https://doi.org/10.31891/2307-5732-2026-365-101
УДК 004.8:622.24

КАСЯНЧУК ІГОР
Івано-Франківський національний технічний університет нафти і газу
https://orcid.org/0009-0008-1980-4339
e-mail: igorkasyanchuk@gmail.com
КАСЯНЧУК ВІТАЛІЙ
Івано-Франківський національний технічний університет нафти і газу
https://orcid.org/0009-0005-9104-933X
e-mail: kasvit93@gmail.com

НАВЧАННЯ З ПІДКРІПЛЕННЯМ У ПРОЦЕСАХ ІНТЕЛЕКТУАЛЬНОГО УПРАВЛІННЯ РЕЖИМАМИ БУРІННЯ СВЕРДЛОВИН

У роботі розглянуто проблему інтелектуального управління режимами буріння нафтогазових свердловин в умовах високої динамічності процесів, невизначеності геологічних умов та жорстких вимог до безпеки та економічної ефективності. Продемонстровано обмеження традиційних підходів, що ґрунтуються на досвіді інженерів і традиційних контурних системах автоматичного регулювання. Запропоновано використання навчання з підкріпленням для оптимізації осьового навантаження на долото, частоти обертання, параметрів промивання та траєкторії на основі схеми «стан – дія – винагорода». Узагальнено сучасні практичні рішення, охарактеризовано роль цифрових двійників і функції винагороди. Обґрунтовано доцільність впровадження таких підходів для підвищення ефективності та безпеки буріння.

Ключові слова: *навчання з підкріпленням, інтелектуальне управління режимами буріння, нафтогазові свердловини, агент «стан – дія – винагорода», архітектура «актор-критик», глибоке навчання з підкріпленням, цифрові двійники бурових комплексів, функція винагороди, коефіцієнт дисконтування, оптимізація параметрів буріння.*

KASYANCHUK IGOR, KASYANCHUK VITALII
Ivano Frankivsk National Technical University of Oil and Gas

REINFORCEMENT LEARNING APPLIED TO THE PROCESSES OF INTELLIGENT WELL DRILLING MODE CONTROL

This paper considers the problem of intelligent control of oil and gas well drilling operations under conditions of highly dynamic processes, uncertain geological conditions, and strict constraints regarding safety and economic efficiency. It is noted that traditional approaches, based on engineers' experience and classical control loop controllers, do not fully account for the multidimensional states of the environment and the nonlinear relationships between operating parameters.

The purpose of the scientific research is to analyze the capabilities of reinforcement learning in the context of intelligent drilling mode control, to investigate the suitability of various classes of reinforcement learning algorithms, and to develop a conceptual architecture for an agent capable of optimizing, within a "state–action–reward" framework, the axial load on the drilling bit, rotation speed, flushing parameters, and trajectory, taking into account the well response.

This research summarizes current practical applications of reinforcement learning to wellbore cleaning, directional drilling with bottomhole pressure control within specified limits, geonavigation, trajectory control, and power plant control; it analyzes actor-critic architectures, DQN-like schemes, and offline reinforcement learning with conservative Q-learning, as well as the role of digital twins as a training environment. Special attention is given to the formation and calibration of the reward function, the influence of the discount factor on the balance between immediate benefit and long-term reliability, as well as the optimization of hyperparameters that determine the stability and convergence of agents.

The results of the conceptual analysis suggest that it is feasible to use reinforcement learning as the basis for developing intelligent drilling mode control systems aimed at increasing mechanical drilling speed, reducing drill bit wear, lowering accident rates, and decreasing the proportion of manual labor. Further research prospects include the development and implementation of reinforcement learning agent prototypes in digital twins of drilling systems, the integration of multi-agent and physics-informed solutions, and the experimental verification of their effectiveness using real industrial data.

Keywords: *reinforcement learning, intelligent drilling mode control, oil and gas wells, the «state-action-reward» agent, «actor-critic» architecture, deep reinforcement learning, digital twins of drilling rigs, reward function, discount factor, drilling parameter optimization.*

Стаття надійшла до редакції / Received 17.03.2026
Прийнята до друку / Accepted 24.04.2026
Опубліковано / Published 28.05.2026

Постановка проблеми у загальному вигляді та її зв'язок із важливими науковими чи практичними завданнями

Сучасні процеси спорудження нафтогазових свердловин характеризуються високою динамічністю, невизначеністю геологічних умов та жорсткими вимогами до безпеки та економічної ефективності. Традиційні методи керування режимами буріння здебільшого ґрунтуються на практичному досвіді інженерів та класичних контурних регуляторах, що не дозволяє в повній мірі врахувати багатовимірність станів середовища, нелінійні залежності між параметрами режиму та їх довгострокові наслідки для ресурсу інструменту, якості стовбура та вартості проходки. В умовах змінних обмежень, складної взаємодії «долото – порода – бурильна колона – буровий розчин» та великої кількості можливих комбінацій навантаження на долото, частоти обертання та витрати промивальної рідини постає задача переходу від реактивного та емпіричного керування до інтелектуальних систем, що здатні самостійно створювати оптимальну стратегію дій. Навчання з підкріпленням, що розглядає керування бурінням як послідовний процес прийняття рішень у термінах «стан – дія – винагорода»,

без необхідності точної фізичної моделі середовища, створює передумови для адаптивної оптимізації режимів у реальному часі. Тим самим проблема розробки та впровадження підходів навчання з підкріпленням безпосередньо пов'язана з ключовими практичними завданнями галузі, а саме з підвищенням механічної швидкості проходки, зменшенням зносу доліт, зниженням аварійності, скороченням участі людини в рутинних операціях та забезпеченням стійкого, керованого буріння в ускладнених умовах.

