The University of Tehran Press

Journal of Sustainable Energy Systems

Home Page: https://ses.ut.ac.ir

Online ISSN: 2980-8693

An Integrated Reinforcement Learning-AHP Decision Framework for Optimizing Crude Oil Allocation in Refinery and Petrochemical Systems

Sattar Zavvari[1] | Amir Ali Saifoddin Asl[2*] | Fathollah Pourfayaz[3]

1. PhD Candidate, Department of Energy and Sustainable Resources Engineering, Faculty of Interdisciplinary Sciences and Technologies, University of Tehran, Tehran, Iran. Email: sattar.zavvari@ut.ac.ir
2. Corresponding Author, Assistant Professor, Department of Energy and Sustainable Resources Engineering, Faculty of Interdisciplinary Sciences and Technologies, University of Tehran, Tehran, Iran. Email: saifoddin@ut.ac.ir
3. Professor, Department of Energy and Sustainable Resources Engineering, Faculty of Interdisciplinary Sciences and Technologies, University of Tehran, Tehran, Iran. Email: pourfayaz@ut.ac.ir

ARTICLE INFO

Article type:
Research Paper

Article History:
Received: 19 February 2026
Revised: 16 March 2026
Accepted: 12 May 2026
Published Online: 23 September 2026

Keywords:
Oil Industry,
Energy Policy,
Refinery,
Petrochemical Industry,
Reinforcement Learning.

ABSTRACT

In recent years, creating a robust strategy to allocate crude oil has been proven difficult due to the sudden and harsh changes in product prices, energy markets, and trading volumes. Traditional optimization methods, whether deterministic or heuristic, fail to optimize the allocation of crude oil, mainly because of various economic goals and the nonlinear dynamics of refinery and petrochemical markets. To address these difficulties, this study aims to develop a policy-driven framework that combines Reinforcement Learning (RL) with the Analytic Hierarchy Process (AHP). We will use AHP to assess expert preferences through pairwise comparisons. This process will generate structured weights for multiple economic and operational criteria. These weights are then embedded directly into the RL reward function, which will enable the agent to learn allocation strategies both adaptive to market transitions and aligned with managerial priorities. The proposed RL–AHP decision framework is trained and evaluated using live transaction, price, and volume data from the Energy Exchange. Simulation results demonstrate that the agent reliably converges to a stable policy, with high agreement between expert optimal and learned decisions (Accuracy: 94.44%). We will gain substantial economic growth by implementing the learned framework. This growth leads to improved total profitability by 10.23%, increased revenue by 1.32%, and reduced operational costs by 8.91%. As a result, the general profit increases from 4.12 billion USD to 4.54 billion USD.

Cite this article: Zavvari, S.; Saifoddin Asl, A. A. & Pourfayaz, F. (2026). An Integrated Reinforcement Learning-AHP Decision Framework for Optimizing Crude Oil Allocation in Refinery and Petrochemical Systems. *Journal of Sustainable Energy Systems*, 5 (4), 741-761. DOI: http//doi.org/10.22059/ses.2026.413963.1238

DOI: http//doi.org/10.22059/ses.2026.413963.1238

Publisher: University of Tehran Press.

1. Introduction

Optimal crude oil allocation across refinery and petrochemical systems is a strategic challenge in energy governance, shaped by complex economic, operational, and policy-driven considerations. Under highly volatile energy markets, conventional rule-based or static optimization approaches often fail to adapt to dynamic transitions and cannot adequately incorporate policymakers' preferences into the decision structure. Reinforcement learning (RL), with its ability to learn from experience and respond to real market conditions, provides a promising foundation for developing adaptive allocation strategies. However, RL alone may overlook essential managerial priorities and institutional constraints. To bridge this gap, the present study introduces an integrated decision-making framework that combines the Analytic Hierarchy Process (AHP) with reinforcement learning. Expert preferences

are transformed into policy weights and embedded directly into the agent's reward function, ensuring that the learned policy aligns with both market realities and managerial logic.

2. Materials and Methods

The proposed framework is built upon an MDP consisting of 243 states and three primary allocation actions. States are defined based on five key market variables: crude oil price, refinery product price, petrochemical product price, refinery trading volume, and petrochemical trading volume. Policy weights derived from pairwise comparisons conducted by 15 experts (with acceptable consistency ratios) are incorporated into a weighted reward function which introduces a policy layer within the learning process. The agent is trained using Q-learning and real transaction data from the Iran Energy Exchange. Through iterative temporal-difference updates, the learner gradually converges toward a policy jointly driven by data-based evidence and expert-informed priority weights.

3. Results

Convergence curves indicate that after several hundred episodes, the cumulative reward increases steadily and its variance decreases, signaling stabilization of the learning process and the emergence of a robust policy. A balanced confusion matrix reveals that the learned policy matches expert policy in 93.67 percent of cases, achieving an F-score of 91.83 percent. Examination of the reward matrix illustrates that expert-driven policy preferences—such as favoring allocations toward more profitable or strategically stable downstream pathways—are strongly reflected in the agent's behavioral patterns. Allocation trajectories over the 40-day simulation horizon also show smooth, interpretable, and market-consistent adjustment patterns. After an initial exploration phase, the agent consistently selects actions that not only improve financial return but also adhere to expert logic.

Economic performance analysis shows that the expert-designed policy yields 4.12 billion USD in profit, whereas the RL-learned policy increases profit to 4.54 billion USD, representing a 10.23 percent improvement. This enhancement arises from a 1.32 percent increase in revenues—achieved through redirecting resources toward more profitable market states—and an 8.91 percent reduction in costs through lower exposure to high-cost production structures. These outcomes demonstrate that the agent successfully captured the intricate interactions among market states, weighted policy rewards, and operational returns, ultimately generating a significantly higher economic value.

4. Discussion and Conclusion

The findings confirm that integrating AHP with reinforcement learning creates an effective approach for designing adaptive, value-aligned allocation policies. The agent not only converged to a numerically optimal solution but also acquired a policy consistent with expert reasoning and institutional priorities. Enhanced profitability, stable allocation patterns, and strong policy convergence collectively highlight the potential of this hybrid framework to support transparent, explainable, and economically beneficial decision-making in energy governance. This methodology can serve as a foundation for intelligent advisory systems in energy markets, and future research may extend it by incorporating additional environmental variables and employing deep reinforcement learning architectures.

انتشارات دانشگاه تهران

فصلنامهٔ سیستم‌های انرژی پایدار

سایت نشریه: https://ses.ut.ac.ir

شاپا الکترونیکی: ۸۶۹۳-۲۹۸۰

چارچوب یکپارچهٔ تصمیم‌گیری مبتنی بر یادگیری تقویتی برای بهینه‌سازی تخصیص نفت خام در سامانه‌های پالایشی و پتروشیمی

ستار زواری[۱] | امیرعلی سیف‌الدین اصل[۲*] | فتح‌الله پورفیاض[۳]

۱. دانشجوی دکتری، گروه مهندسی انرژی و منابع پایدار، دانشکدگان علوم و فناوری های میان‌رشته‌ای، دانشگاه تهران، تهران، ایران. رایانامه: sattar.zavvari@ut.ac.ir

۲. نویسندهٔ مسئول، استادیار، گروه مهندسی انرژی و منابع پایدار، دانشکدگان علوم و فناوری‌های میان‌رشته‌ای، دانشگاه تهران، تهران، ایران. رایانامه: saifoddin@ut.ac.ir

۳. استاد، گروه مهندسی انرژی و منابع پایدار، دانشکدگان علوم و فناوری‌های میان‌رشته‌ای، دانشگاه تهران، تهران، ایران. رایانامه: pourfayaz@ut.ac.ir

اطلاعات مقاله

نوع مقاله:
پژوهشی

تاریخ‌های مقاله:
تاریخ دریافت: ۱۴۰۴/۱۱/۳۰
تاریخ بازنگری: ۱۴۰۴/۱۲/۲۵
تاریخ پذیرش: ۱۴۰۵/۰۲/۲۲
تاریخ انتشار: ۱۴۰۵/۰۷/۰۱

کلیدواژه:
صنعت نفت،
سیاست انرژی،
پالایشگاه،
صنعت پتروشیمی،
یادگیری تقویتی.

چکیده

تدوین راهبردی پایدار برای تخصیص نفت خام همواره با چالش های ساختاری قابل توجهی مواجه بوده است؛ چالش‌هایی که اغلب از نوسانات پیش‌بینی‌ناپذیر قیمت نفت خام، فرآورده‌های نفتی و حجم معاملات آن‌ها در بازارهای انرژی ناشی می‌شوند. این نوسانات، همراه با پویایی‌های غیرخطی و تعاملات پیچیده میان بازارهای نفتی، کارآمدی رویکردهای سنتی بهینه‌سازی (مدل‌های قطعی یا روش‌های ابتکاری) را در ایجاد راه‌حل‌های پایدار محدود می‌کند. پژوهش حاضر یک چارچوب سیاست‌محور مبتنی بر تلفیق یادگیری تقویتی و فرایند تحلیل سلسله‌مراتبی ارائه می‌کند. در این چارچوب، ترجیحات خبرگان استخراج و به وزن‌های کمی برای معیارهای عملیاتی تبدیل می‌شود. با استفاده از این وزن‌ها، تابع پاداش مدل یادگیری تقویتی تولید می‌شود تا عامل یادگیرنده سیاست‌های تخصیص نفت خام را در شرایط متغیر بازار و منطبق با اولویت‌های سیاستی بیاموزد. چارچوب پیشنهادی با داده‌های واقعی قیمت و حجم معاملات بورس انرژی ارزیابی شده است. نتایج شبیه‌سازی نشان می‌دهد عامل به یک سیاست بهینهٔ همگرا می‌شود و تصمیم‌های آن همخوانی معناداری با ارزیابی‌های کارشناسی دارد. افزون بر این، اجرای چارچوب موجب بهبود قابل ملاحظه در عملکرد اقتصادی شده و افزایش ۱۰/۲۳ درصد در سودآوری کل را به همراه داشته است؛ به گونه ای که سود خالص سالانه از ۴/۱۲ میلیارد دلار به ۴/۵۴ میلیارد دلار افزایش می‌یابد.

استناد: زواری، ستار؛ سیف‌الدین اصل، امیرعلی و پورفیاض، فتح‌الله (۱۴۰۵). چارچوب یکپارچهٔ تصمیم‌گیری مبتنی بر یادگیری تقویتی برای بهینه‌سازی تخصیص نفت خام در سامانه‌های پالایشی و پتروشیمی. *فصلنامهٔ سیستم‌های انرژی پایدار*، ۵ (۴) ۷۴۱-۷۶۱.

DOI: http//doi.org/10.22059/ses.2026.413963.1238

ناشر: مؤسسه انتشارات دانشگاه تهران.

DOI: http//doi.org/10.22059/ses.2026.413963.1238

۱. مقدمه

نفت و محصولات پتروشیمی از راهبردی‌ترین بخش‌های نظام‌های انرژی در جهان به شمار می‌آیند و در بسیاری از اقتصادها نقشی محوری در تنوع‌بخشی صنعتی و ایجاد ثبات مالی ایفا می‌کنند [۱]. در کشورهای وابسته به منابع هیدروکربوری، تصمیم‌گیری دربارهٔ تخصیص نفت خام به واحدهای پالایشی یا صنایع پتروشیمی، تنها بر سودآوری کوتاه‌مدت فعالیت‌های عملیاتی اثر نمی‌گذارد، بلکه بر پایداری زنجیره‌های تأمین انرژی در بلندمدت نیز تأثیری تعیین‌کننده دارد [۲].

با توجه به تغییرات سریع و پویای بازار انرژی، همسوسازی تصمیم‌های عملیاتی با اهداف سیاستی مختلف و گاه متعارض (همچون کاهش هزینه، ثبات تولید و ...) به یکی از محورهای اصلی پژوهش‌های نوین تبدیل شده است [۳]. روش‌های متداول برای مدیریت تخصیص نفت خام از واحدهای تولیدی معمولاً مبتنی بر بهینه‌سازی ایستا یا رویکردهای ابتکاری هستند؛ در این روش‌ها تصمیم‌ها بر اساس مجموعه‌ای از قواعد ثابت اتخاذ می‌شوند که از پارامترهای اقتصادی، رفتارهای قطعی و قضاوت‌های شهودی استخراج شده‌اند [۴]. هرچند این رویکردها در شرایط بازار پایدار عملکرد قابل قبولی داشته‌اند، اما عوامل متعدد ناشی از عدم قطعیت‌های متغیر در تقاضا، نوسان ظرفیت تولید طی زمان و پیامدهای ناشی از تغییر محدودیت‌های سیاستی را در نظر نمی‌گیرند [۵].

