Islamic Azad University , Shiraz Branch
نشریه تحلیل مدارها، داده ها و سامانه ها
Journal of Circuits, Data and Systems Analysis
sanad.iau.ir/journal/jcdsa

A survey on intrinsic motivation models in machine learning

Saeed Jamali[1], Saeed Setayeshi[2*], Mohsen Jahanshahi[3], Sajjad Taghvaei[4]

[1] Department of Computer Engineering, Central Tehran Branch, Islamic Azad University, Tehran, Iran
Saeed.Jamali.eng@iauctb.ac.ir
[2] Department of Medical Radiation Engineering, Amirkabir University of Technology, Tehran, Iran.
Setayesh@aut.ac.ir
[3] Department of Computer Engineering, Central Tehran Branch, Islamic Azad University, Tehran, Iran
mjahanshahi@iauctb.ac.ir
4 Department of Mechanical Engineering, Shiraz University, Shiraz, Iran
sj.Taghvaei@shirazu.ac.ir

Abstract: Intrinsic motivation has garnered significant attention in recent years, empowering both living beings and robots to learn autonomously and cumulatively, even without extrinsic motivation (rewards from the environment). This concept, drawing inspiration from psychology and neuroscience, has opened up new avenues in artificial intelligence. Algorithmic architectures for intrinsic motivation facilitate exploration and the effective acquisition of motor skills in scenarios where environment rewards are sparse or absent. This is particularly relevant for many real-world problems, where large portions of the environment offer no explicit rewards. Consequently, intrinsic motivation holds not only theoretical significance for enhancing artificial intelligence algorithms, particularly in exploration tasks, but also practical implications for real-world or near-real-world applications. In this paper, we delve into the significance of intrinsic motivation, providing a brief overview of its origins in psychology. We then systematically categorize and examine research on intrinsic motivation in artificial intelligence. Additionally, we discuss the reinforcement learning method as a successful approach for incorporating intrinsic motivation. Finally, we explore the practical applications, limitations, and future intrinsic motivation research.

Keywords: Cognitive science, Intrinsic motivation, Machine learning, Reinforcement learning, Sparse environment, Search

JCDSA, Vol. 2, No. 4, Winter 2025	Online ISSN: 2981-1295	Journal Homepage: https://sanad.iau.ir/en/Journal/jcdsa
Received: 2024-07-02	Accepted: 2024-12-14	Published: 2025-03-20

CITATION

Jamali, S., Setayeshi, S., Jahanshahi, M., Taghvaei, S., " A survey on intrinsic motivation models in machine learning", Journal of Circuits, Data and Systems Analysis (JCDSA), Vol. 2, No. 4, pp. 24-53, 2025.
DOI: 00.00000/0000

* Corresponding author

Extended Abstract

1- Introduction

Intrinsic motivation is a foundational concept in the realm of artificial intelligenc, inspired by human cognition and behavioral sciences. Unlike extrinsic motivation, which relies on explicit rewards from the environment, intrinsic motivation propels agents to engage in exploration and learning for the sake of curiosity and novelty. This quality is particularly crucial in environments where external rewards are sparse or non-existent, challenging traditional artificial intelligence models to learn effectively. Recent developments have incorporated intrinsic motivation into machine learning, aiming to enhance adaptability and autonomous problem-solving. This paper aims to delve into the role of intrinsic motivation within artificial intelligence, tracing its origins, evaluating existing models, and discussing its practical implications. Furthermore, intrinsic motivation facilitates continuous learning and adaptation, allowing agents to refine their decision-making capabilities autonomously. This approach is increasingly being explored in unsupervised and reinforcement learning scenarios to stimulate exploratory behavior that leads to better generalization and task performance. The integration of such models has potential applications across a range of fields, including robotics, complex system navigation, and cognitive modeling.

Of course, examining the position of intrinsic motivation in psychology and intrinsic motivation in reinforcement learning is also among the objectives of this review paper. The structure of this paper is as follows: In Section 2, we show that motivation is a response to the question of why behavior occurs. Section 3 discusses the different types of motivation. Next, the position of intrinsic motivation in psychology and various related theories are reviewed in Section 4. In Section 5, the diverse research in this field is categorized, and each study is analyzed and evaluated. Given the importance and effective performance of reinforcement learning algorithms in intrinsic motivation, Section 6 introduces this method and the intrinsic motivation in reinforcement learning algorithm. Section 7 examines the limitations, applications, and future work in this domain. Finally, the conclusion is presented in Section 8.

2- Methodology

The methodology section provides an in-depth examination of how intrinsic motivation is modeled and implemented in artificial intelligence systems. This involves categorizing existing algorithmic frameworks that simulate intrinsic motivation through predictive models, curiosity-driven mechanisms, and novelty-based exploration strategies. Key methodologies include the use of self-supervised learning techniques, which enable agents to build internal representations of their environment and use prediction errors as intrinsic rewards. The research also reviews various computational approaches such as reinforcement learning algorithms embedded with intrinsic motivation. These algorithms use auxiliary tasks and prediction models to drive learning, even in the absence of external feedback. The evaluation focuses on the impact of these models on learning efficiency, task performance, and scalability. Theoretical concepts are combined with empirical results to illustrate how these models contribute to more effective exploratory strategies.

3- Results and discussion

Our findings indicate that intrinsic motivation significantly enhances exploratory learning, particularly in tasks where external rewards are infrequent or sparse. Through experiments involving various intrinsic motivation frameworks, we observed marked improvements in task performance and learning rates when compared to models driven solely by extrinsic rewards. For instance, models utilizing novelty detection and prediction error as intrinsic signals demonstrated superior adaptation capabilities in complex and unfamiliar environments.

The discussion extends to compare intrinsic and extrinsic motivation models, noting that while intrinsic motivation fosters better initial exploration, it may require balanced integration with external incentives for optimal task completion. Challenges such as computational overhead and fine-tuning for real-world applications are also addressed. The results underscore the practical implications of applying intrinsic motivation in artificial intelligence, particularly in fields that require adaptability and independent learning, such as autonomous robotics and cognitive simulations.

4- Conclusion

The integration of intrinsic motivation into artificial intelligence systems has proven to be a powerful tool for fostering exploration and learning in environments where explicit guidance is minimal. This paper has highlighted the key methodologies and impacts of intrinsic motivation in enhancing machine learning models. While promising, the approach still faces challenges, including the computational cost of training and ensuring alignment with specific task goals. Future research should aim to develop more efficient and scalable models, as well as explore how intrinsic and extrinsic motivations can be effectively balanced to enhance performance in real-world scenarios.