Аналіз останніх джерел та публікацій

В оглянутих наукових дослідженнях [1-8] продемонстровано, що глибоке навчання з підкріпленням та споріднені підходи дають змогу враховувати складну структуру системи «родовище – свердловини» й оптимізувати політику розробки в умовах нелінійної та нестаціонарної динаміки. Дані роботи демонструють застосування методів машинного навчання та навчання з підкріпленням для оптимізації траєкторій та режимів похило-скерованого і спрямованого буріння. Дослідження за темою розвивають архітектури «актор – критик», гібридні AI-physics моделі та консервативне офлайн-Q-навчання, що є основою для побудови стійких агентів керування. Навчання з підкріпленням використовується для оптимізації горизонтальних траєкторій та політики розробки родовищ з акцентом на послідовному прийнятті рішень. Роботи присвячені глибинному навчанню з підкріпленням для геостерингу, будівництва свердловин і керованого буріння з підтриманням вибійного тиску, що безпосередньо споріднено з тематикою дослідження. Розглянуто інтегровані функції винагороди, хмарні системи геонавігації, багатоагентні схеми та задачі діагностики (зокрема втрати циркуляції). Дослідження фокусуються на керуванні режимами буріння (динамічні тести, оптимізація режимів за невизначеної геології, очищення стовбура, налаштування контурів тиску) на основі цифрових двійників і гібридних схем «регулятор + навчання з підкріпленням».

Формулювання цілей статті

Метою наукового дослідження є аналіз можливостей навчання з підкріпленням у задачах інтелектуального керування режимами буріння свердловин, зокрема оцінка придатності основних класів алгоритмів (методи Монте-Карло, часових розрізнень, динамічне програмування, глибокі Q-мережі, підходи типу «актор – критик» та багатоагентні рішення) в умовах високої невизначеності та нестаціонарності середовища. Передбачається розроблення концептуальної архітектури агента, здатного у форматі послідовного прийняття рішень оптимізувати осьове навантаження на долото, частоту обертання та параметри промивання з урахуванням реакції свердловини, а також формування та калібрування функції винагороди за показниками швидкості проходки, зносу долота, вібрацій, ризику аварій та технологічних обмежень. Досягнення цієї мети ґрунтується на узагальненні практичних кейсів, формалізації простору станів, дій та винагород та обґрунтуванні доцільності впровадження глибокого навчання з підкріпленням як інструмента підвищення ефективності й безпеки процесів спорудження свердловин.

Виклад основного матеріалу

Навчання з підкріпленням – це тип машинного навчання, при якому агент вчиться приймати рішення (оптимальну стратегію – набір правил дій), взаємодіючи зі складним навколишнім середовищем. Адаптивність та ефективність навчання з підкріпленням (RL), навіть у сценаріях із змінними операційними обмеженнями, роблять його надзвичайно цінним для динамічного планування.

Навчання з підкріпленням складається з п'яти основних частин: агент, середовище, дія, стан і винагорода. На відміну від традиційного підходу до оптимізації, навчання з підкріпленням використовує проміжні стани, згенеровані симулятором, і забезпечує оптимізовану стратегію, застосовну до цілої низки різних сценаріїв. Агент виконує дії, отримує винагороди або штрафи залежно від своїх дій і коригує свою стратегію (оновлює свій стан) для максимізації сукупної винагороди з плином часу) (рисунок 1).

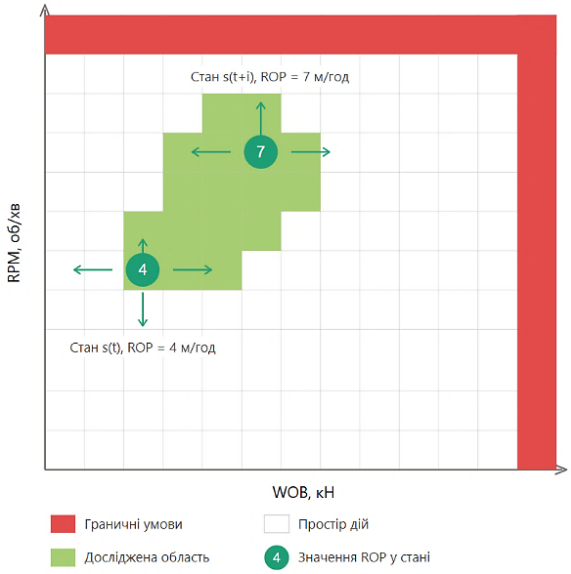

Рис. 1. Схема оптимізації і управління параметрами буріння як процесу навчання з підкріпленням

Сукупність усіх допустимих дій у середовищі відома як простір дій. Саме тому на відміну від методу планування, заснованого на моделюванні, навчання з підкріпленням не вимагає повної геологічної моделі. Системі штучного інтелекту не обов'язково знати точні диференціальні рівняння руйнування гірської породи під різцями PDC-долота. Агент аналізує безпосередньо реакцію свердловини (зміни крутного моменту, MSE, тиск на стояку та ін.) і визначає оптимальну дію, спираючись на стратегію, вироблену під час мільйонів ітерацій тренування. Він вчиться на наслідках, а не на першопричинах.