در این‌میان، پیشرفت‌های اخیر در یادگیری تقویتی چشم‌انداز جدیدی برای مسئلهٔ تخصیص نفت ایجاد کرده است. یادگیری تقویتی این امکان را فراهم می‌کند که عامل تصمیم‌ساز از طریق تعامل مستمر با محیط، به‌تدریج سیاست‌های بهینه را بیاموزد [۶]. به خلاف رویکردهای سنتی که بر معادلات ثابت تکیه دارند، یادگیری تقویتی قادر است بر اساس گذار میان حالت‌ها و پاداش‌هایی که به عنوان بازخورد دریافت می‌کند، راهبردهای تصمیم‌گیری سازگار را ایجاد کند. این ویژگی یادگیری تقویتی را به گزینه‌ای مناسب برای تصمیم گیری در سامانه‌های صنعتی بزرگ، پیچیده و چندبعدی تبدیل می‌کند. با این‌حال، اتکای صرف به الگوریتم‌های یادگیری خودکار ممکن است به نادیده گرفته شدن برخی اولویت‌های راهبردی یا الزامات مقرراتی منجر شود که از دیدگاه خبرگان حوزه اهمیت ویژه‌ای دارند. برای کاهش این فاصله میان خودمختاری ماشین و قصد و راهبرد مدیریتی، فرایند تحلیل سلسله‌مراتبی می‌تواند به عنوان لایه‌ای مکمل از استدلال انسانی در کنار تصمیم‌گیری الگوریتمی به کار گرفته شود [۷].

تحلیل سلسله‌مراتبی چارچوبی نظام‌مند برای مقایسه‌های زوجی میان معیارهای ارزیابی فراهم می‌آورد و این امکان را به مدیران می‌دهد تا اهمیت نسبی شاخص‌هایی همچون هزینه‌های عملیاتی، ذخایر راهبردی و انطباق زیست‌محیطی را تعیین کنند. به این‌ترتیب، قضاوت‌های کیفی به وزن‌های کمی تبدیل می‌شوند و این وزن‌ها می‌توانند مستقیم در ساختار تصمیم‌گیری عامل‌های یادگیری تقویتی که با محیط تعامل دارند، ادغام شوند. وزن‌های استخراج‌شده از تحلیل سسلسله‌مراتبی همچنین می‌توانند جهت‌دهی لازم را برای سیاست‌های تصمیم‌گیری فراهم کنند و در شرایطی که پاداش‌های حاصل از تعامل عامل با محیط با اولویت‌های خبرگان یا سیاست‌های ملی همسو نباشد، ساختار تابع پاداش را اصلاح و تنظیم کنند [۸]. هم‌افزایی میان یادگیری تقویتی و تحلیل سلسله‌مراتبی در قالب یک چارچوب تصمیم‌گیری سیاست‌محور، پیوند مستقیمی میان یادگیری خودکار و دانش نهادی برقرار می‌کند. چنین چارچوبی امکان تنظیم لحظه‌ای سیاست‌های تخصیص منابع را فراهم می‌آورد، بدون آنکه شفافیت و قابلیت تفسیر در فرایند حکمرانی تضعیف شود [۹]. چارچوب پیشنهادی با ادغام دانش خبرگان در یک الگوریتم یادگیری تطبیقی، می‌تواند یکی از کاستی‌های اساسی مدل‌های موجود در مدیریت انرژی یعنی نبود پیوند میان بیشینه‌سازی مبتنی بر داده و فرایند سیاست‌گذاری راهبردی را برطرف کند [۱۰]. افزون بر این، مدل ترکیبی پژوهش حاضر قابلیت تعمیم به سایر مسائل تخصیص منابع فراتر از نفت خام را نیز دارد؛ از جمله مسائلی مانند برنامه‌ریزی و توزیع برق [۱۱]، ادغام منابع انرژی تجدیدپذیر در شبکه [۱۲] و سامانه‌های پاسخگویی تقاضا در صنایع مختلف [۱۳]. با توجه به اهمیت راهبردی حفظ رقابت‌پذیری و پایداری در زنجیرهٔ ارزش صنعت نفت، پژوهش حاضر به طراحی یک معماری جامع مبتنی بر یادگیری تقویتی می‌پردازد که بر پایهٔ پارامترهای وزنی استخراج‌شده از نظر خبرگان شکل گرفته است. هدف این چارچوب، دستیابی به سازوکاری بهینه، منصفانه و منطبق با سیاست‌های کلان برای توزیع نفت خام میان بخش‌های مختلف است. چارچوب پیشنهادی سهمی در گسترش ادبیات روبه‌رشد سامانه‌های تصمیم‌گیری تطبیقی چندمعیاره دارد [۱۴].

در بخش دوم، مروری بر ادبیات جدید مربوط به یادگیری تقویتی و سامانه‌های تصمیم‌گیری چندمعیاره در حوزهٔ بهینه‌سازی انرژی ارائه می‌شود. بخش سوم به تبیین چارچوب روش‌شناسی پژوهش اختصاص دارد و طراحی تابع پاداش مبتنی بر تحلیل سلسله‌مراتبی و ساختار سیاستی را تشریح می‌کند. در بخش چهارم، نتایج شبیه‌سازی ارائه‌شده و یافته‌ها از منظر عملیاتی و سیاست‌گذاری مورد تحلیل قرار می‌گیرند. در نهایت، بخش پنجم به جمع‌بندی پژوهش و ارائهٔ پیشنهادهایی برای مسیرهای آتی تحقیق در حوزهٔ حکمرانی تطبیقی انرژی اختصاص دارد.

۲. مروری بر ادبیات پژوهش

۲–۱. مدل‌های قطعی و ریاضی در تخصیص منابع

رویکردهای بهینه‌سازی در مدیریت منابع نفت و پتروشیمی عمدتاً بر استفاده از روش‌های قطعی و مدل‌های مبتنی بر فرمول‌بندی‌های ریاضی استوار بوده‌اند؛ رویکردهایی که در آن‌ها به جای استفاده از سازوکارهای تطبیقی یا مبتنی بر سیاست، مجموعه‌ای از قواعد ثابت برای تصمیم‌گیری تدوین می‌شد. در سال ۱۹۶۲، آرونافسکی و ویلیامز [۱۵] نخستین کسانی بودند که از برنامه‌ریزی خطی برای بهینه‌سازی تولید نفت خام استفاده کردند. پژوهش آنان با معرفی محدودیت‌های ریاضی در مدل‌سازی، زمینهٔ شکل‌گیری چارچوب‌های تحلیلی در مدل‌های اولیه برنامه‌ریزی پالایشگاهی را فراهم ساخت. در ادامه، براون و برنهام [۱۶] با توسعهٔ مدل‌های ریاضی مبتنی بر سینتیک واکنش‌ها، ارتباط میان سازوکارهای واکنشی و نرخ‌های تولید را تبیین کردند و به این‌ترتیب، مبنایی استاندارد برای مدل‌سازی فرایندهای پالایشگاهی ارائه دادند.

اوایل دههٔ ۲۰۰۰، پارادایم مدل‌های قطعی به سمت روش‌هایی با توان بهینه‌سازی در ابعاد بالاتر حرکت کرد؛ این تحول با استفاده از برنامه‌ریزی خطی عدد صحیح آمیخته و گسترش‌های تصادفی آن برای مدیریت عدم‌قطعیت در عملیات پالایشگاهی همراه بود. در این‌زمینه، ریباس و همکاران [۱۷] و نیز شاه و میشرا [۱۸] با توسعهٔ تکنیک‌های پیشرفته و برنامه‌ریزی تصادفی، نقش مهمی در بهبود روش‌های مدیریت عدم‌قطعیت در عملیات پالایشگاهی ایفا کردند. به طور مشابه، ایوانف و ری [۱۹] با معرفی الگوریتم‌های ژنتیک چندهدفه، رویکردی نوین برای بهینه‌سازی مبادله‌های ایستا میان سودآوری و عملکرد زیست‌محیطی پالایشگاه‌ها ارائه کردند و در واقع، الگوریتم‌های جست‌وجوی ابتکاری را در مراحل اولیه به مدل‌های کلاسیک پالایشگاهی وارد ساختند. در ادامه، لی و همکاران [۲۰] یک مدل برنامه‌ریزی خطی تصادفی دومرحله‌ای ارائه کردند که به طور صریح عدم‌قطعیت در تقاضای بازار را در نظر می‌گرفت و به این‌ترتیب پلی میان بهینه‌سازی قطعی و رویکردهای احتمالی ایجاد کرد. در سال‌های اخیر نیز سونی و همکاران [۲۱] با استفاده از برنامه‌ریزی خطی به بهینه‌سازی برنامه‌های توسعهٔ میادین غیرمتعارف پرداختند و کراسنیوک و همکاران [۲۲] مفهوم جامعی از یکپارچه‌سازی اقتصادی مبتنی بر مدل‌های ریاضی ارائه دادند که عملیات بالادستی و پایین‌دستی صنعت نفت را به طور هم‌زمان در بر می‌گرفت. با وجود قطعیت ریاضی و پایداری محاسباتی این مطالعات، آن‌ها در مجموع نمایانگر لایه‌ای از رویکردهای قطعی در بهینه‌سازی منابع نفتی هستند که محدودیت‌هایی اساسی در مواجهه با پویایی‌های سیاست‌گذاری، وزن‌دهی مبتنی بر نظر خبرگان و بازخوردهای بلادرنگ بازار دارند. افزون بر این، این مدل‌ها به افق‌های بهینه‌سازی بلندمدت و نسبتاً پایدار متکی هستند؛ امری که ضرورت حرکت به سوی رویکردهای منعطف‌تر و سیاست‌محور را برجسته می‌سازد، موضوعی که در بخش‌های ۲.۲ و ۳.۲ مورد بررسی قرار خواهد گرفت.

۲–۲. گذار سیاستی و مدل‌های رفتاریِ تصمیم‌گیری چندمعیاره

از اوایل دههٔ ۱۹۹۰ تا اواسط دههٔ ۲۰۱۰، شیوه‌های تخصیص منابع از سوی سیاست‌گذاران دستخوش تحول قابل توجهی شد؛ به گونه‌ای که مدل‌سازی تخصیص منابع بیش از پیش با تفسیر بسترهای پیچیده اجتماعی و سیاسی پیوند خورد. مدل‌های قطعی اولیه، هرچند از دقت ریاضی بالایی برخوردار بودند، در بازنمایی متغیرهایی که به‌سادگی قابل کمی‌سازی نبودند (مانند قضاوت خبرگان، اختیار مدیریتی و ترجیحات سیاستی در سامانه‌های پیچیدهٔ انرژی) ناتوان بودند. با آشکار شدن این واقعیت که فرایند تصمیم‌گیری متأثر از عوامل شناختی و کیفی نیز هست، کیفر [۲۳] مفهوم ریسک‌گریزی و وابستگی احتمالی را وارد مسئلهٔ تخصیص منابع کرد و نشان داد گرایش‌های روان‌شناختی تصمیم‌گیرندگان تأثیر قابل توجهی بر تعیین نتایج بهینه دارند. به همین

ترتیب، کوپرسمیت و همکاران [۲۴] بررسی کردند که ارزش اطلاعات درون‌سازمانی چگونه بر انسجام تصمیم‌های مدیریتی اثر می‌گذارد و چگونه تصمیم‌های مدیران تحت تأثیر ارزیابی خبرگان و کیفیت اطلاعات دریافتی آن‌ها شکل می‌گیرد. در ادامه، با ورود صریح‌تر اهداف سیاستی به مدل‌ها، استفاده از استدلال چندمعیاره نیز برجسته‌تر شد. در این رویکردها، به معیارهای ارزیابی وزن‌هایی اختصاص داده می‌شد که از طریق آن‌ها اهداف راهبردی ملی بازتاب می‌یافت [۲۵]. در همین راستا، یک چارچوب تصمیم‌گیری مبتنی بر سیاست برای برنامه‌ریزی توسعهٔ پایدار پیشنهاد شد. همچنین، مک کنزی و همکاران [۲۶] مفهوم تخصیص پویا در سامانه‌های منابعِ به‌هم‌وابسته را مطرح کردند؛ مفهومی که به‌نوعی گذار به مدل‌هایی را پیش‌بینی می‌کرد که در آن‌ها انتقال میان حالت‌ها در محیط‌های پیچیده به رسمیت شناخته می‌شود.

مجموع این مطالعات، زمینه‌ساز نوعی گذار سیاستی شد که تصمیم‌گیری را از اتکای صرف بر بهینه‌سازی ریاضی، به سمت تلفیقی از بهینه‌سازی ریاضی، قضاوت انسانی، تحمل عدم‌قطعیت و اولویت‌های سیاستی نهادها سوق داد. بنیانی که این پژوهش‌ها فراهم کردند، در ادامه برای توسعهٔ چارچوب‌های تصمیم‌گیری تطبیقی و داده‌محور مورد استفاده قرار گرفت؛ چارچوب‌هایی که در زیربخش ۳.۲ تشریح خواهند شد.