The potential applications of intrinsic motivation span from robotics to interactive systems, signaling a promising frontier for adaptive and human-like learning processes in artificial intelligence. By addressing current limitations and focusing on advancements in algorithm design and computational efficiency, intrinsic motivation can continue to shape the evolution of artificial intelligence, making it more robust, adaptable, and capable of learning autonomously.

Islamic Azad University , Shiraz Branch
نشریه تحلیل مدارها، داده ها و سامانه ها
Journal of Circuits, Data and Systems Analysis
sanad.iau.ir/journal/jcdsa

مروری بر مدل‌های انگیزش درونی یادگیری ماشین

سعید جمالی[۱]، سعید ستایشی[۲*]، محسن جهانشاهی[۳]، سجاد تقوایی[۴]

۱- گروه مهندسی کامپیوتر و فناوری اطلاعات، دانشکده فنی و مهندسی، واحد تهران مرکزی، دانشگاه آزاد اسلامی، تهران، ایران (Saeed.Jamali.eng@iauctb.ac.ir)

۲- دانشکده مهندسی فیزیک و انرژی، دانشگاه صنعتی امیرکبیر، تهران (Setayesh@aut.ac.ir)

۳- گروه مهندسی کامپیوتر و فناوری اطلاعات، دانشکده فنی و مهندسی، واحد تهران مرکزی، دانشگاه آزاد اسلامی، تهران، ایران (mJahanshahi@iauctb.ac.ir)

۴- دانشکده مهندسی مکانیک، دانشگاه شیراز، شیراز، ایران (sj.Taghvaei@shirazu.ac.ir)

چکیده: امروزه انگیزش، توجه فزاینده‌ای را به خود جلب کرده است؛ زیرا که به موجود زنده و ربات اجازه می‌دهد در فقدان راهنماییِ انگیزش بیرونی (پاداش از محیط)، دانش و مهارت را به‌صورت تجمعی و کاملاً خودمختار اخذ نماید. علاوه بر روان‌شناسی و علوم اعصاب، این مدل‌ها چشم‌انداز جدیدی را در هوش مصنوعی ایجاد کرده‌اند. معماری‌های الگوریتمی انگیزش درونی در اکتشاف، امکان اخذ مهارت‌های حرکتیِ مؤثری را در مسائل به وجود می‌آورند. از طرفی بسیاری از مسائل دنیای واقعی دارای چنین خصوصیتی هستند که در بخش‌های زیادی از محیط، پاداشی از محیط وجود ندارد. در نتیجه، این موضوع نه‌تنها از لحاظ نظری اهمیت بسزایی در بهبود الگوریتم‌های هوش مصنوعی مخصوصاً در بحث اکتشاف دارد، بلکه به لحاظ کاربردی و عملی نیز می‌تواند در کاربردهای واقعی و یا نزدیک به واقعیت مورداستفاده قرار گیرد. این مقاله به اهمیت انگیزش درونی پرداخته و نگاه کوتاهی به جایگاه آن در روان‌شناسی دارد. سپس تحقیقات پیرامون انگیزش درونی در هوش مصنوعی دسته‌بندی‌شده و مورد بررسی قرار گرفته‌اند. همچنین، روش یادگیری تقویتی به‌عنوان رویکردی موفق در ترکیب انگیزش درونی بررسی شده است. در نهایت، به برخی از کاربردهای عملی انگیزش درونی، محدودیت‌ها و تحقیقات آینده اشاره شده است.

واژه‌های کلیدی: علوم‌شناختی، انگیزش درونی، یادگیری ماشین، یادگیری تقویتی، محیط‌های خلوت (اسپارس)، جستجو

نوع مقاله: مروری

DOI: 00.00000/0000

تاریخ ارسال مقاله: ۱۴۰۳/۰۴/۱۳ تاریخ پذیرش مقاله: ۱۴۰۳/۰۹/۲۴ تاریخ چاپ مقاله: ۱۴۰۳/۱۲/۳۰

۱- مقدمه

شناخت به‌وسیله یک موجود زنده عموماً به توانمندی آن برای پردازش اطلاعات ادراکی و متعاقب آن دست‌کاری رفتارش معنا می‌یابد. شناخت انسان شامل یک مجموعه‌ی بزرگ از فرایندهای موجود در ذهن انسان است. توانایی شناختی انسان‌ها به‌طورکلی در خودآگاهی، ادراک، یادگیری، دانش، استدلال، برنامه‌ریزی و تصمیم‌گیری ظاهر می‌شود [۱]. در کنار هم قراردادن همه این توانمندی‌ها در یک عامل مصنوعی یا یک ربات یکی از بزرگ‌ترین چالش‌های عصر حاضر است. رویکردهای مختلف به این موضوع، تحت عنوان «رشد ذهنی خودمختار» معرفی شده‌اند که حوزه‌های تحقیقاتی آن دو هدف اصلی را دنبال می‌کنند: الف) بهبود توانایی سیستم‌های مصنوعی ب) پیشرفت درک ما از ریشه‌های اساسی هوش در سیستم‌های طبیعی [۲]. منظور از هوش، توسعه خودمختار و یادگیری مادام‌العمر و بی‌پایان است [۳]. در نتیجه رباتیک توسعه‌ای (رشدی) حوزه‌ی علمی است که به مطالعه مکانیزم‌ها، معماری و محدودیت‌هایی می‌پردازد که اجازه می‌دهد یادگیری مهارت‌های جدید و دانش در ماشین‌های مجسم به‌صورت مادام‌العمر و بی‌پایان صورت گیرد [۴،۵]. این موضوع باعث می‌شود تا قابلیت‌های بهبود ادراکی، استدلال، برنامه‌ریزی، اجتماعی، استخراج دانش و مهارت‌های تصمیم‌گیری در ماشین‌های مجسم به وجود آید [۶]. علم پایه‌ای برای رباتیک شناختی توسعه‌ای از درک و دریافت ما از فرایند رشد کودک نشئت می‌گیرد. این رشد با توسعه نگاشت حسی - حرکتی جنینی در رحم آغاز شده و از طریق نمایش بدن، توسعه مهارت‌های حرکتی و ادراک فضایی به سمت توسعه رفتارهای اجتماعی پیشرفت کرده است. در نتیجه، یک برنامه توسعه‌ای تعاملات بلادرنگ را با محیط داخلی و خارجی خود به‌وسیله استفاده از حسگرها و اثرگذارهایش برقرار می‌کند. در کودکان، پیشرفتِ تجمعی و تدریجی باعث افزایش پیچیدگی می‌شود [۴]. این یادگیری خود راهبر، ریشه در انگیزش درونی[۲] (ذاتی، اندرونی) دارد. انگیزش

* نویسنده مسئول

[2] Intrinsic motivation (IM)

درونی به‌عنوان یک مکانیسم اساسی به‌گونه‌ای عمل می‌کند که یادگیری و اکتشاف انسان در محیط پیرامونش را هدایت کند. هدف اصلی این مقاله، مروری بر تحقیقات انگیزش درونی در حوزه هوش مصنوعی[1] است.