На відміну від статичних алгоритмів, навчання з підкріпленням базується на постійному циклі зворотного зв'язку, а саме на гнучкості **(**агент постійно коригує свою поведінку залежно від стану середовища (s_t), орієнтації на результат (замість чітких інструкцій «як робити», агент максимізує сумарну винагороду (R)) та на стійкості до змін (навіть якщо умови (обмеження) змінюються в процесі, модель здатна знайти новий оптимальний шлях через ітеративне навчання).

Алгоритми навчання з підкріпленням можуть базуватися на функціях цінності або на стратегіях, а також на гібридних підходах (актор-критик):

1. Алгоритм заснований на цінності, починається з довільного початкового значення. Спочатку агент навчання з підкріпленням обирає спонтанну дію для конкретного стану та обчислює функцію цінності. Агент може знайти найкращу стратегію, використовуючи функцію цінності. Він не зберігає саму стратегію, а лише функцію цінності. Навчання з підкріпленням на основі цінності потребує багато часу, оскільки Q-таблиця містить велику кількість станів і дій (у бурінні це означало б, що ми можемо змінювати навантаження на долото лише фіксованими кроками (наприклад: 50 кН, 60 кН, 70 кН). Спроба створити щільну таблицю для всіх комбінацій WOB, RPM та Q призводить до «прокляття розмірності», через що модель вчиться занадто повільно для реального часу). До поширених алгоритмів цього типу належить алгоритм SARSA (State-Action-Reward-State-Action) ідентичний Q-навчанню, за винятком того, що він не шукає максимальне значення Q і є внутрішньостратегічним. Використання поточного набору дій, що виконуються в поточному стані, покращує процес навчання агента, а попередні стани та винагороди не враховуються для нового набору станів.

2. Алгоритм на основі стратегій починається з вибору агентом випадкової стратегії (він обирає дію для конкретного стану відповідно до цієї стратегії). Потім він обчислює функцію цінності для випадково обраної стратегії. Алгоритм, заснований на стратегіях, має вищу швидкість збіжності. Навчання з підкріпленням на основі стратегій підходить для застосунків із великим простором станів. За допомогою цього підходу можна вивчати стохастичні стратегії. Однак навчання з підкріпленням на основі стратегій страждає від високої дисперсії. Для збіжності потрібно мало ітерацій, але сам алгоритм є складним. Алгоритм «актор-критик» (Actor-Critic) є комбінацією підходів, заснованих на функціях цінності та на стратегіях».

3. Гібридна архітектура Актор-Критик об'єднує найкраще з двох попередніх варіантів, щоб подолати проблему високої дисперсії градієнта стратегії: Актор реалізує підхід на основі стратегій і відповідає безпосередньо за керування – аналізує поточний стан свердловини та генерує оптимальні параметри буріння. Критик реалізує підхід на основі цінності. Він оцінює дію Актора – обчислює функцію цінності, прогнозуючи, до яких довгострокових наслідків призведе цей режим (наприклад, чи не викличе це раптове зростання вартості метра C_{pm} або критичний стрибок MSE). Актор-Критик може балансувати між дослідженням та використанням застосовуючи політику та функцію цінності: політика використовується для дослідження середовища та пошуку нових рішень, а функція цінності – для використання наявних знань про середовище та прийняття рішень на основі очікуваної віддачі від кожної дії. Завдяки Критику Актор вчиться набагато стабільніше та швидше.

Одним із варіантів реалізації гібридної архітектури є м'який актор-критик (SAC – Soft Actor-Critic) – це ще один тип алгоритму навчання з підкріпленням (RL), який інтегрує архітектуру "актор-критик" із регуляризацією ентропії. Мережа актора (actor-network) обирає дії залежно від поточного стану, тоді як мережа критика (critic network) використовується для оцінки цінності цих дій. Елемент регуляризації ентропії вводиться в цільову функцію для стимулювання дослідження середовища (exploration) та запобігання передчасній збіжності.

Автономне навчання з підкріпленням (Offline RL) пропонує практичний підхід до навчання агентів без необхідності взаємодії з симулятором у режимі реального часу. Однак на практиці офлайн-навчання з підкріпленням зазвичай страждає від переоцінки Q-значень при роботі зі складними навчальними даними. Тому автори запропонували використовувати консервативне Q-навчання (Conservative Q-Learning – CQL) – один із варіантів алгоритмів офлайн-навчання з підкріпленням, спрямований на консервативне прогнозування власної продуктивності після виконання рекомендованої дії. CQL може ефективно покращити нижню межу вивченої Q-функції. CQL використовує помилку Беллмана і навчається Q-значенню за допомогою глибокої структури Q-навчання. Він розв'язує проблему переоцінки в офлайн-навчанні з підкріпленням, вводячи консервативний фактор регуляризації в цільову функцію Q-навчання. Це допомагає гарантувати, що вивчена Q-функція є нижньою межею функції істинного значення. CQL побудований на архітектурі «актор-критик». В даному кейсі актор – це нейронна мережа, яка передбачає дії, і ці дії оцінюються критиком.