۲–۳. هوش مصنوعی در تصمیم‌گیری برای انرژی و تخصیص منابع

ورود هوش مصنوعی به بخش انرژی نقطه عطفی مهم در گذار از مدل‌های سنتی مبتنی بر بهینه‌سازی و سیاست‌گذاری کلاسیک به شمار می‌آید. سامانه‌های هوشمند توانستند بدون اتکا به قواعد قطعی ازپیش‌تعریف‌شده، تعاملات پویا میان بازار، تولید و محیط‌های عملیاتی را در فرایند تصمیم‌گیری لحاظ کنند. این تحول روش‌شناختی، عنصر استقلال تصمیم‌گیری و سازگاری‌پذیری را وارد مطالعات تخصیص منابع کرد؛ عناصری که به طور بنیادین بر الگوهای کنترل و مدیریت در صنعت نفت و پتروشیمی اثر گذاشتند. یافته‌های اولیه پژوهش‌ها نشان می‌داد هوش مصنوعی قادر است پیچیدگی‌های بیشتر و الگوهای متغیر مصرف انرژی را با دقت بیشتری پردازش کند. سودهاکار [۲۷] روش‌های یادگیری محاسباتی را برای برنامه‌ریزی و زمان‌بندی پالایشگاه‌ها توسعه داد تا امکان واکنش به‌موقع در برابر شرایط غیرقابل پیش‌بینی فراهم شود. در همین راستا، کوون و همکاران [۲۸] نوعی سامانهٔ تصمیم‌گیری مبتنی بر هوش مصنوعی در زنجیرهٔ تأمین پتروشیمی ارائه کردند که از الگوریتم‌های بهینه‌سازی هوشمند برای بهینه‌سازی هم‌زمان جریان مواد، ساختار هزینه‌ها و روش‌های پردازش بهره می‌برد. این دستاورد از نخستین نمونه‌های ادغام هوش مصنوعی در منطق مدیریتی بود و نشان داد تصمیم‌گیری در صنعت انرژی فقط به بهینه‌سازی عددی محدود نمی‌شود، بلکه شناخت، تفسیر و سازگاری با شرایط پیچیده نیز بخشی از ماهیت آن است.

پیشرفت‌های این فناوری دیگر تنها به افزایش کارایی الگوریتم‌ها محدود نمی‌شود، بلکه ابعاد جدیدی از استدلال راهبردی مبتنی بر هوش مصنوعی را نیز در بر گرفته است. همان‌گونه که راجا و همکاران [۲۹] و اودیمارها و همکاران [۳۰] نشان داده‌اند، عامل‌های یادگیرنده می‌توانند در هماهنگ‌سازی منابع و کنترل فرایندها نقش مؤثری ایفا کنند؛ به گونه‌ای که چرخه‌های بازخوردی تطبیقی و خوداصلاح‌گر به اجزای اصلی سامانه‌های مبتنی بر هوش مصنوعی تبدیل شوند. افزون بر این، کوئی و یائو [۳۱] با بهره‌گیری از استدلال ترکیبی عصبی، مسئلهٔ خودکارسازی تصمیم‌گیری در عملیات صنعتی مقیاس بزرگ را بررسی کردند و نشان دادند سامانه‌های هوش مصنوعی می‌توانند از طریق تعبیهٔ قواعد درون ساختار سیستم، زمینه و محدودیت‌های حاکم بر تصمیم‌های سیاستی را درک کنند. این قابلیت می‌تواند فاصلهٔ میان تصمیم‌گیرندهٔ انسانی و بهینه‌سازی خودکار را تا حد زیادی کاهش دهد. روند تکامل این حوزه همچنین شواهد روشنی از بلوغ فناوری‌های هوش مصنوعی در اختیار قرار می‌دهد. برای مثال، ما و همکاران [۳۲] از عامل‌های هوشمند برای هماهنگ‌سازی تعاملات پتروشیمی طی زنجیرهٔ ارزش پالایشگاه‌ها استفاده کردند؛ عامل‌هایی که نه بر پایهٔ برنامه‌ریزی صریح قبلی، بلکه بر اساس تجربه و یادگیری تدریجی توسعه یافته بودند. همچنین، در مطالعات لی و همکاران [۳۳] و شی و همکاران [۳۴] مشهود است که حوزهٔ یادشده در حال عبور از تصمیم‌گیری‌های صرفاً مبتنی بر تجربه‌های گذشته به سوی چارچوب‌های تصمیم‌گیری است که بر پایهٔ سیاست‌گذاری و ملاحظات انسانی شکل گرفته‌اند.

مجموع این پژوهش‌ها به‌روشنی نشان می‌دهد ورود هوش مصنوعی به عرصهٔ تخصیص منابع، نوعی جدید از هوشمندی تصمیم‌گیری تطبیقی را پدید آورده است که وابستگی به پیش‌بینی‌ها را به عنوان مبنای اصلی تصمیم‌گیری کاهش می‌دهد و در مقابل، امکان یادگیری مستمر، خودسازمان‌دهی و تصمیم‌گیری حساس به زمینه را در اجتماعات پیچیدهٔ انرژی فراهم می‌سازد. از این‌رو، کانون توجه پژوهش‌ها از مدل‌های پیش‌بینی‌محور و محدود به چارچوب‌های نهادی، به سمت معماری‌های هوش مصنوعی تصمیم‌محور و مبتنی بر یادگیری از تجربه سوق یافته است؛ موضوعی که در بخش ۴.۲ تشریح خواهد شد.

۲-۴. مطالعات ترکیبی و نوآوری پژوهش

پژوهش‌های جدید، از جمله مطالعهٔ لی و همکاران [۳۳]، کارایی یک مدل تلفیقی یادگیری تقویتی و فرایند تحلیل سلسله‌مراتبی را در اولویت‌بندی گزینه‌های پتروشیمی نشان داده‌اند و به این‌ترتیب شواهدی ارائه کرده‌اند مبنی بر اینکه الگوریتم‌های تطبیقی می‌توانند با سازوکارهای وزن‌دهی چندمعیاره ادغام شوند. به طور مشابه، وانگ و همکاران [۳۵] و کوئی و همکاران [۳۴] با بهره‌گیری از تحلیل سلسله‌مراتبی و سایر چارچوب‌های تصمیم‌گیری چندمعیاره، این رویکردها را در حوزهٔ سرمایه‌گذاری، اولویت‌بندی و برنامه‌ریزی ظرفیت انرژی پیزوالکتریک به کار برده‌اند و از ساختارهای منطقی سلسله‌مراتبی برای ارزیابی سیاست‌ها استفاده کرده‌اند. در امتداد این پژوهش‌ها، تاتار و همکاران [۳۶] و بهول و همکاران [۳۷] سامانه‌های هوش مصنوعی سیاست‌محوری طراحی کرده‌اند که اصول حکمرانی را در معماری‌های یادگیری خود لحاظ می‌کنند تا شفافیت فرایندهای تصمیم‌گیری را ارتقا دهند. در مجموع، این مطالعات از شکل‌گیری چشم‌اندازی نوظهوری خبر می‌دهد که در آن الگوریتم‌ها با چارچوب‌های تصمیم‌گیری ترکیب می‌شود تا بهترین نتیجهٔ نظری به دست آید. با این‌حال، این دستاوردها هنوز به طور کامل به ترجمهٔ عملی و پیاده‌سازی مؤثر در نظام‌های موجود تخصیص منابع بازار منجر نشده‌اند و شکافی میان نوآوری نظری و کاربرد عملی در مقیاس بازار همچنان پابرجاست.

این پژوهش برای نخستین‌بار، کاربرد هم‌زمان دو روش یادگیری تقویتی و فرایند تحلیل سلسله‌مراتبی را به منظور توسعهٔ یک چارچوب سیاست‌محور برای پشتیبانی از تصمیم‌گیری چندمعیاره در فرایندهای یادگیری تطبیقی مرتبط با تخصیص منابع نفتی و پتروشیمیایی ارائه می‌کند. در این مدل، سامانهٔ پاداش‌دهی یادگیری تقویتی با ارائهٔ سیگنال‌های کمی مبتنی بر ترجیحات و دیدگاه‌های خبرگان عمل می‌کند و از طریق داده‌های تجربی واقعی داده‌های بازار بورس انرژی ایران اعتبارسنجی شده است. ترکیب استدلال سیاست‌محور (کیفی) با رویکرد یادگیری تقویتی (کمی) موجب شکل‌گیری قابلیتی دوگانه در سیستم می‌شود: از یک‌سو، قابلیت تبیین‌پذیری و انطباق با سیاست‌ها را فراهم می‌سازد و از سوی دیگر، امکان پویایی و سازگاری الگوریتمی را در فرایند یادگیری حفظ می‌کند. به همین‌دلیل، چارچوب پیشنهادی می‌تواند به صورت مشابه در تمامی جنبه‌های صنایع دارای تخصیص منابع و قانون‌گذاری شده به کار رود. در نتیجه، این مدل نوآوری محسوسی را در عرصهٔ حکمرانی منابع مبتنی بر سیاست‌گذاری معرفی می‌کند و به عنوان یک رویکرد روش‌شناختی جدید، ارتباط میان سیاست، یادگیری تطبیقی و تصمیم‌گیری چندمعیاره را به شکلی منسجم و قابل کاربرد برقرار می‌سازد.

۳. روش‌شناسی

۳-۱. فرایند تحلیل سلسله‌مراتبی

فرایند تحلیل سلسله‌مراتبی یک روش تصمیم‌گیری چندمعیاره است که یک مسئلهٔ پیچیده را به ساختاری سلسله‌مراتبی شامل معیارها، زیرمعیارها و گزینه‌ها تجزیه می‌کند. در این روش، وزن نسبی معیارها از طریق مقایسه‌های زوجی تعیین می‌شود. نتیجهٔ این مقایسه‌ها، با نرمال‌سازی ماتریس مقایسه، به یک بردار وزنی تبدیل می‌شود. برای ارزیابی میزان سازگاری قضاوت‌ها از نسبت سازگاری استفاده می‌شود که نشان می‌دهد ارزیابی‌های خبرگان تا چه حد منطقی و سازگار هستند. معمولاً اگر مقدار این شاخص کمتر از ۰/۱ باشد، سطح سازگاری قابل قبول تلقی می‌شود. در این مطالعه، مقایسه‌های زوجی میان معیارهای سیاستی بازار توسط یک پنل خبرگان متشکل از ۱۵ نفر انجام شد. حاصل این فرایند، وزن‌های سیاستی کمی است که به طور مستقیم در تابع

پاداش الگوریتم یادگیری تقویتی مورد استفاده قرار می‌گیرد. به این‌ترتیب، فرایند تحلیل سلسله‌مراتبی این امکان را فراهم می‌کند که دانش کیفی خبرگان به راهنمای عددی و کدگذاری‌شده تبدیل شود و در فرایند یادگیری تقویتی به کار گرفته شود.

۳-۲. یادگیری تقویتی

یادگیری تقویتی نوعی چارچوب یادگیری است که در آن یک عامل با محیط تعامل می‌کند؛ به این‌صورت که در حالت S_t، اقدام A_t را انتخاب می‌کند و در قبال آن، پاداشی به صورت $R_{AHP}(S_t, A_t)$ دریافت می‌کند. هدف یادگیری تقویتی یادگیری بهترین سیاست تصمیم‌گیری $\pi^*(A,S)$ است؛ سیاستی که بتواند در هر وضعیت بازار، بهترین اقدام تخصیصی را انتخاب کند، به گونه‌ای که پاداش وزنی بلندمدت (مشتق‌شده از تحلیل سلسله‌مراتبی) بیشینه شود (شکل ۱). محیط تصمیم‌گیری به صورت یک فرایند تصمیم‌گیری مارکوف تعریف می‌شود (رابطهٔ ۱):

$$MDP = (S, A, P, R_{AHP}) \quad (۱)$$

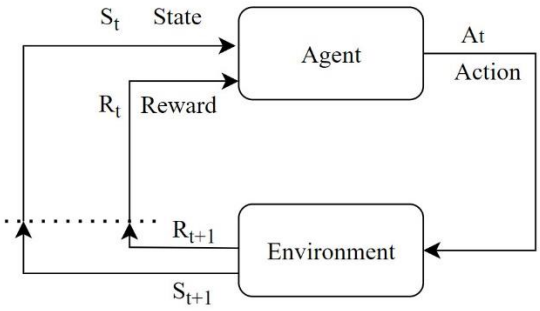
که در آن S فضای حالت‌های بازار، A مجموعه اقدامات قابل انجام در تخصیص منابع، $P(S' | S, A)$ احتمال انتقال از حالت S به حالت S' بر اثر اقدام A و $R_{AHP}(S,A)$ تابع پاداش مشتق‌شده از تحلیل سلسله‌مراتبی که بر اساس وزن‌های سیاستی خبرگان تعریف شده است. در هر اجرای الگوریتم، عامل یادگیری تقویتی دو مؤلفه کلیدی را ارزیابی می‌کند. یکی $V_\pi(S)$ به معنای ارزش مورد انتظار حالت S در صورت پیروی از سیاست π و دیگری $Q_\pi(S,A)$ به معنای ارزش اقدام، که نشان می‌دهد با انتخاب اقدام A در حالت S، و ادامه پیروی از سیاست π، چه میزان ارزش یا بازده مورد انتظار حاصل می‌شود. این توابع به صورت تکراری و بر اساس روابط تعریف‌شده در فرمول بازگشتی بلمن به‌روزرسانی می‌شوند تا در نهایت، به سیاست بهینهٔ π^* همگرا شوند. مقدار Q با استفاده از روش اختلاف زمانی که یکی از تکنیک‌های رایج برای به‌روزرسانی مقادیر Q در یادگیری تقویتی است، اصلاح می‌شود. در این روش، به‌روزرسانی به صورت رابطهٔ ۲ انجام می‌گیرد:

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha\left[R_{AHP}(S_t, A_t) + \gamma \max_{A'} Q(S_{t+1}, A') - Q(S_t, A_t)\right] \quad (۲)$$

که در آن α نرخ یادگیری، γ ضریب تنزیل پاداش‌های آتی، S_{t+1} حالت بعدی بازار پس از اقدام و $\max_{A'} Q(S_{t+1}, A')$ بیشینهٔ ارزش اقدام‌های ممکن در حالت بعدی است. به این‌ترتیب، یادگیری تقویتی با اتکا بر پاداش‌های وزن‌دهی‌شده مبتنی بر تحلیل سلسله‌مراتبی، به‌صورت پویا و تجربه‌محور سیاستی را می‌آموزد که با ترجیحات سیاستی خبرگان هم‌راستا بوده و در عین حال، توان سازگاری با شرایط متغیر بازار را حفظ می‌کند و رابطهٔ نهایی آن به صورت رابطهٔ ۳ خواهد بود:

$$Q_{t+1}(S_t, A_t) = Q_t(S_t, A_t) + \alpha\left[R_{AHP}(S_t, A_t) + \gamma \max Q_t(S_{t+1}, A_{t+1}) - Q_t(S_t, A_t)\right] \quad (۳)$$

در اینجا، α نشان‌دهندهٔ نرخ یادگیری است که میزان یا شدت به‌روزرسانی‌ها را مشخص می‌کند و γ معرف ضریب تنزیل است که بر میزان تأثیر پاداش‌های آتی در محاسبات فعلی نظارت دارد. الگوریتم در نهایت به سمت بهینه‌ترین سیاست تخصیص میل می‌کند و میان جست‌وجو برای یافتن اقدامات جدید و بهره‌برداری از راهبردهای دارای پاداش بیشتر، تعادل برقرار می‌سازد.