باتوجه به تنوع و حجم بالای کارهای تحقیقاتی، دسته‌بندی پیشنهادی برای انواع این تحقیقات نیز ارائه شده و سپس به بررسی ایده اصلی، مزایا و معایب و الگوریتم پایه مورداستفاده آن‌ها پرداخته شده است. البته بررسی جایگاه انگیزش درونی در روان‌شناسی و یادگیری تقویتی انگیزش درونی[2] نیز از سایر اهداف این مقاله مروری است. ساختار این مقاله بدین صورت است که: در قسمت ۲، نشان می‌دهیم که انگیزش پاسخی به چرایی رفتار است. قسمت ۳، به انواع انگیزش می‌پردازد. سپس جایگاه انگیزش درونی در روان‌شناسی و نظریه‌های مختلف پیرامون آن در قسمت ۴ مورد بررسی قرار می‌گیرد. در قسمت ۵ تحقیقات بسیار متنوع این حوزه دسته‌بندی شده و هر تحقیق مورد بررسی و ارزیابی قرار گرفته است. با توجه به اهمیت و عملکرد مناسب الگوریتم یادگیری تقویتی[3] در انگیزش درونی، قسمت ۶ به معرفی این روش و الگوریتم یادگیری تقویتی انگیزش درونی می‌پردازد. قسمت ۷ محدودیت‌ها، کاربردها و کارهای آینده در این حوزه را بررسی می‌کند. نهایتاً نتیجه‌گیری در قسمت ۸ ارائه شده است.

۲- رفتار و انگیزش

فصل مشترک و زیربنای علوم‌شناختی،هوش مصنوعی و رباتیک را می‌توان در بررسی چرایی به‌وجود آمدن رفتار خلاصه نمود. پاسخ به دو سؤال کلیدی می‌تواند در درک ما از رفتار بسیار کمک‌کننده باشد. چه چیزی موجب رفتار می‌شود؟ و چرا شدت رفتار تغییر می‌کند؟ [۷] پاسخ این سؤالات در انگیزش نهفته است. انگیزش یک ساختار روانی یکپارچه مبتنی بر سیستم پیچیده است که فرد را به انجام انواع فعالیت‌ها ترغیب می‌کند و به شیوه خاصی بر رفتار انسان تأثیر می‌گذارد [۲۷]. انگیزش، در طول زمان نوسان داشته و با تغییر آن، رفتار نیز تغییر می‌کند [۷]. انگیزش حالت مجزا یا ثابتی نیست، بلکه فرایندی پویا بوده و همیشه تغییر می‌کند. شماری از نظریه‌پردازان انگیزش معتقدند که انگیزش انواع مختلفی دارد [۸]. برای مثال انگیزش بیرونی[4] با انگیزش درونی تفاوت دارد [۹]. نظریه‌های مختلف و متنوعی پیرامون انگیزش ارائه شده است، از جمله آن‌ها می‌توان به غریزه [۱۰]، سایق[5] [۱۱،۱۲]، انگیزش پیشرفت [۱۳]، ناهمگونی شناختی[6][۱۴]، انگیزش کارایی [۱۵]، غوطه‌وری (جریان)[7] [۱۶]، تعیین هدف [۱۷]، درماندگی آموخته شده[8] [۱۸]، احساس کارایی [۱۹] و طرحواره خویشتن[9] [۲۰] اشاره کرد.

نیازها، شناخت‌ها، و هیجان‌ها تجربیات درونی‌ای هستند که انگیزه می‌تواند از آن‌ها نشئت بگیرد. انگیزش رفتار را نیرومند و هدایت می‌کند. از طرفی رویدادهای بیرونی نیز چنین توانایی را دارند. آن‌ها در نقش مشوق‌های محیطی‌ای هستند که باعث می‌شوند فرد به انجام عمل خاصی روی آورده یا از آن اجتناب کند. مشوق‌ها قبل از رفتار واقع می‌شوند. آن‌ها عاملی هستند که فرد را به سمت رویدادهای بیرونی که حاصل آن تجربیات خوشایند است، می‌کشانند یا او را از رویدادهای بیرونی که باعث تجربیات ناخوشایند می‌شوند، دور می‌کنند [۷].

۳- انگیزش بیرونی و درونی

تجربه به ما یاد داده است که دو راه برای لذت‌بردن از یک فعالیت وجود دارد: به‌صورت درونی یا بیرونی. در واقع، هر فعالیتی را می‌توان با انگیزش درونی یا انگیزش بیرونی یا ترکیبی از هر دو انجام داد [۲۱]. انگیزش بیرونی از علت محیطی برای آغازکردن یک عمل ناشی می‌شود. انگیزش بیرونی برای هدایتِ یادگیریِ رفتارهایی استفاده می‌شود که مرتبط با نیازهای اولیه موجود زنده باشد. منظور از نیازهای اولیه، نیازهایی است که در ارتباط با بقا و تولیدمثل است [۲۲]. رویدادهای محیطی تااندازه‌ای انگیزش بیرونی به وجود می‌آورند که وابستگی «وسیله برای هدف» را در ذهن فرد ایجاد کنند که به‌موجب آن، وسیله، رفتار است و هدف پیامدی جالب است. بررسی انگیزش بیرونی بر سه مفهوم مهم مشوق‌ها، تقویت‌کننده‌ها (مثبت و منفی) و تنبیه‌کننده‌ها استوار است [۷]. از طرفی انگیزش درونی، گرایش فطری پرداختن به تمایلات و به‌کاربردن توانایی‌ها و جستجوکردن چالش‌های بهینه و تسلط یافتن بر آن‌ها در انجام کار است [۲۳]. انگیزش درونی به طور خودانگیخته از نیازهای روان‌شناختی، کنجکاوی و تلاش‌های فطری برای رشد، حاصل می‌شود. البته باید اشاره نمود، تمایزی سخت و سریع بین سیگنال پاداش درونی[10] و بیرونی وجود ندارد [۲۲]. انگیزش درونی از احساس شایسته بودن و خودمختار بودن در حین انجام‌دادن یک فعالیت، حاصل می‌شود.