Для навчання критика використовується функція втрат, що включає помилку Беллмана, модифікатор Soft Actor-Critic (SAC) та модифікатор CQL. Модифікатор SAC дозволяє мережі політики видавати дії з розподілом ймовірностей і спрямовує критика на прийняття дій з високою ентропією. Модифікатор CQL штрафує значення Q, що відповідають політиці, і підвищує значення Q, отримані з навчального набору даних.

Перевага підходу актор-критик полягає у тому що, він може обробляти нестаціонарні середовища, де розподіл даних змінюється з часом. Це відбувається тому, що політика оновлюється на основі власного досвіду,

а функція цінності критика оновлюється з використанням помилки часової різниці. Тому CQL – ефективний підхід для навчання на великих і статичних наборах даних у якому дані про систему класифікуються за станом (інформацією, що відома про систему на момент виконання дії), дією та винагородою.

До інших підходів до навчання з підкріпленням також відносять (рисунок) метод Монте-Карло (MC), метод часових розрізнень (TD) і динамічне програмування (DP) та ін. Методи Монте-Карло та часових розрізнень вимагають наявності досвіду і можуть навчатися шляхом дослідження (методом спроб і помилок) без апріорної моделі середовища. Якщо підсумувати досвід авторів, то можна зробити такі висновки:

1. Динамічне програмування (DP) вимагає, щоб модель навчалася та слідувала рекурсивному підходу. Модель повинна визначати стани, дії, винагороди та ймовірність переходу. Динамічне програмування використовується для розв'язання складних задач. Воно вирішує задачу шляхом її розбиття на підзадачі (етапи) з наступним пошуком рішення для кожної з них та їх об'єднанням для формування загального розв'язку. Воно може розв'язувати взаємопов'язані підзадачі, задачі з оптимальною підструктурою та марковські процеси прийняття рішень (MDP).
2. Метод Монте-Карло (MC) слідує підходу навчання на основі завершених епізодів і використовується в умовах невизначеності.
3. Метод часових розрізнень (TD) визначає оцінки цінності на основі оцінок інших значень – цей процес називається бутстрапуванням (bootstrapping). TD підходить для безперервних задач (що не завершуються) і не вимагає знання моделі. Він оновлює оцінки цінності після кожного кроку. Йому не потрібно чекати завершення всього експерименту (епізоду) – він оновлює значення тільки для фактично пройденого шляху. Метод TD використовується в онлайн- та інкрементному навчанні, має низьку дисперсію та деяку зміщеність (похибку).

Окрім вибору самого підходу до навчання, на поведінку RL-агента суттєво впливають його ключові параметри. Винагорода також тісно пов'язана із коефіцієнтом дисконтування γ (коефіцієнтом передбачення), що по суті, характеризує, наскільки агент навчання з підкріпленням зацікавлений у винагородах у далекому майбутньому порівняно з винагородами в найближчому майбутньому. При $\gamma < 0.85$ агент RL стає «жадібним» і за через те, що бачить лише поточний метр проходки він екстремально збільшує WOB та RPM, щоб отримати максимальну нагороду за швидкість просто зараз. Це призводить до інтенсивного зносу долота, вібрацій та погіршення «стану стовбура свердловини». При $\gamma > 0.85$ агент RL діє стратегічно та обирає більш консервативні та безпечні режими буріння. При $\gamma > 0.88$ поведінка системи більше не змінюється, тобто що алгоритм RL досягає свого глобального оптимуму і настає ефект плато.

Проаналізуємо алгоритми глибокого навчання з підкріпленням, що набули найбільшого поширення в задачах керування бурінням. Проксимальна оптимізація стратегії (PPO – Proximal Policy Optimization) – це алгоритм, що працює безпосередньо зі стратегією (policy-based підхід) і є одним із найефективніших методів глибокого навчання з підкріпленням (DRL) для керування складними динамічними об'єктами, якими є бурові установки. Він генерує дані та оновлює стратегію, використовуючи наявну (поточну) стратегію. PPO складається з двох частин: мережі стратегії (policy network), яка відображає стани на дії, та мережі цінності (value network), яка обчислює цінність стану або комбінації стану та дії. Мережа стратегії навчається максимізувати прогнозовану сукупну винагороду, тоді як мережа цінності навчається оцінювати цінність станів і дій. Цей алгоритм є досить простим у розгортанні та стійким до вибору гіперпараметрів, проте критичним для його успішного навчання є етап min-max нормалізації (масштабування) для усунення проблем домінування великих чисел. До того ж, за даними авторів PPO забезпечує найшвидшу збіжність до майже оптимального рішення порівняно з іншими алгоритмами на основі моделі «актор-критик».