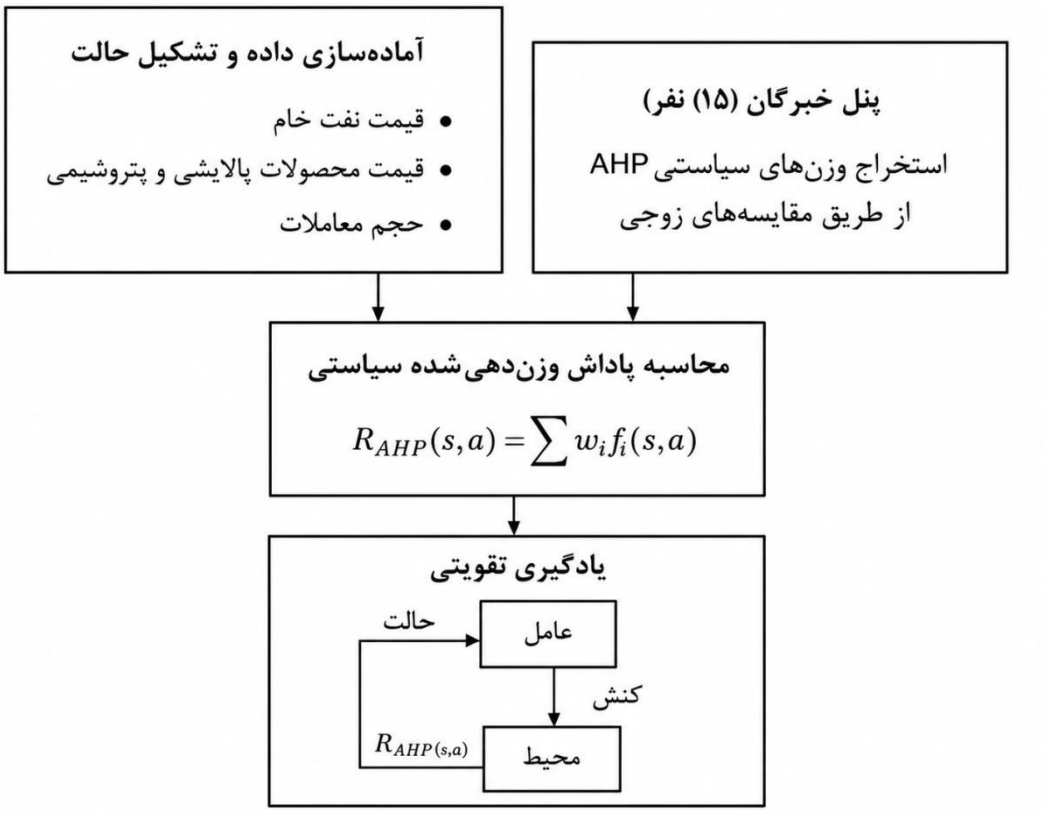


شکل ۱. ساختار یادگیری تقویتی

۳-۳. چارچوب و جریان فرایند

چارچوب پیشنهادی در این پژوهش، یک سامانهٔ یادگیری تقویتی سیاست‌محور است که وظیفهٔ تخصیص بهینهٔ نفت خام بین بخش‌های پتروشیمی و پالایشگاهی را به عهده دارد. این ساختار با بهره‌گیری از دانش خبرگانِ موضوعی، اهمیت و وزن سیاست‌های مختلف را از طریق فرایند تحلیل سلسله‌مراتبی تعیین کرده و این داده‌ها را در محاسبات مربوط به تابع پاداشِ یادگیری تقویتی ادغام می‌کند. در نتیجه، هر تصمیم اتخاذشده توسط عامل یادگیری تقویتی، به طور مستقیم با اولویت‌های راهبردی تبیین‌شده توسط خبرگان هم‌سو خواهد بود. این بخش با مدل‌سازی محیط یادگیری تقویتی به صورت یک فرایند تصمیم‌گیری مارکوف تحقق می‌یابد. در این فرایند، حالت S_t مجموعه‌ای از ویژگی‌های قابل مشاهده بازار و متغیرهای عملیاتی است؛ از جمله قیمت نفت خام، قیمت محصولات پالایشگاهی، قیمت محصولات پتروشیمی و میزان محصولات معامله‌شده. در مقابل، اقدامات A_t به معنای نحوهٔ تخصیص محصول به مصرف‌کنندگان مختلف است یعنی افزایش سهم پالایشگاه، حفظ سطح فعلی تخصیص یا افزایش سهم پتروشیمی.

هدف عامل یادگیری تقویتی، یادگیری سیاست بهینه $\pi(A,S)$است؛ سیاستی که بتواند در این مدل مارکوف، با در نظر گرفتن وزن‌های سیاستی که توسط خبرگان تعیین شده‌اند، بیشترین پاداش را تولید کند. به بیان دیگر، عامل باید طوری یاد بگیرد که در هر وضعیت بازار، بهینه‌ترین تصمیم تخصیص را مطابق با معیارهای سیاستی انتخاب کند. عامل یادگیری تقویتی با اطلاعات واقعی معاملات و داده‌های بازار که از بورس انرژی ایران استخراج می‌شود، تعامل مستقیم دارد. تجربه‌ای که عامل یادگیری تقویتی از طریق این داده‌ها به دست می‌آورد به صورت پیوسته و پویا افزایش می‌یابد. عامل یادگیری تقویتی با استفاده از تجربه‌های انباشته‌شده خود، به صورت مستمر تابع ارزش اقدام $Q(s,a)$را به‌روزرسانی می‌کند تا در نهایت سیاستی برای بهینه‌ترین قاعده تخصیص ایجاد کند و تضمین کند که این قاعده، از نظر عملی و سیاستی، بهترین نتیجه را ارائه می‌دهد. تابع پاداش وزن‌دهی‌شده R_{AHP} به عنوان معیار اصلی برای به‌روزرسانی سیاست عامل عمل می‌کند. این تابع یک ساختار ترکیبی دارد که شامل داده‌های عملکردی تاریخی (داده‌های واقعی بازار) و وزن‌های حاصل از AHP (ترجیحات سیاستی خبرگان) (شکل ۲).



شکل ۲. چارچوب یادگیری تقویتی سیاست‌محور برای تخصیص بهینهٔ نفت خام

۳-۴. تعریف فضای حالت – اقدام یادگیری تقویتی

در مدل یادگیری تقویتی تخصیص نفت خام، محیط تعامل عامل از طریق فضای حالت (S) و فضای اقدام (A) نمایش داده می‌شود. این دو مؤلفه، متغیرهای کلیدی بازار و گزینه‌های مدیریتی مرتبط با نحوهٔ توزیع نفت خام میان صنایع پتروشیمی و پالایشگاهی را هدایت می‌کنند. به بیان دیگر، هر وضعیت بازار و هر تصمیم تخصیص، نمایانگر پویایی واقعی بازار انرژی و منطق راهبردی مدیران در انتخاب مسیر بهینهٔ تخصیص است.

۳-۴-۱. فضای حالت (S)

فضای حالت، توصیف‌کننده وضعیت‌های قابل درک بازار انرژی در زمان t بر حسب پنج متغیر اصلی است. این فضا به صورت رابطهٔ ۴ تعریف می‌شود:

$$S = \{P_O, P_R, P_P, V_R, V_P\} \quad (۴)$$

که در آن P_O قیمت نفت خام، P_R قیمت محصولات پالایشگاهی، P_P قیمت محصولات پتروشیمی، V_R حجم معاملات محصولات پالایشگاهی و V_P حجم معاملات محصولات پتروشیمی است. تمام این متغیرها، بر اساس روندهای بازار، در قالب سه وضعیت کیفی خلاصه می‌شوند: در حال افزایش، باثبات، در حال کاهش. به این‌ترتیب، هر یک از پنج متغیر می‌تواند یکی از این سه وضعیت را به خود بگیرد. با توجه به زمان گسسته، مجموعه‌ای متناهی و نماینده از حالت‌های ممکن بازار به دست می‌آید که نشان می‌دهد ۲۴۳ حالت متمایز برای توصیف وضعیت بازار در این مدل تعریف شده است.

۳-۴-۲. فضای اقدام

فضای اقدام، بیانگر تصمیم‌های مدیریتی مرتبط با بازتخصیص نفت خام میان بخش‌ها است. سه اقدام گسسته برای تصمیم‌گیری در نظر گرفته شده است (رابطهٔ ۵):

$$A = \{A_1, A_2, A_3\} \quad (۵)$$

که در آن A_1 تخصیص نفت بیشتر به بخش پتروشیمی، A_2 حفظ وضعیت موجود (بدون تغییر در تخصیص) و A_3 تخصیص منابع بیشتر به بخش پالایشگاهی است. هر اقدام باعث می‌شود حالت فعلی S_t به حالت بعدی S_{t+1} منتقل شود؛ این انتقال‌ها بر اساس داده‌های تاریخی بورس انرژی ایران مدل‌سازی می‌شوند. این ویژگی باعث می‌شود که تعامل عامل یادگیرنده قابل تفسیر باشد و رفتار آن ریشه در واقعیت بازار و نه فقط شبیه‌سازی فرضی داشته باشد برای هر زوج (S,A)، یک پاداش متناظر $R_{AHP}(S,A)$ تعریف می‌شود که مطابق با وزن‌هایی است که در مرحلهٔ ارزیابی خبرگان و از طریق فرایند تحلیل سلسله‌مراتبی استخراج شده‌اند. به این‌ترتیب، هر تصمیم تخصیصی نه‌تنها بازتاب‌دهندهٔ وضعیت بازار، بلکه منعکس‌کنندهٔ ترجیحات و اولویت‌های سیاستی خبرگان نیز هست.

۳-۴-۳. تابع پاداش

عملکرد پاداش در این مطالعه، در واقع تحقق پیوند اصلی میان ترجیحات سیاستی خبرگان و فرایند بهینه‌سازی در یادگیری تقویتی است. این تابع، میزان مطلوبیت انجام یک اقدام خاص (A_t) در وضعیتی مشخص (S_t) را برای تصمیم‌گیرندگان، بر اساس باورهای نظام‌مندی که از طریق فرایند تحلیل سلسله‌مراتبی استخراج شده است، می‌سنجد. طبق تعریف، تابع پاداش به صورت رابطهٔ ۶ است:

$$R_{AHP}(S_t, A_t) = \sum W_i r_i(S_t, A_t) \quad (۶)$$

که در آن $r_i(S_t, A_t)$ به مقدار کمی‌شده معیار iام (نظیر سودآوری، کارایی هزینه، توازن تولید) اشاره دارد و W_i وزنی است که از طریق روش تحلیل سلسله‌مراتبی محاسبه شده و نشان‌دهندهٔ اولویت قضاوت خبرگان است. به منظور تعیین W_i با استفاده از روش تحلیل سلسله‌مراتبی، از نظرات ۱۵ نفر از نخبگان دانشگاهی و خبرگان وزارت نفت در فرایند انجام مقایسات زوجی میان معیارهای مربوطه استفاده شد. این وزن‌ها به گونه‌ای نرمال‌سازی شدند که با آستانه نسبت سازگاری مطابقت داشته باشند؛ امری

که تضمین می‌کند مقایسات سلسله‌مراتبی از نظر آماری قابل اعتماد و دارای سازگاری درونی هستند. هر یک از عناصر پاداش، یعنی $r_i(S,A)$، مجموعه‌ای از متغیرهای مالی و عملیاتیِ مرتبط با تصمیم‌گیری‌های تخصیص را در خود خلاصه و تلفیق می‌کند.