۴- انگیزش درونی در روان‌شناسی

ظرفیت یادگیری تجمعی خودمختاری که به‌وسیله ارگانیسم‌هایی همچون پستانداران به‌ویژه انسان‌ها نشان داده می‌شود، بسیار حیرت‌انگیز است. ریشه این ظرفیت در انگیزش درونی قرار دارد. انگیزشی که مستقیماً به پاداش‌های بیرونی از قبیل غذا یا روابط جنسی ارتباط مستقیمی ندارد، اما می‌توان گفت به چیزی که موجودات می‌دانند (کنجکاوی، تازگی، شگفتی) یا می‌توانند انجام دهند (شایستگی، توان) مرتبط است. روان‌شناسان برای انسان‌ها و حیوانات مدارکی یافتند که انگیزش درونی نقش مهمی در رفتار و یادگیری آن‌ها بازی می‌کند [۲۴]. قبل از بررسی تحقیقات اصلی انگیزش درونی در حوزه روان‌شناسی باید به دو نکته مهم توجه کرد: ابتدا اینکه هر محقق یا تیم‌های تحقیقاتی از زاویه دید خود به بررسی این حوزه پرداخته و از جنبه‌های مختلفی بدان پرداخته‌اند، لذا ما با تنوع دیدگاه و دسته‌بندی مواجه هستیم که

[1] Artificial intelligence (AI)
[2] Intrinsically motivated reinforcement learning (IMRL)
[3] Reinforcement learning (RL)
[4] Extrinsic motivation (EM)
[5] Drive
[6] Cognitive dissonance
[7] Flow
[8] Learned helplessness
[9] Self-schemas
[10] Intrinsic reward (IR)

طبقه‌بندی این پژوهش‌ها خود از موضوعات باز تحقیقاتی است. نکته دیگر اینکه در کنار بررسی این حوزه، به زاویه دید محققان حوزه هوش مصنوعی توجه شده است، چرا که این دیدگاه‌ها پایه و اساس مدل‌های پیشنهادی محاسباتی خواهد بود که در بخش بعدی مورد کنکاش قرار می‌گیرد. بر اساس دیدگاه دسی[1] و ریان[2] انگیزش درونی به معنای انجام فعالیتی برای رضایت ذاتی است و نه نتایجی جدا [۹]. وقتی فردی به‌صورت درونی انگیزه‌مند می‌شود، اعمالی را برای سرگرم‌شدن یا غلبه بر چالش انجام می‌دهد و نه برای به‌دست‌آوردن محصولات، فشارها و پاداش خارجی. این موضوع در کودکان با تلاش برای گرفتن، پرتاب‌کردن، گازگرفتن و له کردن اشیای جدید به‌وضوح قابل‌مشاهده است. حتی در بزرگسالان با تلاش برای حل جدول، خواندن رمان، دیدن فیلم یا نقاشی کشیدن نیز کاملاً مشهود است [۲۵]. برای درک بهتر این موضوع باید تفاوت انگیزش درونی را با انگیزش بیرونی موردتوجه قرار داد. انگیزش بیرونی ساختاری است که هر زمان فعالیتی انجام می‌شود به‌منظور دستیابی به نتایجی جداگانه صورت می‌گیرد اما انگیزش درونی اشاره به انجام فعالیت به صورتی ساده برای لذت‌بردن از فعالیت خود، به‌جای ارزش‌های ابزاری، دارد [۹]. روان‌شناسان تلاش کرده‌اند درباره اینکه چه ویژگی‌هایی از فعالیت‌ها برای برخی مردم (نه همه مردم) در برخی زمان‌ها، انگیزش درونی تولید می‌کند، تئوری‌هایی ارائه دهند. آن‌ها مطالعاتی درباره چگونگی پیاده‌سازی عملکردی در یک ارگانیسم و به طور خاص انسان‌ها انجام داده‌اند. در جدول (۱) مهم‌ترین دیدگاه‌ها در مورد انگیزش درونی به همراه توضیحات مختصری آورده شده است.

