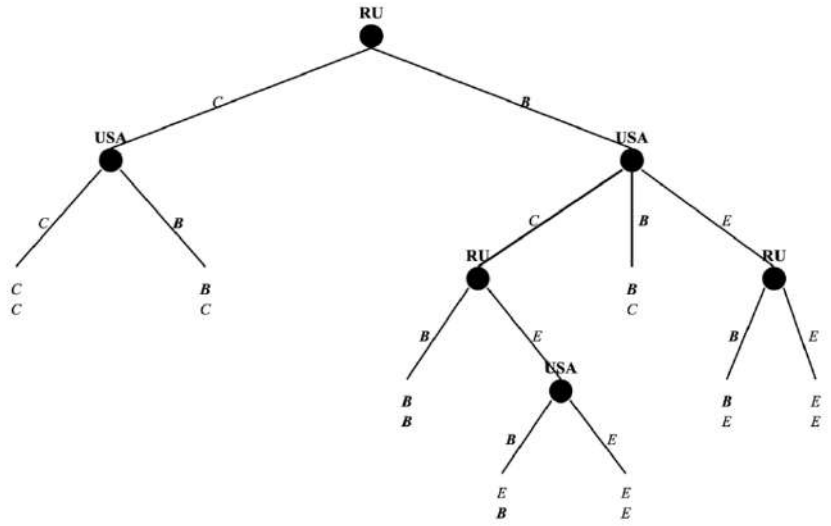
[2] Рисунок був створений за допомогою ChatGPT 5.2 за запитом автора. Усі візуальні метафори відповідають авторському задуму та були відібрані з декількох згенерованих варіантів.
[3] Рисунок був створений за допомогою ChatGPT 5.2 за запитом автора. Усі візуальні метафори відповідають авторському задуму та були відібрані з декількох згенерованих варіантів.

5. Рисунок 2.2 Базова гра стримування з виплатами на основі переваг, пояснених Загаре (1992) та Крайг (1999). Рівновага Неша знаходиться в точці (C, C). Верхні числа результатів вказують на виграш Росії, а числа нижче - на виграш США. Модель демонструє умови досягнення стабільного взаємного стримування за наявності достовірних загроз обох сторін.[4]

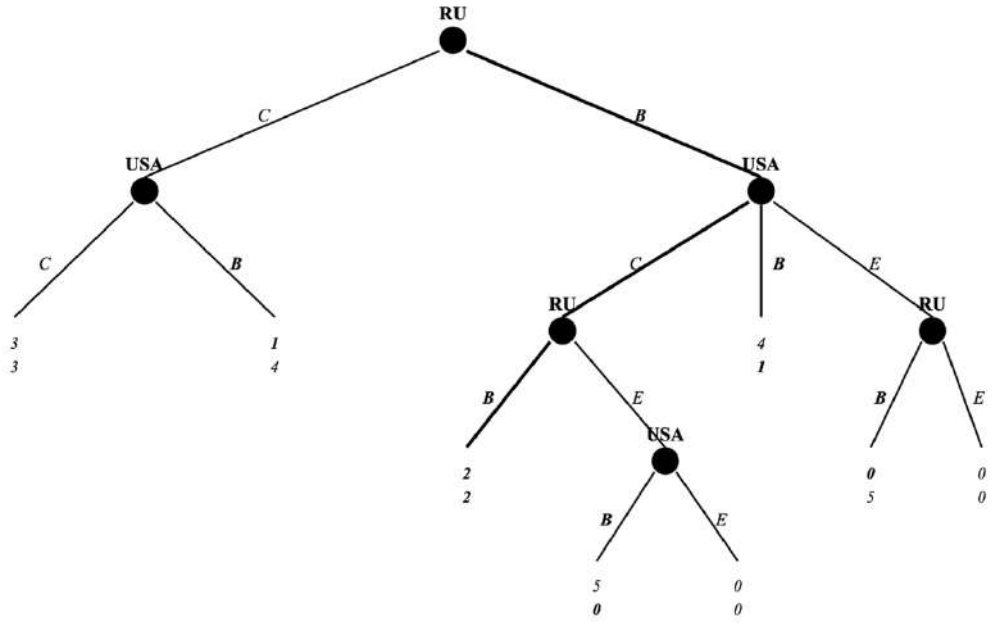


6. Рисунок 2.3 Двоетапна гра стримування з ядерною ескалацією в екстенсивній формі (адаптовано за Загаре, 1992; Крайг, 1999), де C - співпраця (cooperation); B - дезертирство (defection), що інтерпретується як перехід до звичайних (конвенційних) бойових дій; E - ядерна ескалація (nuclear escalation). Жирна лінія позначає рівноважний шлях гри, отриманий методом зворотної індукції.[5]

[4] Рисунок був створений за допомогою ChatGPT 5.2 за запитом автора. Усі візуальні метафори відповідають авторському задуму та були відібрані з декількох згенерованих варіантів.
[5] Рисунок був створений за допомогою ChatGPT 5.2 за запитом автора. Усі візуальні метафори відповідають авторському задуму та були відібрані з декількох згенерованих варіантів.