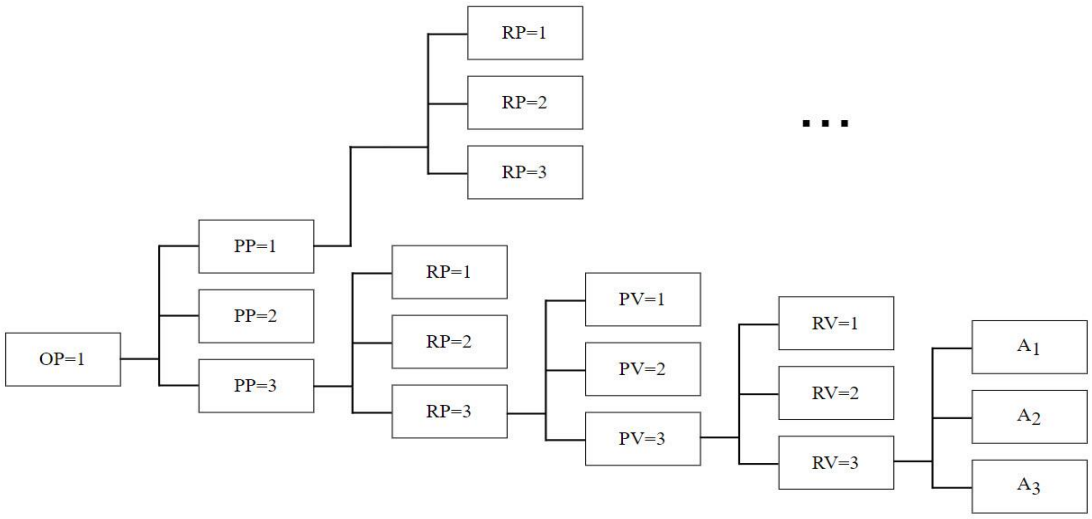
به عنوان مثال، افزایش سودآوری محصولات پتروشیمی (به دلیل عرضه محدود نفت)، به اختصاص پاداش قوی‌تری برای اقدام A_1 (تخصیص بیشتر به پتروشیمی) منجر می‌شود. به همین‌ترتیب، بهبود شرایط و بازدهی بخش پالایشگاهی، پاداش بیشتری را برای اقدام A_3 به همراه خواهد داشت و برقراری توازن میان این دو بخش، به تخصیص یک پاداش میانی برای اقدام A_2 (حفظ وضعیت موجود) منجر می‌شود. در نتیجه، ماتریس پاداش R_{AHP} به یک نمایش کمیِ سازمان‌یافته و قابل خواندن از قضاوت سیاستی انسان تبدیل می‌شود (جدول ۱ در بخش نتایج). این ماتریس به طور مستقیم در فرایند یادگیری تقویتی وارد می‌شود تا الگوریتم Q-Learning بتواند به‌روزرسانی سیاست‌ها را بر مبنای ترجیحات تعریف‌شده توسط خبرگان انجام دهد نه فقط بر پایهٔ پیش‌بینی‌های آماری داده‌محور. به بیان دیگر، سیستم یادگیرنده با اتکا به این ماتریس، سیاست‌های خود را به گونه‌ای بهینه‌سازی می‌کند که با اهداف راهبردی و قضاوت انسانی در حوزهٔ سیاست‌گذاری انرژی هم‌راستا باشند؛ و از این طریق، نقش دانش ضمنی و تجربی تصمیم‌گیرندگان در منطق یادگیری ماشین حفظ می‌شود.

۴. نتایج و تحلیل

در این بخش، یافته‌های حاصل از مدل سیاست‌محور ترکیبی فرایند سلسله‌مراتبی و یادگیری تقویتی برای تخصیص بهینهٔ نفت خام میان واحدهای پتروشیمی و پالایشی بر اساس داده‌های واقعی معاملات و قیمت‌های مبادله‌شده در بورس انرژی ایران ارائه می‌شود. همان‌طور که بیان شد، در این پژوهش یک سیاست تصمیم‌گیری برای تخصیص نفت به صنایع پتروشیمی و پالایشگاهی با بهره‌گیری از روش یادگیری تقویتی و تحلیل سلسله‌مراتبی ارائه می‌شود. مجموعه داده‌های مورد استفاده در این پژوهش، شامل داده‌های قیمت محصولات پتروشیمی و پالایشگاهی و همچنین، شاخص تقاضای این محصولات (بر پایهٔ حجم معاملات) طی سال‌های ۲۰۲۳ تا ۲۰۲۶ است. برای استخراج سیاست بهینه، چارچوب محاسباتی تحقیق بر پایهٔ زبان برنامه‌نویسی پایتون نسخهٔ ۹.۴ و روی یک رایانهٔ شخصی مجهز به پردازنده Intel Core i7 4720 (با فرکانس ۲/۶ گیگاهرتز) و ۱۶ گیگابایت رم پیاده‌سازی شده است. مدل یادگیری تقویتی با تعریف یک حالت ترکیبی (Composite State) متشکل از پنج زیر حالت توسعه یافته است که عبارت‌اند از: قیمت نفت (OP)، قیمت محصولات پتروشیمی (PP)، قیمت محصولات پالایشگاهی (RP)، حجم معاملات محصولات پتروشیمی (PV) و حجم معاملات محصولات پالایشگاهی (RV) که هر حالت بیانگر روند آتی زیر حالت‌ها در مقایسه با گام زمانی پیشین است. هر یک از متغیرهای یادشده می‌توانند یکی از سه وضعیت افزایش، ثبات یا کاهش را اتخاذ کنند. به عنوان مثال، حالت S(۲, ۳, ۱, ۱, ۱) معرف وضعیتی است که در آن قیمت‌های نفت، پتروشیمی و پالایشگاهی در حال افزایش و حجم معاملات محصولات پتروشیمی در حال کاهش است و حجم معاملات محصولات پالایشگاهی در وضعیت ثبات قرار دارند. علاوه بر این، مجموعه اقدامات با نماد A تعریف شده است که شامل سه گزینه: ۱- افزایش تخصیص نفت به بخش پتروشیمی؛ ۲- حفظ سهم تخصیص فعلی؛ ۳- افزایش تخصیص نفت به بخش پالایشگاهی است. با این فرمول‌بندی، مدیریت بهینهٔ تخصیص نفت از طریق اجرای اقدامات بهینه رخ می‌دهد که امکان انتقال از حالت فعلی به حالت آتی را فراهم می‌آورد. این انتقال از طریق مسیرهای طراحی‌شده بر اساس داده‌های تاریخی اجرا می‌شود که در نهایت به محاسبهٔ تابع احتمال انتقال از حالت اولیه به ثانویه منجر می‌شود.

برای تعیین مقادیر پاداش در تابع پاداش، از رویکردی سیستماتیک مبتنی بر فرایند تحلیل سلسله‌مراتبی به منظور استخراج وزن‌های مربوط به هر اقدام از سوی تصمیم‌گیرندگان استفاده شده است. در این راستا، پرسشنامه‌ای برای جمع‌آوری دیدگاه‌ها و ارزیابی‌های پانل کارشناسی طراحی و توزیع شد. این فرایند، تعیین دقیق مقادیر پاداش در تابع یادشده را تضمین می‌کند. فرایند مقایسات زوجی و استخراج وزن‌ها از طریق روش تحلیل سلسله‌مراتبی به شیوه‌ای شفاف و سیستماتیک انجام شد تا ضمن تضمین قابلیت تکرارپذیری، متدولوژی سوگیری‌ها را به حداقل برساند. همچنین، تنوع اعضای پانل کارشناسی مانع از غلبه یک دیدگاه واحد بر فرایند تصمیم‌گیری شده و احتمال تأثیر سوگیری‌های فردی بر چارچوب کلی مدل را کاهش داده است. شکل ۳

نمونه‌ای از پرسشنامهٔ ارائه‌شده به پانل کارشناسی را نمایش می‌دهد. این سناریوی خاص برای نمایش یک حالت فرضی که در آن قیمت نفت رو به افزایش و قیمت محصولات پالایشگاهی و پتروشیمی رو به کاهش است، انتخاب شده است. هدف از انتخاب این مثال، نمایش دادن یک موقعیت پویا و پیچیده (مشابه شرایط واقعی صنعت) و سنجش پاسخ‌های کارشناسان در چنین سناریوهایی بود. در واقع، تحت این شرایط، پنل کارشناسی اولویت خود را به‌ترتیب برای اقدام ۱ نسبت به ۲، اقدام ۲ نسبت به ۳ و اقدام ۱ نسبت به ۳ مشخص می‌کند. وزن‌های مربوط به اقدامات ۱، ۲ و ۳ با استفاده از روش AHP تعیین شده و مقادیر پاداش متناظر از طریق مقایسات زوجی و تحلیل سلسله‌مراتبی استخراج شده‌اند (جدول ۱). نتایج جدول یادشده نشان می‌دهد با افزایش تقاضا برای محصولات پتروشیمی، از دیدگاه کارشناسان، تولید این محصولات به‌صرفه‌تر و سودآورتر شده و در نتیجه تخصیص نفت به صنعت پتروشیمی افزایش می‌یابد. به همین‌ترتیب، در صورت افزایش قیمت محصولات پالایشگاهی (با فرض ثابت ماندن سایر متغیرها)، میزان نفت تخصیص‌یافته به این بخش نیز افزایش خواهد یافت.



شکل ۳. پرسشنامهٔ سناریوی فرضی برای حالت OP=1

جدول ۱. ترجیحات پنل خبرگان و مقادیر پاداش وزن‌دار تحلیل سلسله‌مراتبی

Variables					Weights (the decision maker's preference)		
P_V	R_V	P_P	R_P	O_P	A_1	A_2	A_3
1	1	1	3	3	0.139	0.602	0.258
1	2	1	3	3	0.142	0.428	0.428
1	3	1	3	3	0.109	0.267	0.623
2	1	1	3	3	0.075	0.831	0.092
2	3	1	3	3	0.142	0.428	0.428
2	2	1	3	3	0.139	0.602	0.258
3	1	1	3	3	0.092	0.831	0.075
3	2	1	3	3	0.075	0.831	0.092
3	3	1	3	3	0.139	0.602	0.258

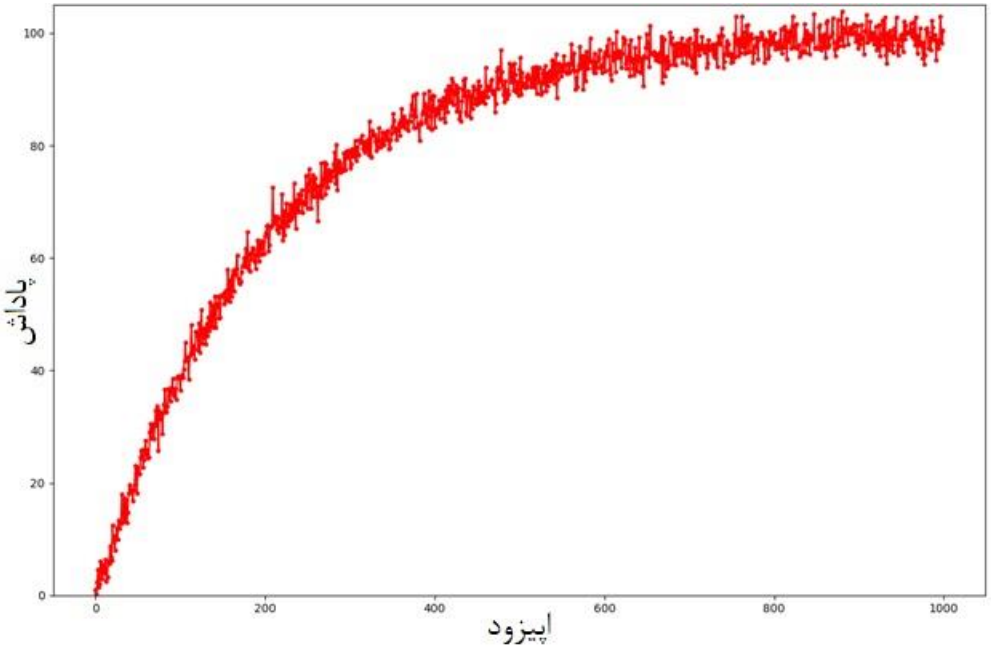
۴-۱. همگرایی پاداش و پایداری یادگیری

پس از تعریف مفاهیم حالت، اقدام و پاداش، بهره‌گیری از فرایند تصمیم‌گیری مارکوف برای فرمول‌بندی یک سیاست حاکم مورد نیاز است. این سیاست به عنوان یک راهنمای مسیر عمل کرده و با هدف نهایی بهینه‌سازی پاداش‌های بلندمدت مورد انتظار، حالت‌ها را به اقدامات متناظر نگاشت می‌کند. در مواردی که وابستگی‌های متقابل میان حالت، اقدام و پاداش به‌خوبی تثبیت شده باشد، استفاده از چارچوب تصمیم گیری مارکوف بسیار توصیه می‌شود؛ با این‌حال، در سناریوهایی که با اطلاعات ناقص یا محیط‌های پیچیده‌ای روبه‌رو هستیم که در آن چنین روابطی به طور دقیق تعریف نشده‌اند، الگوریتم یادگیری Q[1] به عنوان یک

1. Q-learning

جایگزین ارزشمند پدیدار می‌شود. در این پژوهش، نرخ یادگیری α=۰.۹ و ضریب تخفیف γ=۰.۸ در نظر گرفته شده است. شکل ۴ روند معناداری را نشان می‌دهد که در آن، حداکثر پاداش قابل دستیابی به‌سرعت به سوی یک مقدار بالا همگرا می‌شود. این مشاهده بیانگر آن است که عامل[1] توانسته است از طریق تصمیم‌گیری بهینه در هر حالت، به طور مؤثری توانایی بیشینه‌سازی پاداش تجمعی خود را طی یک اپیزود کسب کند.