جدول (۱): برخی از نظریه‌های مختلف انگیزش درونی

دیدگاه	مفهوم	توضیحات
سایق‌هایی برای دست‌کاری و اکتشاف	برای تطبیق نظریه هال[3] بر پدیده‌هایی از قبیل حل پازل مکانیکی توسط میمون‌های ریزوس[4] بدون دریافت پاداش، این مفاهیم ایجاد شد [۲۵].	. نظریه هال: رفتار حیوانات به‌وسیله سایق‌هایی (مانند تشنگی و گرسنگی) تحریک می‌شود که به‌عنوان نقص فیزیولوژیکی موقتی محسوب شده و ارگانیسم به‌منظور دستیابی به تعادل حیاتی، برای کاهش آن تلاش می‌کند [۲۵].
کاهش ناهمگونی شناختی	ارگانیسم‌ها برای کاهش ناهمگونی شناختی انگیزه‌مند می‌شوند [۲۵].	. ناهمگونی بین ساختارهای شناختی داخلی و شرایطی که در حال حاضر دریافت می‌کند. . جستجوی شکل‌هایی از بهینگی بین شرایط کاملاً نامعلوم و شرایط کاملاً معلوم [۲۵].
ناسازگاری بهینه	انسان در جستجوی ناسازگاری بهینه است [۲۵].	. محرک‌هایی جالب‌توجه هستند که بین سطح درک شده و سطح استاندارد اختلافی وجود داشته باشد [۲۵] (پرپاداش‌ترین شرایط: وجود سطح متوسطی از تازگی).
شایستگی، خودمختاری، ارتباط	بر اساس نظریه خود تصمیم‌گیری: سه نیاز مرکزی انسان‌ها شامل: احساس‌های شایستگی، خودمختاری و ارتباطات اجتماعی [۲۱] است.	. خودمختار: تمایلات، ترجیحات، خواسته‌ها و فرایند تصمیم‌گیری را برای انجام‌دادن یا انجام‌ندادن فعالیتی خاص، هدایت کند. . شایستگی: نیازمند مؤثربودن در تعامل با محیط که بیانگر میل به‌کاربردن استعدادها و مهارت‌ها بوده و دنبال حل‌کردن چالش‌های بهینه و تسلط یافتن بر آن‌ها است. . ارتباط: نیاز به برقراری پیوندها و دلبستگی‌های عاطفی با دیگران است [۷].
کنجکاوی، تازگی و علاقه	انسان‌ها رفتار را بر اساس کنجکاوی که احساس علاقه را از طریق ارزیابی تازگی - پیچیدگی و قابلیت فهم استنباط شده تنظیم می‌کند، انجام می‌دهند [۲۶].	. ایجاد علاقه‌مندی: موضوع تازه و/یا به‌اندازه کافی پیچیده تا باعث کنجکاوی شود، اما نه اینکه به طور غیرقابل‌فهم تازه و پیچیده باشد [۲۶]. . کنجکاوی در زمینه‌های متنوعی مورداستفاده قرار می‌گیرد درحالی‌که علاقه، مشغول‌شدن به یک حوزه خاص است [۲۹].
غوطه‌وری	مردم به فعالیت‌های چالش‌برانگیزی جذب می‌شوند که نه خیلی آسان و نه خیلی سخت باشند [۲۸].	. غوطه‌وری: شامل زیرمؤلفه‌های کشف، نقش‌گذاری[5]، شخصی‌سازی و گریزگری[6] [۲۶] . اثرات غوطه‌وری می‌تواند دارای سطوح مختلفی باشد [۳۰].
سلطه	انگیزش، برداری پنج‌بعدی است: دستاورد، اکتشاف، معاشرت‌پذیری، غوطه‌وری و سلطه [۲۶].	. تنها مؤلفه جدید نسبت به نظریه خود تصمیم‌گیری، سلطه است [۲۶]. . همپوشانی جزئی با خودمختاریِ نظریه خود تصمیم‌گیری (خودمختاری: تحت تسلط دیگران نبودن اما نه به معنی نفوذ به دیگران) [۲۶]
دانش، شایستگی	سیستم‌های IM مبتنی بر دانش (ظرفیت پیش‌بینی) یا مبتنی بر شایستگی (ظرفیت انجام کار) هستند.	. مبتنی بر دانش: به‌کارگیری معیارهایی مرتبط با پتانسیل سیستم و مدل‌سازی محیط . مبتنی بر شایستگی: بهره‌برداری از معیارهایی مرتبط با توانایی سیستم در داشتن اثرات خاص بر روی محیط [۳۱].
تنظیم هدف	تلاش افراد برای رسیدن به اهداف: متمرکزکردن توجه افراد بر ناهمخوانیِ سطحِ موجودِ دستاورد، با سطحِ ایده‌آلِ دستاوردِ او [۳۴].	. هدف: هر چیزی که فرد سعی دارد به آن برسد [۱۷]. . هدف‌ها همیشه عملکرد را افزایش نمی‌دهند، فقط هدف‌های دشوار و روشن عملکرد را افزایش می‌دهند [۷]. اهداف آسان و مبهم، انگیزش تولید نمی‌کنند [۳۴].

[1] Deci
[2] Ryan
[3] Hull
[4] Rhesus monkey
[5] Role-playing
[6] Escapism

۵- انگیزش درونی در هوش مصنوعی

خاستگاه مفهوم انگیزش درونی، علوم روان‌شناسی [۹،۲۴] و نوروفیزیولوژی [۳۶] است که به‌خوبی توانسته است به حوزه‌های یادگیری ماشین [۳۷،۳۸] و رباتیک شناختی (یا به طور خاص رباتیک رشدی) [۳۹] راه پیدا کند تا رشد خودمختار عامل‌های هوشمند و یادگیری بی‌پایان مجموعه‌ای از مهارت‌ها را بهبود بخشد. همانند عامل-های زیستی، استخراج خودمختار شایستگی‌های جدید یک گام مهم به‌سوی دستیابی به ربات‌هایی با تطبیق‌پذیری بیشتر است. از سوی دیگر، یکی از ویژگی‌های فوق‌العاده‌ی هوش طبیعی در کودکان، ظرفیت یادگیری خودمختار (مستقل) در مدت‌زمان عمر است. با بررسی‌کردن وضعیت کودکان در عمل، با این واقعیت مواجه می‌شویم که: عملکرد خودبه‌خودی آن‌ها در کشف محیط و تحلیل عملکرد، مهندسی و همراه با آنالیزهای پیچیده نیست [۴۰]. کودکان کنترل تمام حرکات بدن خود را پیکسل به پیکسل یاد نمی‌گیرند؛ بلکه با هم‌کوشی [1] ماهیچه‌ها به دنیا می‌آیند که این هم‌کوشی به‌صورت عصبی در سیستم دینامیکی تعبیه شده است همانند CPG ها[2]. جستجوی آن‌ها به‌صورت خودانگیخته (خودبه‌خود) است. مدل‌های محاسباتی برای درک بهتر اکتشاف در کودکان بر سه نوع مکانیزم تمرکز کرده‌اند: الف) مجسم (بدن‌مندی): هم‌کوشی بدن که ساختاری را بر روی عمل و ادراک فراهم می‌کند. ب) انگیزش درونی: خودسازماندهی برنامه‌ای برای اکتشاف و یادگیری در سطوح چندگانه انتزاع ج) هدایت (راهنمایی) اجتماعی: بهره‌مندی از آنچه دیگران از قبل می‌دانند [۴۱]. مهم‌ترین جزء جستجوی خودکار کودکان، سیستم انگیزش درونی است که سیستم اکتشاف کنجکاوی محور نیز نامیده می‌شود. یکی از ایده‌های کلیدی که به کمک آن می‌توان اکتشاف خودکار کودکان را شبیه‌سازی و پیش‌بینی کرد، این است که کودکان تجربیاتی را انتخاب می‌کنند که IR را بر اساس پیشرفت یادگیری تجربی، بیشینه کند. به‌صورت خودکار مسیر رشدی (توسعه‌ای) را در جایی که وظایف تدریجاً پیچیده انجام می‌شوند، تولید می‌کند [۴۰]. نقش انگیزش درونی، فراهم‌کردن راهنمای کیفی برای اکتشاف است [۲۲].