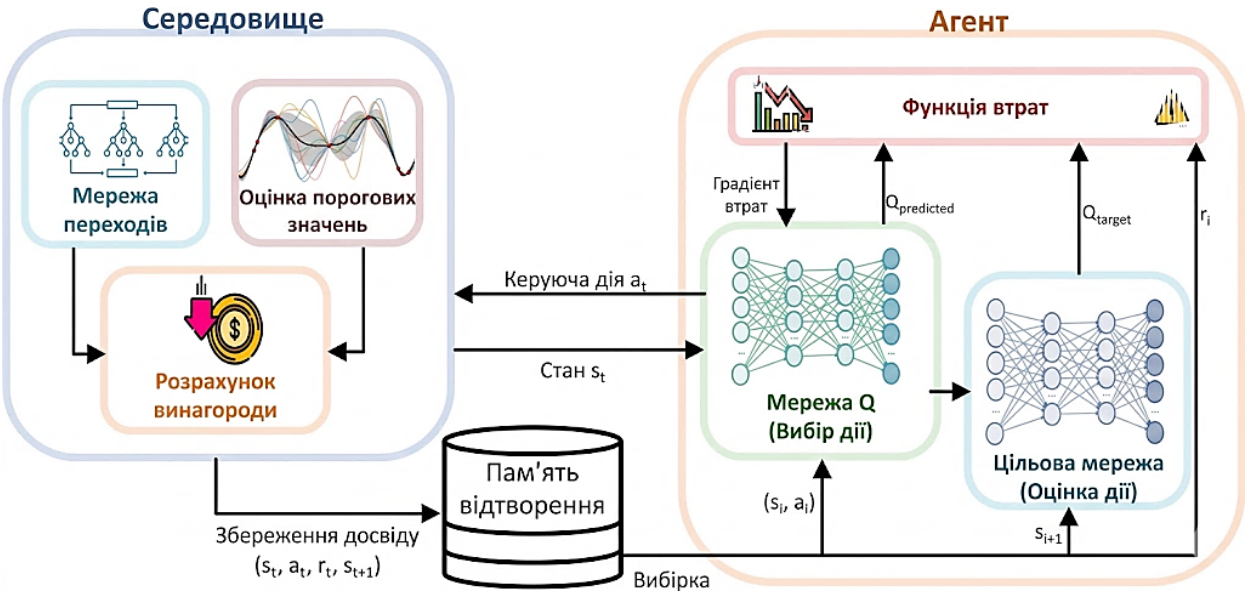

Рис. 2. Ілюстрація алгоритму навчання з підкріпленням

Deep Q-Network (DQN) поєднує Q-навчання з глибокими нейромережами, використовуючи механізм буфера відтворення (Experience Replay) для стабілізації навчання. DQN забезпечує стабільну, перевірювану відповідність між умовами експлуатації та дискретними керуючими діями і може бути легко інтегрований з існуючою інфраструктурою керування буровою установкою. Так, авторами пропонується структура навчання з підкріпленням (рисунок 2) у межах якої середовище моделюється для розрахунку функції винагороди шляхом інтеграції змінних стану та дії з моделлю переходів, побудованою з використанням алгоритму випадкового лісу (Random Forest).

Ця модель відображає ймовірнісну динаміку між операційними вхідними даними та умовами у свердловині, тоді як гаусівський процес встановлює порогові обмеження для кількісної оцінки операційних ризиків.

Середовище взаємодіє з агентом глибокої Q-мережі (DQN), який ітеративно оновлює Q-мережу для мінімізації втрат під час ухвалення рішень та оптимізації продуктивності стратегії. Для забезпечення стабільності під час навчання та запобігання зміщенню, викликаному послідовними кореляціями в даних буріння, агент використовує буфер відтворення досвіду, що дозволяє повторно використовувати обрані траєкторії для оновлення градієнта. Така конструкція забезпечує як надійність (робастність), так і адаптивність, дозволяючи агенту вивчати ефективні стратегії ухвалення рішень в умовах складної та невизначеної динаміки спорудження свердловини.

Суттєвим обмеженням алгоритму DQN є те, що процес навчання агента може стати нестабільним через різкі та значні оновлення вагових коефіцієнтів мережі. Описану проблему, за даними авторів можливо розв'язати шляхом інтеграції механізму «м'якого оновлення» (soft update) в архітектуру DQN – забезпеченні стабільнішого та ефективнішого процесу навчання за рахунок поступового змішування вагових коефіцієнтів цільової мережі (target network) з ваговими коефіцієнтами поточної основної мережі (online network) в режимі реального часу.

Якщо параметрами функції стратегії (політики) є ваги глибокої нейронної мережі, то така задача оптимізації називається глибоким навчанням з підкріпленням (Deep Reinforcement Learning, DRL). У рамках модифікованої структури DQN додається нейронна мережа «стара нейронна мережа» з такою ж структурою. Нейронна мережа використовує історичні параметри, період оновлення є тривалим, і на виході отримується старе оціночне значення $Q_0'(s2, a2)$ для реальності. Оцінка значення:

$$Q_r(s1, a1) = R(s1, a1) + \gamma \cdot Q_0'(s2, a2) \quad (1)$$

де (s, a) – значення керуючого впливу a за умов навколишнього середовища s; γ – значення ослаблення, яке вказує на те, що значення керуючої дії *a1* у стані *s1* ослаблюється з кореляцією між наступним станом та дією *s2*, *a2*.

Система керування на основі подвійної нейронної мережі використовує «історичний досвід» за допомогою «старої нейронної мережі» та застосовує «нову нейронну мережу» для оцінки значень та вибору стратегії керування, що порушує кореляцію між досвідом і підвищує ефективність оновлення параметрів системи.