همان‌گونه که در شکل ۴ نشان داده شده است، مقادیر پاداش اپیزودها طی ۱۰۰۰ اپیزود الگوریتم Q-learning دچار نوسان می‌شوند. در مرحلهٔ اولیه یادگیری، سیگنال پاداش بسیار آشفته و پرنوسان است؛ این امر عمدتاً ناشی از دو عامل است: نخست نرخ بالای اکتشاف در اپیزودهای اولیه و دوم، فقدان اطلاعات پیشین کافی برای عامل در خصوص پیامدهای اقدام‌ها در وضعیت‌های مختلف. با پیشرفت فرایند آموزش طی زمان، می‌توان دو روند متمایز در رفتار یادگیری عامل را شناسایی کرد. میانگین پاداش‌های تجمعی به صورت پیوسته افزایش می‌یابد، زیرا عامل به‌تدریج زوج‌های حالت ـ اقدام را شناسایی می‌کند که بازده‌های بالاتری را بر اساس پاداش وزن‌دار R_{AHP} تولید می‌کنند. این موضوع نشان می‌دهد عامل، سیاستی را می‌آموزد که با ترجیحات وزن‌گذاری‌شده خبرگان همسو است. واریانس پاداش‌های دریافتی عامل، پس از چندصد اپیزود نخست آموزش، به طور قابل ملاحظه‌ای کاهش می‌یابد. این کاهش واریانس بیانگر افزایش پایداری سیاست عامل و انتقال تدریجی از رفتار اکتشافی به سمت رفتار بهره‌بردارانه (استفاده از اقدام‌های بهینه‌شده) است. به این‌ترتیب، الگوی پاداش‌ها نشان می‌دهد فرایند یادگیری نه‌تنها همگرا شده، بلکه در نهایت به سیاستی پایدار و سازگار با ساختار پاداش مبتنی بر تحلیل سلسله‌مراتبی ختم می‌شود. این دو روند نشان می‌دهند تابع پاداش وزن‌گذاری‌شده یک سطح بهینه‌سازی مشخص، منظم و قابل‌یادگیری ایجاد کرده است. پایداری و هموار شدن منحنی همگرایی در اپیزودهای پایانی آموزش بیانگر آن است که عامل، به یک سیاست پایدار برای تخصیص نفت خام دست یافته که با دیدگاه‌ها و ترجیحات خبرگان سازگار است. به بیان دیگر، عامل یادگیرنده پس از طی مراحل اکتشاف، در نهایت روی سیاستی متمرکز می‌شود که نه تنها از منظر ریاضی بهینه است، بلکه از نظر سیاست‌گذاری انسانی و قضاوت کارشناسی نیز تأیید می‌شود.



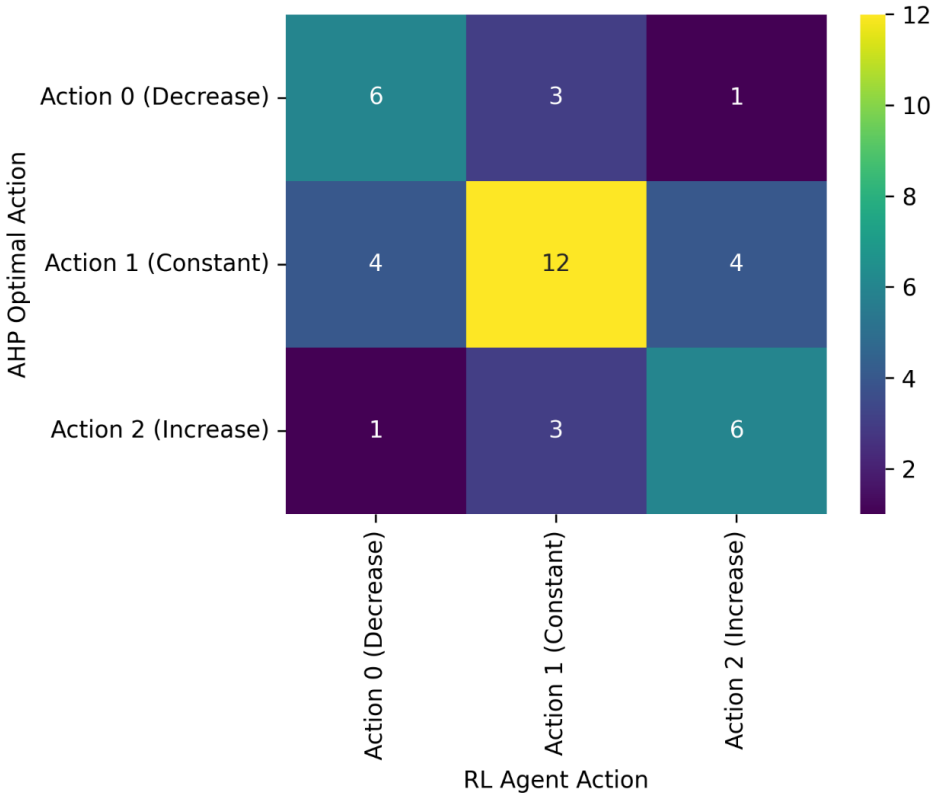
شکل ٤. همگرایی پاداش اپیزودها طی فرایند Q-Learning

1. Agent

۴-۲. ارزیابی سیاست و بررسی همخوانی با منطق خبرگان مبتنی بر تحلیل سلسله‌مراتبی

برای ارزیابی میزان انطباق میان سیاست یادگیری‌شده توسط الگوریتم یادگیری تقویتی و سیاست مرجع استخراج‌شده از نظر خبرگان بر اساس وزن‌های تحلیل سلسله‌مراتبی (جدول ۱)، یک ماتریس آشفتگی متوازن[1] مطابق با شکل ۵ ایجاد شد. در شکل ۵، غلبهٔ عناصر قطری نشان می‌دهد در اغلب وضعیت‌ها، عاملی که توسط یادگیری تقویتی انتخاب می‌شود همان اقدامی است که خبرگان ترجیح داده‌اند. مقادیر خارج از قطر عمدتاً در وضعیت‌های مرزی مشاهده می‌شوند؛ یعنی جایی که اختلاف پاداش میان دو اقدام بسیار اندک است. عملکرد مدل بسیار مطلوب بوده است؛ معیارهای ارزیابی به شرح زیر گزارش شد: دقت[2]: ۹۳/۶۷ درصد، دقت مثبت[3]: ۸۹/۸۱ درصد، بازخوانی[4]: ۹۷/۲۱ درصد و امتیاز F[5]: ۹۱/۸۳ درصد.

این شاخص‌های تجمیعی نشان می‌دهند عامل یادگیری تقویتی توانسته است منطق تصمیم‌گیری خبرگان را در سطح سیاست[6] بازسازی کند. این یافته‌ها تأیید می‌کنند که اولویت‌های کارشناسی قابل انتقال مستقیم به تابع پاداش هستند و می‌توانند در زمینهٔ یادگیری تقویتی، یک چارچوب مؤثر برای تقلید دقیق سیاست[7] ایجاد کنند.



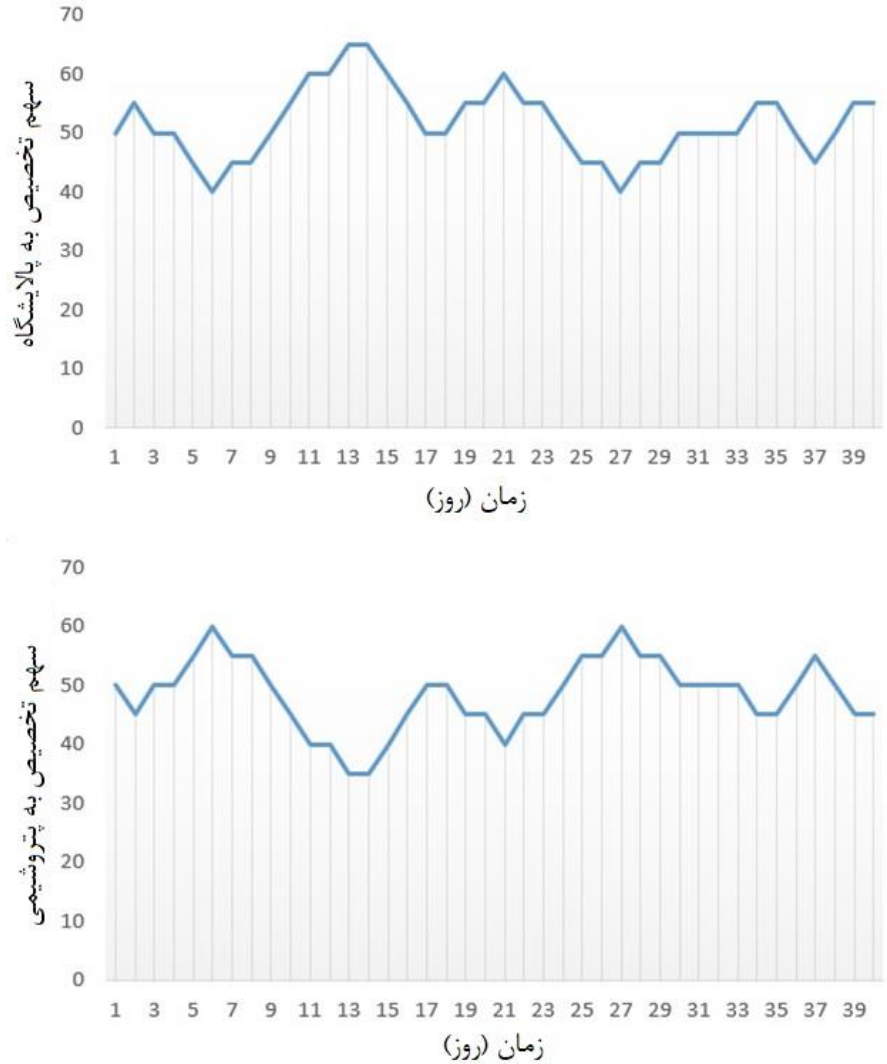
شکل ۵. ماتریس آشفتگی برای مقایسهٔ اقدامات بهینه و آموخته‌شده

۴-۳. مسیر تخصیص تحت سیاست آموخته‌شده

مسیرهای تخصیص از طریق شبیه‌سازی سیاست یادگیری‌شده یادگیری تقویتی روی دنباله‌ای از وضعیت‌های بازار تولید شده‌اند تا پیامدهای سیاست آموخته‌شده بر توزیع نفت خام میان صنایع پالایشی و پتروشیمی بررسی شود. روند تغییر سهم تخصیص طی

1- Balanced Confusion Matrix
2- Accuracy
3- Precision
4- Recall
5- F-score
6- Policy Level
7- Policy Imitation

زمان، برای سیاست یادگیری تقویتی آموخته‌شده، در شکل۶ نمایش داده شده است. مدل با یک تقسیم اولیه ۵۰—۵۰ آغاز می‌شود و در هر وضعیت، تغییرات به صورت گسسته و در قالب گام‌های ±۵ درصدی اعمال می‌شود. در شکل یادشده، برای هر یک از دو بخش (پتروشیمی و پالایشگاهی) دو منحنی ترسیم شده است تا الگوی تعدیل سهم‌ها هم بر اساس منطق خبره و هم بر اساس تصمیمات عامل یادگیری تقویتی پس از آموزش قابل مقایسه باشد. الگوهای حرکتی نشان می‌دهد سیاست حاصل، تغییرات تخصیص را به صورت هموار و پیوسته و در راستای منطق حرفه‌ای ایجاد می‌کند. هرچند سیاست یادگیری تقویتی در مراحل اولیه، به دلیل رفتار اکتشافی، نوسانات جزئی نشان می‌دهد، اما در ادامه به‌تدریج به همان الگوی تعدیل همسو با نظر خبرگان همگرا می‌شود. پس از همگرایی، مسیر تخصیص استخراج‌شده توسط یادگیری تقویتی به طور محسوسی به مسیر مبتنی بر نظر خبرگان نزدیک می‌شود. این نتیجه نشان می‌دهد چارچوب یادگیری تقویتی و تحلیل سلسله‌مراتبی قادر است نه‌تنها جهت تغییرات بخشی، بلکه زمان‌بندی و شدت (دامنه) جابه‌جایی‌های بین‌بخشی را نیز مطابق با ساختار پاداش خبره بازتولید کند. بنابراین، شکل ۶ نمایانگر توزیع سهم پتروشیمی و پالایشگاه در چارچوب سیاست آموخته شده است.