به طور خلاصه می‌توان گفت، اکتشاف آنچه شما را غافلگیر می‌کند. علی‌رغم دشواری هماهنگی سیگنال‌های IR و پاداش بیرونی با محیط-های داخلی و خارجی موجود زنده، محیط داخلی می‌تواند نقش بزرگ‌تر و یا حداقل نقش متفاوتی در تولید سیگنال‌های پاداش مرتبط با انگیزش درونی را بازی کند [۲۲]. به‌عنوان مثال، یک محرک برجسته‌ی خارجی ممکن است پاداشی را به میزان غیرمنتظره‌ای تولید کند، درحالی‌که انتظارات به وسیله‌ی فرایندهایی در محیط داخلی و اطلاعات ذخیره شده آن، ارزیابی می‌شوند. تازگی، غافلگیری، ناسازگاری و سایر ویژگی‌هایی که بر پایه انگیزش درونی فرض شده‌اند، همگی بستگی به چیزی که تاکنون عامل یاد گرفته و تجربه کرده است، دارند یعنی، خاطرات، باورها و وضعیت دانش داخلی که همه آن‌ها اجزای وضعیت محیط داخلی موجود زنده هستند. در نتیجه باید اشاره کرد، تمایز واضحی بین انواع سیگنال‌های پاداش نمی‌باشد: به‌جای آن، بازه‌ای پیوستار از کاملاً بیرونی به کاملاً درونی وجود دارد [۲۲]. محققان حوزه هوش مصنوعی و رباتیک تلاش کرده‌اند تا نظریه‌های معمول در روان‌شناسی را در قالب مدل‌های محاسباتی به حوزه خود وارد کرده و از آن‌ها بهره ببرند.

مدل‌های محاسباتی زیادی تاکنون ارائه شده‌اند که پژوهشگران سعی کرده‌اند تا از جنبه‌های مختلفی این مدل‌ها را دسته‌بندی نمایند. در این مقاله، هفت دسته مختلف برای طبقه‌بندی تحقیقات پیشنهاد شده است که عبارت‌اند از: ۱) مدل‌های مبتنی بر دانش ۲) مدل‌های مبتنی بر شایستگی ۳) تازگی (نوآوری) ۴) توانمندسازی، قدرت و کنترل ۵) تنهایی، وابستگی اجتماعی ۶) هدف ۷) غافلگیری، کنجکاوی، (عدم) قطعیت. البته باید خاطرنشان کرد که این دسته‌بندی‌ها از لحاظ مفهومی، همپوشانی‌هایی با هم دارند و نمی‌شود مرز بسیار دقیق بین آن‌ها کشید. بعلاوه یک کار تحقیقاتی می‌تواند به بیش از یک دسته اختصاص داده شود، منتها در این مقاله مروری سعی شده است که هر تحقیق در دسته-ای قرار گیرد که بیشترین شباهت را به آن داشته و علاوه‌برآن، بیان و دیدگاه و تلقی نویسندگان مقاله به آن دسته‌بندی انگیزش درونی خاص نزدیک‌تر باشد. باید توجه داشت توسعه انگیزش درونی در هوش مصنوعی از طریق چندین نقطه عطف مهم پیشرفت کرده است. در شکل (۱) خلاصه‌ای از تحقیقات کلیدی و پیشرفت‌های این حوزه، همراه با منابع، ارائه شده است.

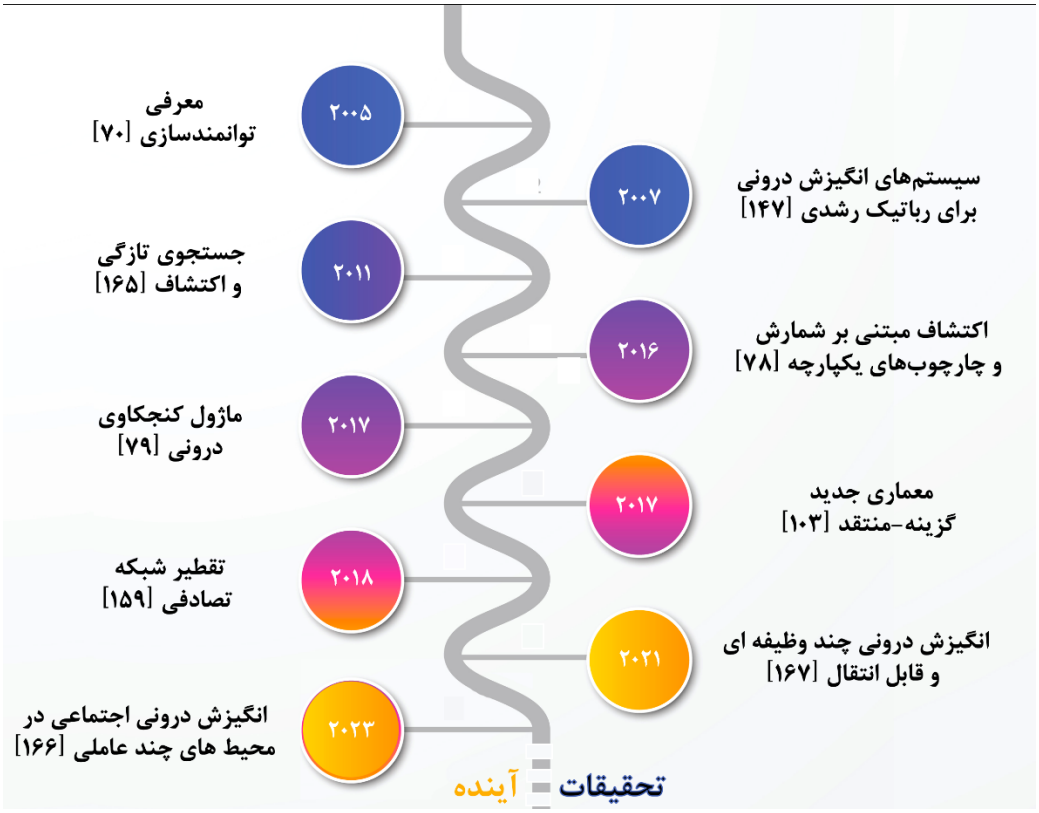
جدول ۲: مدل‌های محاسباتی IM مبتنی بر دانش

سال انتشار و مرجع	ایده اصلی	مزایا	معایب/چالش‌ها	روش یا الگوریتم
۲۰۱۸ [۵۱]	ارائه سیستمی که ترکیبی از تکنیک‌های برنامه‌نویسی منطقی قیاسی با اکتشاف خودکار مفاهیم و RL با IM به‌منظور کشف مفاهیم، وضعیت‌ها و اقدامات جدید و سیاست رفتاری برای حل مسائل	کشف خودکار مفاهیم، یادگیری رفتارها با استفاده از مفاهیم کشف شده، امکان تغییر توصیف حالت در حین یادگیری، سیستم‌های هوش مصنوعی انعطاف‌پذیرتر و توانمندتر با	احتمال کشف مفاهیم اضافی و غیرضروری، کندی فرایند یادگیری با یادگیری مفاهیم و IM، نیاز به طراحی روش‌های کارآمد برای فیلترکردن مفاهیم غیرمرتبط	ترکیب تکنیک‌های برنامه‌نویسی منطقی قیاسی با اختراع مسند و RL با IM