Залежно від характеру керованих параметрів, у сучасній інженерній практиці застосовується низка специфічних RL-алгоритмів: DDQN (Double DQN) – вирішує проблему переоцінки (overestimation) значень Q-функції, використовуючи дві окремі мережі для вибору та оцінки дії, Dueling DDQN – розділяє оцінку стану середовища та перевагу конкретної дії, що дозволяє агенту краще розуміти цінність перебування в певній ситуації незалежно від дій, DDPG (Deep Deterministic Policy Gradient) – на відміну від попередніх, цей алгоритм призначений для середовищ із безперервним простором дій, працює поза політикою та використовує детерміновану політику й Q-функцію, якій він навчається за допомогою рівняння Беллмана.

Передовими методами DRL є Soft Actor-Critic (SAC), самокероване навчання з підкріпленням (Self-RL) та багатоагентне навчання з підкріпленням, засноване на фізичних принципах (MAPIRL), що призначені для оптимізації виробництва при одночасному явному управлінні ризиками. Self-RL (Self-Learning Reinforcement Learning): концепція, де агент здатний самостійно генерувати цілі або тренувальні сценарії, що дозволяє моделі ефективно навчатися навіть за відсутності чіткої зовнішньої винагороди на початкових етапах.

MAPIRL (Multi-Agent Physics-Informed Reinforcement Learning) включає: багатоагентність – система складається з кількох інтелектуальних одиниць, що взаємодіють між собою; фізичні принципи (Physics-Informed) – у функцію навчання інтегровані фізичні закони (наприклад, рівняння збереження маси, енергії або динаміки потоків), що гарантує, що рішення нейромережі не суперечитимуть реальності; явне управління ризиками (Explicit Risk Management) – на відміну від стандартного RL, який максимізує середній прибуток, ці підходи враховують «безпечні коридори», мінімізуючи ймовірність критичних збоїв або аварійних ситуацій під час оптимізації.

Незалежно від обраної архітектури – від базового DQN до багатоагентних фізично-інформованих систем – практична працездатність RL-агента критично залежить від коректного налаштування його гіперпараметрів та якості вхідних даних. Розгляньмо ці два аспекти докладніше.

По відношенню до оптимізації гіперпараметрів, то гіперпараметри DQN суттєво впливають на швидкість збіжності моделі, загальну продуктивність та стабільність навчання. Ініціалізація мережі гіперпараметрів можлива на основі емпіричних значень з обранням сумарної винагороди як метрики оцінки їх ефективності для отримання максимальної сумарної винагороди. Різниця між алгоритмом навчання з підкріпленням з оптимізованими гіперпараметрами та алгоритмом без них може означати різницю між практичною та надійною моделлю і моделлю, яка взагалі не сходиться.

Процес оптимізації зазвичай має таку структуру:

1. Ініціалізується обраний алгоритм оптимізації, і з попередньо заданого розподілу випадковим чином обирається набір (або набори) гіперпараметрів.

2. Алгоритму навчання з підкріпленням надається набір (або набори) гіперпараметрів, після чого здійснюється виклик симулятора для початку процесу навчання.

3. Процес навчання проводиться зі збором усієї необхідної інформації через задані інтервали часу. Результати передаються назад до алгоритму оптимізації.

4. На основі результатів навчання та вибору алгоритму оптимізації набір гіперпараметрів відповідним чином розширюється (оновлюється).

5. Кроки 2-4 повторюються доти, доки не буде досягнуто критерію зупинки.

6. Збір даних та обробка результатів.

Якщо акцентувати увагу на розрахунку значення винагороди, то у навчанні з підкріпленням значення винагороди відіграє вирішальну роль в оптимізації параметрів моделі, вимірюючи ефективність методу або дії. Точний розрахунок значення винагороди є необхідним, оскільки він безпосередньо впливає на оцінку градієнта мережі, що, зі свого боку, впливає на стабільність і швидкість збіжності моделі. Неправильний метод розрахунку значення винагороди може призвести до помилок у налаштуванні параметрів моделі, що сповільнить збіжність або навіть призведе до її збою.

Для розрахунку базового значення винагороди часто використовують метод простого ковзного середнього (SMA). Проте ковзне середнє надає однакову вагу всім даним у вікні і якщо бурове долото раптово переходить із м'якої породи у твердий вапняк, параметри (наприклад, ROP або крутний момент) змінюються миттєво, а алгоритм SMA використовує стару інформацію про м'які породи (ті що були 10 хвилин тому), тому модель видаватиме нерелевантні команди (наприклад, надмірне WOB). Формування винагороди може відбуватися довкола «базової лінії» (класична задача безпечного навчання з підкріпленням, де агент максимізує продуктивність, не порушуючи критичних фізичних обмежень) коли позитивна винагорода призначається, якщо ROP, отримана за обраною стратегією дій (WOB і RPM), перевищує базову ROP. Винагорода може видаватися також і за збереження ріжучої здатності долота і його мінімальне зношування, а також за мінімізацію часу непродуктивного буріння.

Запропонований авторами метод розрахунку базового значення винагороди на основі експоненційного ковзного середнього (EMA) на відміну від SMA, надає більшої ваги найостаннішій точності розпізнавання, послаблюючи при цьому вплив попередньої точності. Це дозволяє базовому значенню винагороди швидко відображати зміни в продуктивності моделі, скорочуючи затримки та прискорюючи процес оптимізації.