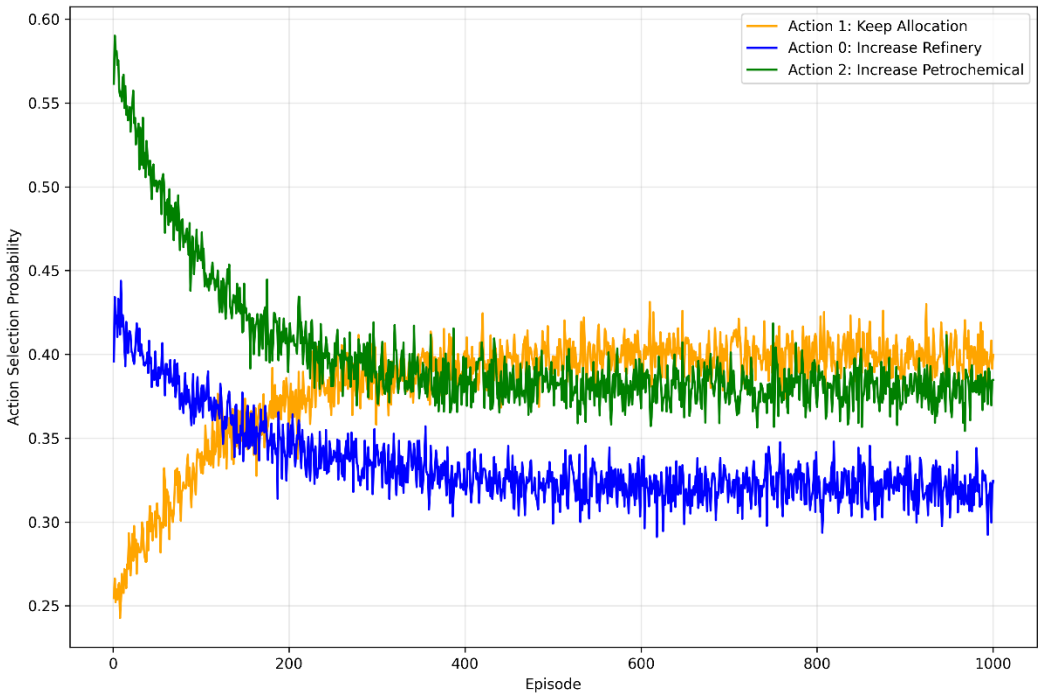


شکل ٦. سهم‌های روزانهٔ تخصیص برای بخش پالایشی (بالا) و بخش پتروشیمی (پایین) در بازه ٤٠روزه

۴-۴. تکامل سیاست در طول مسیر آموزش در یادگیری تقویتی

همان‌گونه که در شکل ۷ مشاهده می‌شود، روند تغییرات احتمال انتخاب اقدامات طی زمان، الگوی مشخصی را دنبال می‌کند که نشان‌دهندهٔ مراحل مختلف یادگیری عامل یادگیری تقویتی است. تحول این سه احتمال انتخاب اقدام را می‌توان در قالب سه فاز

یادگیری متمایز طبقه‌بندی کرد. در فازهای ابتدایی یادگیری (اپیزودهای < ۲۰۰)، رفتار عامل تقریباً کاملاً یکنواخت است؛ به طوری که هر سه اقدام با احتمال تقریباً برابر انتخاب می‌شوند. این وضعیت نشان‌دهندهٔ دوره‌ای است که در آن عامل تعداد زیادی اقدام تصادفی تولید می‌کند و هنوز تجربهٔ کافی از ارتباط حالت – پاداش در اختیار ندارد تا بتواند از دانش خود دربارهٔ محیط به شکل کارآمد استفاده کند. در فاز میانی آموزش (تقریباً اپیزودهای ۲۰۰ تا ۴۰۰)، قدرت ساختار پاداش مبتنی بر R_{AHP} به‌تدریج آشکار می‌شود. در این بازه عامل با انباشته شدن تجربه، برآوردهای دقیق‌تری از پاداش‌های متناظر با اقدامات مختلف به دست می‌آورد و احتمال انتخاب اقدامِ مورد ترجیح خبرگان افزایش می‌یابد. در مقابل، احتمال انتخاب سایر گزینه‌های تخصیص کاهش پیدا می‌کند. پدیدار شدن جهت‌مندی در الگوهای یادگیری نشان می‌دهد عامل در حال تشخیص این است که کدام اقدام‌ها ارزش نسبی بیشتری دارند؛ و به این‌ترتیب، قادر می‌شود طی زمان، تصمیم‌های مناسب‌تری دربارهٔ انتخاب الگوی تخصیص اتخاذ کند. در فاز پایانی (اپیزودهای > ۴۰۰)، منحنی‌های احتمال به یک پیکربندی پایدار نزدیک می‌شوند که در آن، یک اقدامِ غالب به‌وضوح بر سایرین برتری می‌یابد. در این مرحله عامل به طور مداوم و با احتمال بالا همان اقدام را انتخاب می‌کند و می‌توان نتیجه گرفت که عامل در حال اجرای سیاستی تثبیت‌شده است. فرایند تصمیم‌گیری عامل در این نقطه با منطق تصمیم‌گیری نهفته در ساختار پاداش وزن‌دار هم‌سوتر است. به بیان دیگر، عامل نه فقط به یک جواب عددی، بلکه به یک سیاست هم‌راستا با منطق و اولویت‌های خبرگان دست یافته است.



شکل ۷. تحول احتمالات انتخاب اقدام طی زمان

۴–۵. پیامدهای عملیاتی و اقتصادی

به منظور ارزیابی پیامدهای عملیاتی و مالی به‌کارگیری چارچوب جدید ترکیبی، هر دو سیاست یعنی سیاست مبتنی بر نظر خبرگان (خط مبنا) و سیاست استخراج‌شده توسط الگوریتم روی یک مجموعه یکسان از داده‌های تاریخی بازار در دورهٔ ارزیابی اجرا شدند. در این چارچوب ارزیابی درآمدها و هزینه‌های مرتبط با هر دو راهبرد تخصیص، به صورت روزانه و بر اساس قیمت‌های واقعی محصولات پتروشیمی و پالایشی و همچنین، حجم معاملات تحقق‌یافته در هر روز از دورهٔ آزمون محاسبه شد. سپس، جمع سودهای خالص روزانه، مبنای محاسبهٔ شاخص‌های سودآوری تجمیعی گزارش‌شده در این بخش قرار گرفت. بر اساس این روش ارزیابی، مشخص شد که سودآوری کل سیاست پایه مبتنی بر خبرگان در دورهٔ مورد بررسی، حدود ۴/۱۲ میلیارد

دلار بوده است. این مقدار، حاصل تجمیع سود خالص روزانه‌ای است که در نتیجهٔ به‌روزرسانی تخصیص نفت مطابق با قواعد تعریف‌شده توسط خبرگان به دست آمده است.

در مقابل، زمانی که همان توالی شرایط روزانهٔ بازار بر اساس سیاست آموخته‌شده شبیه‌سازی شد، سودآوری تجمعی به ۴/۵۴ میلیارد دلار افزایش یافت. این بهبود، بازتاب‌دهندهٔ تفاوت در تصمیمات تخصیص روزانه‌ای است که توسط سیاست یادگرفته‌شده تولید شده‌اند. تحلیل جزئی‌تر منشأ این افزایش سودآوری نشان می‌دهد دو عامل اصلی در این بهبود نقش داشته‌اند: افزایش درآمدها به میزان ۱/۳۲ درصد که این افزایش ناشی از گرایش سیاست این پژوهش به بازتوزیع تخصیص‌ها میان جریان‌های محصولاتی است که در روزهای خاص، حاشیهٔ سود تحقق‌یافتهٔ بالاتری نشان داده‌اند. به بیان دیگر، سیاست یادگرفته‌شده توانسته است در شرایط مساعد بازار، به طور هوشمندانه منابع را به سمت گزینه‌های با بازده بالاتر هدایت کند. کاهش ۸/۹۱ درصدی هزینه‌های عملیاتی که توالی اقدامات منتخب توسط عامل یادگیری تقویتی به شکل‌گیری ساختارهای تولید کم‌هزینه‌تری منجر شده، به‌ویژه در مقاطعی که سیاست مبتنی بر خبرگان از الگوی تخصیص کارآمدتری برخوردار نبوده است. ترکیب این دو اثر افزایش درآمد و کاهش هزینه در مجموع به افزایش ۱۰/۲۳ درصدی سودآوری کل نسبت به سیاست پایه منجر شده است. این یافته نشان می‌دهد چارچوب ترکیبی یادگیری تقویتی و تحلیل سلسله‌مراتبی نه‌تنها از منظر نظری همسو با منطق تصمیم‌گیری خبرگان عمل می‌کند، بلکه در عمل نیز قادر است ارزش اقتصادی قابل توجهی خلق کند. شایان یادآوری است که شاخص‌های اقتصادی برای هر دو سیاست (سیاست پایه فعلی تخصیص و ترکیبی یادگیری تقویتی و تحلیل سلسله مراتبی) با استفاده از یک مجموعه یکسان از داده‌های قیمت حجم و ساختارهای هزینهٔ قیمت یکسان محاسبه شده‌اند. تنها عامل تمایزدهنده بین این دو نتیجه، الگوی تخصیص تولیدشده توسط سیاست آموخته شده است.

بنابراین، افزایش سودآوری از ۴/۱۲ میلیارد دلار به ۴/۵۴ میلیارد دلار را می‌توان فقط با توالی تصمیمات ترکیبی یادگیری تقویتی و تحلیل سلسله‌مراتبی تبیین کرد و نه با تغییرات در ورودی‌های بازار یا مفروضات خارجی. این رویکرد تضمین می‌کند که بهبود گزارش‌شده، به‌مثابه عملکرد واقعی سیاست آموخته‌شده در شرایط واقعی بازار در نظر گرفته شود. این یافته‌ها نشان می‌دهند چارچوب ترکیبی یادگیری تقویتی و تحلیل سلسله‌مراتبی قادر است تعدیلات تخصیصی پایدار و منسجمی را ایجاد کند و این تعدیلات را در دورهٔ بررسی، به مزایای مالی قابل اندازه‌گیری تبدیل کند. سیاست توسعه‌یافته، کارآمدی بیشتری را در بهره‌برداری از نوسانات روزانه بازار به اثبات رسانده و در نتیجه، به سود تجمعی بهتری نسبت به مدل تخصیص قدیمی مبتنی بر خبرگان منجر شده است.

۵. بحث

یافته‌های پژوهش حاضر نشان می‌دهد چارچوب تصمیم‌گیری پیشنهادی ترکیبی یادگیری تقویتی و تحلیل سلسله‌مراتبی در تخصیص نفت خام، واجد ویژگی‌های رفتاری و ساختاری قابل توجهی است. نخست، همگرایی در الگوهای تحول سیاست نشان می‌دهد الگوریتم یادگیری تقویتی قادر است اولویت‌های تعیین‌شده توسط خبرگان را در شرایط داده‌های پرنوسان بازار با تغییرات روزانه تثبیت کند. این امر بیانگر آن است که ادغام وزن‌های تحلیل سلسله‌مراتبی در تابع پاداش، به‌مثابه یک ابزار سیاست‌گذاری مؤثر عمل می‌کند؛ به این‌معنا که فضای پاسخ‌های ممکن را به گونه‌ای مقید می‌سازد که تصمیمات آموخته‌شده با واقعیت‌های مدیریتی سازگار باشند، نه آنکه فقط به سمت بیشینه‌سازی عددی و بهینه‌سازی ریاضی حرکت کنند. در اینجا یک سازوکار اقتصادی دوگانه وجود دارد که افزایش سود از ۴/۱۲ میلیارد دلار به ۴/۵۴ میلیارد دلار را سبب می‌شود. افزایش تخصیص نفت خام به بخش‌هایی با حاشیهٔ سود تحقق‌یافته بالاتر در دوره‌های صعودی بازار و کاهش مواجههٔ عرضه با ساختارهای پرهزینهٔ تولید در دوره‌های نزولی یا رکودی بازار نتیجه این تصمیم‌گیری‌ها بوده است. این نتایج بیانگر آن است که عامل یادگیری تقویتی توانسته است درکی نسبتاً پیچیده از گذارهای روزانه و پویایی‌های درون‌دوره‌ای بازار به دست آورد، درکی که معمولاً در سامانه‌های مبتنی بر قواعد ثابت یا مدل‌های بهینه‌سازی خطی، به دلیل ماهیت ایستا یا قطعی آن‌ها، غافل می‌ماند.

از منظر روش‌شناختی، این پژوهش نشان می‌دهد در حکمرانی سیاست‌های انرژی همچنان نیاز به چارچوب‌های هوش مصنوعی وجود دارد که با ملاحظات سیاستی همسو باشند و از دانش خبرگان بهره ببرند؛ امری که در این مطالعه از طریق

ترکیب فرایند تحلیل سلسله‌مراتبی با یادگیری تطبیقی مبتنی بر یادگیری تقویتی محقق شده است. این رویکرد، شکاف میان قواعد سیاستی سنتی مبتنی بر مدل‌های خطی و قطعی و بهینه‌سازی خودمختار مبتنی بر هوش مصنوعی را پر می‌کند. افزون بر این، چنین چارچوبی به تنظیم‌گران امکان می‌دهد تا با درک بهتر فرایندهای تصمیم‌گیری درون این سامانه‌های هوشمند، منطق شکل‌گیری تصمیمات را مشاهده و تحلیل کنند؛ قابلیتی که به تداوم تکامل سیاست‌ها در چارچوب محدودیت‌های نهادی کمک می‌کند. با این‌وجود، مطالعهٔ حاضر همچنان با محدودیت‌هایی مواجه است. نخست آنکه داده‌های محیط یادگیری تقویتی فقط بر مبنای اطلاعات بورس انرژی ایران شکل گرفته و وضعیت محیط تنها از طریق تغییرات قیمت و حجم معاملات توصیف شده است. در پژوهش‌های آینده می‌توان متغیرهای دیگری را نیز در مدل لحاظ کرد؛ از جمله عوامل سیاسی، سطح موجودی‌ها، محدودیت‌های زیست‌محیطی یا ظرفیت‌های تولید.