[1] Synergy
[2] Central pattern generators

		هدایت فرایند یادگیری توسط خود عامل		
۲۰۱۹ [۵۵]	روشی جدید برای آموزش‌دادن به یادگیرنده‌ها با IM برای کاوش در محیط. این روش بر به حداکثر رساندن کارایی کاوش تمرکز دارد و منجر به یادگیرنده‌هایی می‌شود که می‌توانند به طور کامل محیط را کاوش کرده و درعین‌حال زمان لازم برای دستیابی به آن کاوش را به حداقل برسانند.	کاوش کارآمدتر در مدت‌زمان کمتر و پوشش بخش بیشتری از محیط	طراحی اهداف جانشین مؤثر که به طور دقیق کارایی کاوش را به تصویر می‌کشند، نیاز به تحقیقات بیشتر برای تأیید اثربخشی آن در سناریوهای مختلف دنیای واقعی دارد.	الگوریتم‌های RL با تابع پاداش تغییریافته برای تمرکز بر آنتروپی کاوش به‌جای سیگنال‌های پاداش سنتی
۲۰۲۱ [۲۹]	بررسی IM در عامل مصنوعی به‌طوری‌که این عامل بدون نیاز به پاداش‌های بیرونی قادر به اکتشاف محیط خود و کسب مهارت‌های مفید باشند. هدف، یادگیری کلی و فشرده‌ای که عامل را به کاهش آنتروپی بازدیدهای حالت‌های خود تشویق می‌کند.	توانایی عامل در اکتشاف و کنترل محیط‌های پیچیده و پویا بدون نیاز به پاداش‌های بیرونی، یادگیری رفتار پیچیده و کاهش عدم قطعیت	نیاز به تحقیقات بیشتر برای رسمی‌سازی ارتباط بین ترمودینامیک و نظریه اطلاعات و محدودیت‌های مدل‌های فضای حالت نهفته در مدل‌سازی صحیح باورها	استفاده از فیلتر بیزین عمیق و مدل‌سازی فضای حالت نهفته با استفاده از سیگنال IR برای کاهش آنتروپی بازدیدهای حالت عامل
۲۰۲۱ [۳۰]	ارائه روشی جدید برای کشف در RL آنلاین از طریق حداکثرکردن انحراف از مناطق کاوش شده با اضافه‌کردن یک تنظیم‌کننده تطبیقی به هدف استاندارد RL که تعادلی بین اکتشاف و بهره‌برداری از تجربیات ایجاد می‌کند.	بهبود کارایی نمونه‌ها، سازگاری با روش‌های مختلف یادگیری تقویتی، و اجرای آسان. بهبود عملکرد روش‌های پاداش UCB[1] سنتی بدون نیاز به محاسبات پیچیده	محدودیت در استفاده از نمایش‌های ساده حالت‌ها و نیاز به ارتقای برای استفاده از نمایش‌های فشرده‌تر. عملکرد پایین‌تر در محیط‌های با تغییرات غیرقابل‌پیش‌بینی	افزودن یک تنظیم‌کننده تطبیقی به هدف استاندارد RL همراه با یک IR جدید
۲۰۲۲ [۳۵]	بررسی و ارائه روش‌های RL سلسله‌مراتبی برای توسعه مهارت‌های حل مسئله پیشرفته در عامل مصنوعی با الهام از مکانیسم‌های سلسله‌مراتبی مشاهده شده در حیوانات و انسان‌ها	بهبود توانایی حل مسئله با استفاده از یادگیری سلسله‌مراتبی، انتقال دانش از یک وظیفه به وظایف جدید و مشابه و توانایی برنامه‌ریزی تصمیم‌گیری در زمان واقعی	دشواری یکپارچه‌سازی مکانیسم‌های شناختی زیستی به یک معماری واحد، پیچیدگی محاسباتی و تحقق عملی این روش‌ها در عامل‌های مصنوعی	RL سلسله‌مراتبی شامل مکانیزم‌های شناختی؛ مانند انتزاع ترکیبی و پردازش پیش‌بینی‌کننده
۲۰۲۲ [۴۶]	اهمیت استفاده از غیرمارکوفی بودن در بهینه‌سازی اکتشاف حداکثر آنتروپی حالت‌ها باوجود تعداد نمونه‌های محدود	بهبود کارایی نمونه‌ها در RL آنلاین و اکتشاف یکنواخت‌تری در فضای حالت در هر تجربه واحد	یافتن سیاست غیرمارکوفی بهینه برای حداکثر آنتروپی حالت نمونه محدود، به‌طورکلی مشکل NP-hard است.	استفاده از سیاست‌های غیرمارکوفی
۲۰۲۳ [۴۸]	معرفی یک تکنیک اکتشافی جدید به نام «حداکثرسازی آنتروپی حالت وابسته به ارزش». این تکنیک به طور جداگانه آنتروپی حالات را که بر اساس ارزش تخمین زده شده هر حالت است، محاسبه و سپس میانگین آنها را حداکثر می‌کند تا مشکل توزیع نامتوازن در روش‌های اکتشاف قبلی را حل کند.	تسریع آموزش الگوریتم‌های RL، بهبود کارایی نمونه‌گیری، و حل مشکل توزیع نامتوازن حالات باارزش بالا و پایین	نبود درک نظری کامل از نحوه عملکرد آنتروپی حالت وابسته به ارزش، محدودیت در استفاده از تخمین‌گر آنتروپی حالت غیرپارامتریک و نیاز به توسعه روشی که بدون نیاز به پاداش وظیفه فضاهای حالت را به زیر فضاهای معنادار تقسیم کند.	تکنیک حداکثرسازی آنتروپی حالت وابسته به ارزش که از تخمین‌گر آنتروپی و یک طرح نرمال‌سازی ارزش برای حفظ یکنواختی توزیع ارزش‌ها استفاده می‌کند.
۲۰۲۳ [۴۹]	پیشنهاد الگوریتمی به نام حداکثرسازی آنتروپی محدودشده برای حل مشکلات اکتشاف ایمن بدون توجه به وظایف خاص. این الگوریتم با استفاده از یک تخمین‌گر آنتروپی K نزدیک‌ترین همسایه برای ارزیابی کارایی اکتشاف و به طور هم‌زمان کاهش هزینه‌های ایمنی، به یادگیری یک سیاست که آنتروپی حالت را تحت محدودیت‌های ایمنی حداکثر می‌کند، می‌پردازد.	دستیابی به یک سیاست اکتشافی ایمن در محیط‌های پیچیده، بهبود کارایی نمونه‌گیری برای وظایف هدف و ارائه یک روش عملی و تقریباً همگرا برای اکتشاف ایمن بدون توجه به وظایف خاص	پیچیدگی محاسباتی تخمین آنتروپی حالت در دامنه‌های پیچیده، چالش‌های ایجاد تعادل مناسب بین اکتشاف و ایمنی، و نیاز به تنظیم دقیق وزن‌های ایمنی در محیط‌های واقعی	الگوریتم حداکثرسازی آنتروپی محدودشده با استفاده از تخمین‌گر آنتروپی K نزدیک‌ترین همسایه برای ارزیابی اکتشاف
۲۰۲۴ [۵۲]	ارائه سیستم یادگیری که می‌تواند با استفاده از مقدار کمی داده‌های دنیای واقعی، یک مدل شبیه‌سازی را بهبود بخشد و سپس یک استراتژی کنترل دقیق طراحی کند که در دنیای واقعی قابل‌اجرا باشد. این سیستم از طریق یک فرایند اکتشاف فعال در دنیای واقعی، داده‌های باکیفیت بالا جمع‌آوری کرده و از آنها برای شناسایی پارامترهای فیزیکی استفاده می‌کند.	نیاز به مقدار کمی داده‌های دنیای واقعی برای بهبود مدل شبیه‌سازی، امکان انتقال موفق از شبیه‌سازی به دنیای واقعی و کاهش نیاز به تنظیم دستی پارامترهای شبیه‌سازی	پیچیدگی فرایند شناسایی سیستم، نیاز به اجرای دقیق اکتشاف فعال و محدودیت در توانایی مدل شبیه‌سازی در بازسازی صحیح محیط‌های پیچیده	استفاده از الگوریتمی برای به حداکثر رساندن اطلاعات فیشر