Розрахунок базового значення на основі EMA формулюється таким чином:

$$\text{baseline}_t = \left(\frac{2}{N_e+1}\right) \cdot A_t + \left(1 - \frac{2}{N_e+1}\right) \cdot \text{baseline}_{t-1} \quad (2)$$

де t позначає час (номер) поточної епохи, а baseline_t та baseline_{t-1} – це базове значення винагороди для циклів t та $t-1$ відповідно. A_t позначає точність розпізнавання на виході інтелектуальної моделі протягом циклу t. N_e позначає довжину вікна (кількість періодів) для EMA. Дріб $\frac{2}{N_e+1}$ є ваговим коефіцієнтом, який гарантує, що найостанніші показники точності мають більший вплив на базове значення. За такого підходу експоненційне згладжування дозволяє RL-агенту миттєво «скидати» застарілий досвід при різкій зміні літології та швидко адаптувати свою стратегію до нових умов. Оптимізацію процесу навчання з підкріпленням можна розглядати як оптимізацію будь-якої сильно нелінійної функції для досягнення оптимального набору параметрів, що приведе до найкращого можливого результату. Буріння – це класичний приклад багатовимірної, сильно нелінійної динамічної системи, і саме тому RL працює тут краще за традиційні PID-регулятори. Для вибору напрямків і постановки задач досліджень важливим є розглянути існуючі практичні кейси реалізації навчання з підкріпленням (RL) при спорудження нафтогазових свердловин, особливо при інтелектуальному управлінні режимами буріння. На практиці навчання з підкріпленням, у тому числі і глибоке навчання з підкріпленням (у багатовимірних просторах станів та складних задачах) можна розглядати як засіб дослідження оптимальної стратегії дій у безперервному процесі послідовного прийняття рішень. Стратегія послідовного ухвалення рішень передбачає баланс між безпосередньою вигодою (максимальна проходка) та безпекою буріння, а також надійністю інструменту. Отже, для знаходження оптимального рішення необхідний ретельний компроміс між поточними вигодами та майбутніми перевагами у вигляді довговічності інструменту й бездоганного (безаварійного) виконання робіт.

Сьогодні можливості використання алгоритмів навчання з підкріпленням при спорудженні нафтогазових свердловин також вивчають для підвищення ефективності похило-скерованого буріння (оптимізації механічної швидкості проходки, зменшення звивистості стовбура свердловини, скорочення чисельності персоналу на об'єкті та забезпечення узгодженості операцій. У задачах геонавігації навчання з підкріпленням як метод ухвалення рішень використовує релевантні масиви даних, зокрема результати каротажу (LWD) та відстань до меж пласта, для генерування обґрунтованих геонавігаційних команд. Головна мета агента – стратегічне позиціювання стовбура свердловини в межах продуктивного пласта для підвищення її кінцевої продуктивності. Функція винагороди в алгоритмах RL для геонавігації проєктується таким чином, щоб відображати операційні цілі, які зазвичай включають максимізацію довжини контакту з цільовим пластом та мінімізацію загальних експлуатаційних витрат. Узгоджуючи функцію винагороди із цими цілями, агент адаптується та навчається на кожному кроці ухвалення рішення, поступово вдосконалюючи свої стратегії для оптимізації просторових траєкторій свердловин у мінливих геологічних умовах), для підтримання постійного вибійного тиску при MPD (Managed Pressure Drilling) шляхом поєднання агента глибокого Q-навчання із пропорційно-інтегральним

регулятором та для оптимізації рішень щодо керування бурінням (автори розробили агента глибокого навчання з підкріпленням (DRL), заснованого на алгоритмі DDPG, для оптимізації рішень щодо керування бурінням шляхом вибору осьового навантаження на долото (WOB) та частоти обертання (RPM) з використанням вхідних даних у режимі реального часу, таких як механічна швидкість проходки (ROP), крутний момент і властивості пласта. У їхньому віртуальному середовищі були інтегровані фізичні моделі ROP та зносу, а також функція винагороди, що враховує довговічність долота, рівень вібрацій та ризики відмов (аварій)).

Зміна параметрів буріння оптимізується з огляду на абразивність породи та цільову (проектну) глибину буріння: у м'яких породах модель RL застосовує верхню межу осьового навантаження на долото та частоти обертання по всій глибині буріння для максимізації механічної швидкості проходки та скорочення часу буріння, у твердих і абразивних породах модель RL поступово змінює частоту обертання та навантаження на долото, щоб запобігти передчасному зносу різців долота та у нестабільних умовах, модель RL обмежує співвідношення навантаження на долото до частоти обертання (WOB/RPM) для запобігання вібраціям та проковзуванню. Також навчання з підкріпленням активно використовують в таких областях , як управління і оптимізація буріння, при проводці похилоскерованих свердловин, при вирішенні задач геонавгації в автономному скерованому бурінні, під час прогнозування проблем пов'язаних із прихопленням бурильної колони у процесі буріння, чи втратою циркуляції; при плануванні розробки родовищ (розміщення свердловин, операціях із заводнення), під час предиктивного керування генераторами для бурових робіт.