از منظر عملیاتی، یافته‌های این پژوهش نشان می‌دهد بهره‌گیری از چارچوب تلفیقی تحلیل سلسله‌مراتبی (AHP) و یادگیری تقویتی (RL) می‌تواند به عنوان یک ابزار تصمیم‌سازی کارآمد برای تخصیص بهینهٔ نفت خام در زنجیرهٔ پالایش و پتروشیمی مورد استفاده قرار گیرد. از منظر اجرایی، به مدیران و سیاست‌گذاران توصیه می‌شود که در طراحی سیاست‌های تخصیص، فقط به شاخص‌های مالی کوتاه‌مدت اتکا نکنند، بلکه ترجیحات راهبردی، محدودیت‌های عملیاتی و اولویت‌های نهادی را نیز در قالب وزن‌های تصمیم‌گیری وارد سازوکار تخصیص کنند. نتایج این مطالعه نشان می‌دهد ادغام این ترجیحات در تابع پاداش، می‌تواند به سیاست‌هایی منجر شود که علاوه بر افزایش سودآوری، از نظر انطباق با اهداف مدیریتی نیز کارآمدتر باشند.

همچنین پیشنهاد می‌شود بورس‌های انرژی، شرکت‌های پالایش و نهادهای تنظیم‌گر بازار، از مدل‌های داده‌محور تطبیقی برای پایش مستمر شرایط بازار و بازتنظیم سیاست‌های تخصیص استفاده کنند. در چنین چارچوبی، سیستم تصمیم‌یار می‌تواند با دریافت داده‌های به‌روز از قیمت‌ها، حجم معاملات و سایر متغیرهای اثرگذار، به صورت پویا الگوی تخصیص را اصلاح کرده و از تصمیم‌گیری‌های ایستا و کم‌انعطاف جلوگیری کند. افزون بر این، به‌کارگیری این چارچوب در محیط‌های واقعی مستلزم ایجاد زیرساخت‌های داده‌ای شفاف، استانداردسازی داده‌های معاملاتی و تعامل مؤثر میان خبرگان صنعتی و تیم‌های تحلیل داده است.

علی‌رغم دستاوردهای این پژوهش در توسعهٔ یک چارچوب تلفیقی مبتنی بر تحلیل سلسله‌مراتبی و یادگیری تقویتی جهت بهینه‌سازی تخصیص نفت خام، مسیرهای گسترده‌ای برای پژوهش‌های آتی باز است. از منظر فنی، ارتقای الگوریتم‌های مورد استفاده از یادگیری تقویتی کلاسیک به روش‌های یادگیری تقویتی عمیق نظیر الگوریتم‌های PPO می‌تواند توانمندی مدل را در مدل‌سازی فضاهای حالت با ابعاد بالا و روابط غیرخطی پیچیدهٔ میان متغیرهای محیطی به طور قابل توجهی افزایش دهد. علاوه بر این، ادغام متغیرهای برون‌زا نظیر نوسانات کلان اقتصادی، ریسک‌های ژئوپلیتیکی و محدودیت‌های لجستیکی زنجیرهٔ تأمین، برای افزایش تاب‌آوری و استحکام مدل در برابر شوک‌های ناگهانی بازار، ضرورتی انکارناپذیر خواهد بود.

شایان یادآوری است که یکی از ویژگی‌های برجسته و کلیدی چارچوب پیشنهادی در این مطالعه، قابلیت تعمیم‌پذیری بالای آن به حوزه‌های دیگر است. رویکرد تلفیقی توسعه‌یافته در این پژوهش، صرف‌نظر از ماهیت تخصصی صنعت پالایش و پتروشیمی، یک ساختار تصمیم‌یار چندمنظوره محسوب می‌شود که می‌تواند در سایر صنایع فرایندی و زنجیره‌های تأمین پیچیده نیز به کار گرفته شود. به عنوان نمونه، صنایع فولاد، تولیدات دارویی، مدیریت شبکه‌های توزیع انرژی و سیستم‌های لجستیک پیشرفته که با چالش‌های مشابهی در زمینهٔ تخصیص بهینهٔ منابع و مدیریت اهداف متعارض مواجه هستند، می‌توانند از این رویکرد برای بهبود کیفیت تصمیم‌گیری در محیط‌های پویا و تصادفی بهره‌مند شوند. در راستای این هدف، پژوهش‌های آتی می‌توانند بر تلفیق تکنیک‌های هوش مصنوعی تبیین‌پذیر با فرایند یادگیری تمرکز کنند تا با شفاف‌سازی منطق تصمیمات اتخاذ شده توسط عامل هوشمند، اعتماد مدیران را به سیستم‌های خودکار در سطوح راهبردی افزایش دهند. در نهایت، انجام تحلیل‌های حساسیت جامع بر وزن‌های معیارها در مدل تحلیل سلسله‌مراتبی، می‌تواند به درک دقیق‌تر اثرگذاری ترجیحات ذهنی خبرگان بر پایداری سیاست‌های نهایی و ارتقای قابلیت اطمینان این چارچوب در کاربردهای صنعتی متنوع کمک کند.

۶. نتیجه‌گیری

این پژوهش یک مدل یادگیری تقویتی سیاست‌محور ارائه کرد که با استفاده از وزن‌های استخراج‌شده از فرایند تحلیل

سلسله‌مراتبی تقویت شده است تا بتواند در زمان واقعی، دوگانگی تصمیم‌گیری میان صنایع پتروشیمی و پالایشگاهی را در زمینهٔ تخصیص نفت خام مدیریت کند. مدل پیشنهادی، ترجیحات خبرگان را به یک تابع پاداش تبدیل می‌کند و از این طریق امکان می‌دهد که یادگیری سیاست بدون آسیب به سازگاری با اهداف مدیریتی و راهبردی صورت گیرد. اجرای سیاست ترکیبی پژوهش حاضر بر داده‌های واقعی بازار نشان‌دهندهٔ مزایای عملیاتی و مالی قابل توجه است. برنامهٔ تخصیص به‌دست‌آمده موجب افزایش ۱۰/۲۳ درصدی سودآوری تجمعی تا سطح ۴/۵۴ میلیارد دلار شده که شامل افزایش ۱/۳۲ درصدی درآمد و کاهش ۸/۹۱ درصدی هزینه‌های عملیاتی است. مسیرهای سیاست همچنین همگرایی منسجمی را نشان می‌دهند؛ به این معنا که استراتژی آموخته‌شده از نظر عملیاتی قادر است به طور مؤثر با چرخه‌های تصمیم‌گیری روزانه مواجه شود. این چارچوب، روشی عمومی برای توسعهٔ سامانه‌های هوش مصنوعی تصمیم‌محور در بازارهای انرژی ارائه می‌دهد؛ به‌ویژه از طریق ادغام مستقیم بینش خبرگان در الگوریتم‌های یادگیری تطبیقی، به جای اتکا به روش‌های سنتی مبتنی بر بهینه‌سازی ایستا. این امر شفافیت و قابلیت توضیح سیاست‌های تولیدشده را به طور محسوسی افزایش می‌دهد.

منابع

[1] Azizi V, Javadi SM, Razavi SA. The Relationship between Oil Price Uncertainty and Earnings Management in Refining and Petrochemical Companies of Tehran Stock Exchange. Information Sciences and Technological Innovations. 2025 Mar 25;2(1):35-47.

[2] Jiang G, Chen F, Gu M. Supply chain digitization and energy resilience: Evidence from China. Energy Economics. 2025 Apr 1; 144:108420.

[3] Wang T, Cai X, Xu Q. Energy market price forecasting and financial technology risk management based on generative AI. APPLIED AND COMPUTATIONAL ENGINEERING Учредители: EWA Publishing. 2025;116(1):29-34.

[4] Kallrath J, Pardalos PM, Rebennack S, Scheidt M, editors. Optimization in the energy industry. Berlin: Springer; 2009.

[5] Xu B, Fu R, Lau CK. Energy market uncertainty and the impact on the crude oil prices. Journal of Environmental Management. 2021 Nov 15; 298:113403.

[6] Lu J, Zhao R, Yu Z, Dai Y, Zeng K. RL and AHP-Based Multi-Timescale Multi-Clock Source Time Synchronization for Distribution Power Internet of Things. Computers, Materials & Continua. 2024 Mar 1;78(3).

[7] Li C, Zheng P, Yin Y, Wang B, Wang L. Deep reinforcement learning in smart manufacturing: A review and prospects. CIRP Journal of Manufacturing Science and Technology. 2023 Feb 1; 40:75-101.

[8] Aksakal E, Dağdeviren M. Analyzing reward management framework with multi criteria decision making methods. Procedia-Social and Behavioral Sciences. 2014 Aug 25; 147:147-52.

[9] He Z, Tran KP, Thomassey S, Zeng X, Xu J, Yi C. A deep reinforcement learning based multi-criteria decision support system for optimizing textile chemical process. Computers in Industry. 2021 Feb 1; 125:103373.

[10] Yu Y, Zhang N. Revealing the power of market-based energy policy: Evidence from China's energy quota trading system using machine learning. Energy Policy. 2026 Jan 1; 208:114905.

[11] Lara-Perez JI, Trejo-Caballero G, Tapia-Tinoco G, Raya-González LE, Garcia-Perez A. Deep Reinforcement Learning-Based Voltage Regulation Using Electric Springs in Active Distribution Networks. Technologies. 2026 Feb 1;14(2):87.

[12] Shoaei M, Noorollahi Y, Yousefi H. Scenario-based long-term planning for a 100% renewable energy system at the regional scale. Renewable Energy. 2026 May 24:125986.

[13] Wang X, Huang Y, Liu Y, Liang D, Zhou Y. Deep reinforcement learning-based coordinated optimization between distribution networks and microgrids towards demand response uncertainty. Energy and AI. 2026 Mar 17:100717.

[14] Aruldoss M, Lakshmi TM, Venkatesan VP. A survey on multi criteria decision making methods and its applications. American Journal of Information Systems. 2013 Dec;1(1):31-43.

[15] Aronofsky, J. S., and Williams, T. S., "Refinery scheduling by linear programming," *Operations Re search,* 1962.

[16] Braun, M. A., and Burnham, D., "Mathematical programming in process coordination," *AIChE Jou rnal,* 1990.

[17] Ribas, G., Leiras, A., and Rebello, R., "Optimization in integrated refinery–petrochemical planning using MILP," *Energy,* 2010.

[18] Shah, P., and Mishra, S., "Optimization of refinery operations with environmental constraints," *Jou rnal of Cleaner Production,* 2013.

[19] Ivanov, D., and Ray, P., "Multi-objective genetic algorithm for resilient production and environme ntal performance," *International Journal of Production Economics,* 2014.

[20] Li, H., Ma, J., and Chen, Z., "A two-stage stochastic MILP for refinery supply under demand uncert ainty," *Energy Reports,* 2020.

[21] Soni, A., Chaudhary, R., and Yadav, D., "Multi-objective MILP framework for sustainable oil-refin ery scheduling," *Computers & Chemical Engineering,* 2021.

[22] Krasnyuk, S., et al., "Integrated economic–mathematical modeling for refinery–petrochemical coordination," *Energy,* 2022.

[23] Keefer, D. L., "Risk analysis in decision making: The role of risk attitudes," *Decision Sciences,* 1991.

[24] Coopersmith, E., Kutz, J., and Voss, C., "Value-of-information in managerial decisions," *European Journal of Operational Research,* 2000.
[25] Batubara, J., et al., "Policy-based decision framework for sustainable resource development," *Energy Policy,* 2016.
[26] MacKenzie, A., "Dynamic allocation and Markov decision optimization in complex systems," *Journal of Industrial Engineering,* 2016.
[27] Sudhakar, B., "Reinforcement learning for dynamic refinery scheduling," *Expert Systems with Applications,* 2020.
[28] Kwon, J. S.-I., et al., "Decision-supporting platform for petrochemical supply chain management," *Computers & Chemical Engineering,* 2022.
[29] Raja, M., et al., "Reinforcement learning for adaptive resource coordination," *Applied Energy,* 2023.
[30] Odimarha, A., et al., "Process control through self-learning adaptive agents," *Energy AI,* 2024.
[31] Cui, T., and Yao, L., "Hybrid neuro-symbolic automation in industrial decision systems," *Applied Energy,* 2024.
[32] Ma, L., Zhong, R., and Hu, C., "Adaptive coordination between refinery and petrochemical units via a deep RL," *Energy Reports,* 2023.
[33] Lee, C.-Y., Chou, B.-J., and Huang, C.-F., "Data-science and reinforcement learning for raw-material procurement in petrochemical industry," *Advanced Engineering Informatics,* 2022.
[34] Cui, T., Zhou, J., and Liu, Y., "Deep reinforcement learning for multi-criteria resource control," *Applied Energy,* 2024.
[35] Wang, Y., and Zhang, S., "Analytic hierarchy process in renewable energy policy evaluation," *Renewable Energy,* 2020.
[36] Tatar, O., and Kara, M., "AI-supported policy learning in governance models," *Energy Economics,* 2024.
[37] Bhol, S. K., and Reddy, P. V., "Multi-criteria AI design for sustainable decisions," *Expert Systems with Applications,* 2024.