7. Рисунок 2.4 Двоетапна гра стримування з ядерною ескалацією в екстенсивній формі з виплатами (адаптовано за Загаре, 1992; Крайг, 1999), де верхнє число - виграш Росії (RU), нижнє - виграш США (USA). CC = (3,3) - мир; BB = (2,2) - звичайна війна; BC = (1,4) / CB = (4,1) - одностороннє дезертирство; EB = (5,0) - одностороння ядерна ескалація Росії; BE = (0,5) - одностороння ядерна ескалація США; EE = (0,0) - взаємне ядерне знищення. Жирна лінія позначає рівноважний шлях гри, визначений методом зворотної індукції; за умов повної інформації та заданої структури переваг рівноважним результатом є CC (мир, взаємне стримування).[6]



[6] Рисунок був створений за допомогою ChatGPT 5.2 за запитом автора. Усі візуальні метафори відповідають авторському задуму та були відібрані з декількох згенерованих варіантів.

АНОТАЦІЯ

Кваліфікаційної роботи

Тема. «Dead hand 2.0 – автоматизоване ядерне стримування в епоху штучного інтелекту»

Здобувачка Боженко Анна Євгеніївна

Рік навчання 4-ий рік навчання

Факультет соціальних наук та соціальних технологій

Науковий керівник Тараненко Ганна Геннадіївна, канд.політ.наук, доцентка кафедри міжнародних відносин

Рецензент _____

Захищена “____”_____2026 р.

Короткий зміст роботи

Робота присвячена дослідженню трансформації логіки ядерного стримування під впливом розвитку штучного інтелекту та автоматизованих систем управління на прикладі концепту «Dead Hand 2.0». У роботі виявлено прогалину між класичними теоретичними моделями ядерного стримування, що ґрунтуються на припущенні про людський контроль над рішеннями, та сучасними технологічними реаліями алгоритмічного стримування. У теоретичному розділі розглянуто базові концепції безпеки та ядерного стримування крізь призму реалізму, лібералізму та конструктивізму, проаналізовано інструментарій теорії ігор - рівновагу Неша, раціональність, зобов'язання, достовірність загрози - та обґрунтовано методологію поєднання кейс-стаді з теоретико-ігровим моделюванням. У другому розділі досліджено радянську систему «Периметр» як перший практичний прототип автоматизованого стримування, проаналізовано базові та двоступеневі моделі стримування Загаре та Крайга, а також механізми впровадження штучного інтелекту в ядерні системи командування. У третьому розділі побудовано потенційну модель «Dead Hand 2.0» як трирівневу систему, виявлено чотири фундаментальні дилеми алгоритмічного стримування та чотири конкретні механізми ескалаційних ризиків, а також визначено межі контролю, за якими

автоматизоване стримування стає структурно некерованим. Доведено, що «Dead Hand 2.0» здатна забезпечити стратегічну стабільність лише за одночасного виконання чотирьох умов: транспарентності параметрів активації, реального людського контролю, достатнього кіберзахисту та стратегічної симетрії в рівнях автоматизації між ядерними державами.

Short summary:

The paper examines the transformation of nuclear deterrence logic under the influence of artificial intelligence and automated control systems, using the concept of «Dead Hand 2.0» as a case study. The paper identifies a gap between classical theoretical models of nuclear deterrence, which assume human control over decisions, and the contemporary technological realities of algorithmic deterrence. The theoretical section analyses the basic concepts of security and nuclear deterrence through the lenses of realism, liberalism, and constructivism, examines game theory tools - Nash equilibrium, rationality, commitment, and credibility of threats - and justifies a methodology combining case study with game-theoretic modelling. The second section investigates the Soviet Perimeter system as the first practical prototype of automated deterrence, analyses Zagare's and Craig's basic and two-stage deterrence models, and explores the mechanisms of AI integration into nuclear command systems. The third section constructs a potential model of «Dead Hand 2.0» as a three-tier system, identifies four fundamental dilemmas of algorithmic deterrence and four specific escalation risk mechanisms, and defines the limits of control beyond which automated deterrence becomes structurally unmanageable. It is argued that «Dead Hand 2.0» can ensure strategic stability only if four conditions are met simultaneously: transparency of activation parameters, genuine human control, sufficient cyber protection, and strategic symmetry in the levels of automation between nuclear states.