У процесі буріння система геонавігації в режимі реального часу оцінює стан буріння та приймає рішення щодо його напрямку, отримуючи інформацію про буріння в реальному часі, спрямовуючи бурове долото в цільовий пласт і збільшуючи швидкість буріння. У процесі інтенсивного навчання інтелектуальної системи геонавігації під час буріння кожна керуюча дія отримує винагороду від стану навколишнього середовища: перехід в ідеальний стан повинен винагороджуватися позитивно, а відхилення від ідеального стану – негативно. Зі збільшенням часу роботи на конкретному нафтовому родовищі зростає досвід навчання системи керування, і запропонована стратегія керування стає більш відповідною умовам буріння на родовищі. Однак кореляція між безперервним досвідом призводить до того, що система керування стає нечутливою до реакції на конкретні ситуації, і стає неможливим своєчасно надавати відповідні стратегії керування для впливу на процес буріння і для виправлення ситуації можна застосувати метод фіксованих Q-цілей.

У задачі проєктування траєкторії свердловини навчання з підкріпленням має такі переваги: агент може самостійно видавати керуючі команди відповідно до поточного стану та історичного досвіду, реалізуючи самокероване формування траєкторії, агент може коригувати напрямок відповідно до стану в режимі реального часу в процесі виконання траєкторії та має здатність до відстеження траєкторії і динамічного оновлення та функція винагороди є гнучкою і може враховувати різні інженерні обмеження, такі як обмеження інтенсивності викривлення, вимоги до зенітного кута та азимута свердловини, геологічні перешкоди тощо, що підвищує стійкість та узагальнюючу здатність стратегії.

Висновки з даного дослідження і перспективи подальших розвідок у даному напрямі

У запропонованому дослідженні проаналізовано наявні підходи навчання з підкріпленням у завданнях інтелектуального управління режимами буріння свердловин та показано їхню перевагу над традиційними емпіричними методами та класичними контурними регуляторами в умовах невизначеності та нестаціонарності середовища. На основі узагальнення результатів вітчизняних і закордонних робіт дослідження встановлено, що алгоритми RL і DRL (DQN, DDPG, PPO, TRPO, SAC, CQL, багатоагентні та фізично-орієнтовані архітектури) здатні забезпечувати послідовне прийняття рішень у термінах «стан–дія–винагорода», оптимізуючи осьове навантаження на долото, частоту обертання, параметри промивання, траєкторію та пов'язані з ними показники ROP, зносу інструменту, вібрацій і ризиків аварій.

У роботі порівняно різні класи алгоритмів RL, *розроблено узагальнену концептуальну схему агента* для керування режимами буріння, запропоновано використання цифрових двійників і симуляторів як базового середовища тренування, а також окреслено підходи до формування й калібрування функції винагороди (зокрема з використанням SMA та EMA) та оптимізації гіперпараметрів. Представлено, що коректний вибір коєфіцієнта дисконтування, структури мережі «актор-критик» та стратегії оновлення цільових параметрів є критично важливим для стабільності й збіжності навчання.

Перспективи подальших досліджень пов'язані з поглибленою розробкою та реалізацією прототипів RL-агентів у промислових цифрових двійниках бурових систем, інтеграцією багатоагентних і physics-informed підходів, розширенням функції винагороди.

Література

1. Advanced oil field development planning using deep reinforcement learning with reservoir-invariant graph neural network transfer learning / E. Ikponmwoba et al. *Offshore technology conference*, Houston, Texas, USA, 4–7 May 2026. 2026. URL: https://doi.org/10.4043/36830-ms

2. Maximizing efficiency of deep-reinforcement learning agents in autonomous directional drilling with hyperparameter optimization / V. Kesireddy et al. *Unconventional resources technology conference*, Colorado

Convention Center, Denver, Colorado, US, 13–15 June 2023. Tulsa, OK, USA, 2023. URL: https://doi.org/10.15530/urtec-2023-3865879

3. A reinforcement learning method for optimal control of oil well production using cropped well group samples / Y. Ding et al. *Heliyon*. 2023. P. e17919. URL: https://doi.org/10.1016/j.heliyon.2023.e17919

4. Deep reinforcement learning for constrained field development optimization in subsurface two-phase flow / Y. Nasir et al. *Frontiers in applied mathematics and statistics*. 2021. Vol. 7. URL: https://doi.org/10.3389/fams.2021.689934

5. Ismailov S. Z., Shmoncheva Y. Y., Jabbarova G. V. Application of machine learning algorithms for optimizing the trajectory of inclined wells. *SOCAR proceedings*. 2024. SI1. P. 89–94. URL: https://doi.org/10.5510/ogp2024si101004

6. Machine learning for improved directional drilling / J. Pollock et al. *Offshore technology conference*, Houston, Texas, USA. 2018. URL: https://doi.org/10.4043/28633-ms

7. Zhao X., Yin H., Li Q. An autonomous optimization and control method for drilling parameters based on reinforcement learning. *Ssrn*. 2024. 28 October. P. 1–16. URL: https://ssrn.com/abstract=5001631.

8. Research on wellbore trajectory optimization and drilling control based on the TD3 algorithm / H. Gu et al. *Applied sciences*. 2025. Vol. 15, no. 13. P. 7258. URL: https://doi.org/10.3390/app15137258