[1] Upper Confidence Bound



شکل ۱: خلاصه‌ای از تحقیقات کلیدی و پیشرفت‌های انگیزش درونی به همراه منابع

۵-۱- مدل‌های مبتنی بر دانش

سیستم‌های IM مبتنی بر دانش (ظرفیت پیش‌بینی)، برای هدایت یادگیری بر اساس سطح یا تغییرات دانش مناسب هستند [۴۲]. بیشتر مدل‌های پیشنهادی IM مبتنی بر دانش هستند و به محرک دریافت شده به وسیله سیستم یادگیری وابسته هستند و نه به مهارت‌های سیستم. در مدل مبتنی بر دانش، جالب بودن به مقایسه بین جریان پیش‌بینی شده از مقدار حسگر حرکتی، بر اساس یک مدل داخلی درونی، با جریان واقعی مقدارها بدست می‌آید. چنانچه مدل آموخته شود، این مدل به طور خاص منجر به انگیزه سازگاری می‌شود (انگیزه سازگارانه به مکانیزمی اشاره دارد که سطوح مختلفی از علاقه به یک موقعیت/فعالیت را با توجه به لحظه خاص رشد(توسعه) که در آن قرار می‌گیرد، تعیین می‌کند). اولین رویکرد محاسباتی به IM براساس معیارهای ناسازگاری (یا رزونانس[1]) بین وضعیت‌های تجربه شده با دانش و انتظارات ربات در مورد این شرایط، تعریف می‌شود. این امر به عنوان یک فعالیت فعال که در آن ربات اعمالی را انجام می‌دهد و خروجی واقعی اعمال خود را با دانش و انتظاراتاش از نتیجه این اعمال مقایسه می‌کند [۲۵]، محسوب می‌شود. از اولین تحقیقات می توان به مقالات اشمیدهوبر[2] اشاره کرد [۴۳،۴۴]. مبنای هر دو مقاله بر اساس کمینه کردن عدم قطعیت یا خطای عامل در پیش بینی اش می باشد. الگوریتم پایه هر دو روش بر اساس الگوریتم RL است. البته پیاده سازی پیچیده، نیاز به محاسبات زیاد و تنظیم پارامترها از چالش های این تحقیقات بود. روی[3] و مک‌کالم[4] رویکردی مبتنی بر بیزین برای یادگیری فعال بهینه با استفاده از تخمین مونت کارلو ارائه نمودند که با چالش نیاز به نمونه برداری دقیق از توزیع پسین بیزین مواجه بود [۴۵]. اودیر[5] با تمرکز روی مفهوم پیشرفت یادگیری، مکانیزمی به نام «کنجکاوی هوشمند و سازگار» معرفی کرد که محرکی درونی برای ربات ایجاد می کند تا بر روی موقعیت هایی تمرکز کند که پیشرفت یادگیری او را به حداکثر می رساند [۴۷]. او نیز ایده اش را بر روی الگوریتم RL پیاده سازی کرد. مریک[6] و ماهر[7] مدل محاسباتی انگیزش برای عامل یادگیری برای تحقق یادگیری چند وظیفه ای و سازگار در محیط های پیچیده و پویا معرفی کردند. این مدل برای وظایفی که نمی تواند به طور کامل قبل از یادگیری پیش بینی شود، با استفاده از الگوریتم های RL، ارائه شد. البته مقیاس پذیری آن برای وظایف پیچیده تر و شکست در این شرایط به علت عدم

[1] Resonances
[2] Schmidhuber
[3] Roy
[4] McCallum
[5] Oudeyer
[6] Merrick
[7] Maher

تضمین باقی ماندن در یک رویداد به اندازه کافی جالب، از چالش های اصلی این پژوهش می باشد. در جدول ۲ برخی از پژوهش های مرتبط اخیر با این بخش آورده شده است. علاوه بر ارائه ایده اصلی مقاله، مزایا، معایب یا چالش های پیش رو، روش یا الگوریتم پایه آن نیز آورده شده است.

۵-۲- مدل‌های مبتنی بر شایستگی

سیستم‌های انگیزش درونی مبتنی بر شایستگی (ظرفیت انجام کار) برای هدایت یادگیری بر اساس سطح یا بهبود شایستگی به کار برده می‌شوند [۵۶]. مدل‌های مبتنی بر شایستگی، سیستم‌های مصنوعی هستند که یادگیری آن‌ها به وسیله شکل‌های مختلفی از انگیزش درونی مبتنی بر شایستگی هدایت می‌شود. در مدل‌های مبتنی بر شایستگی، جالب بودن به مقایسه بین اهداف خودساخته (که پیکربندی‌های خاصی در فضای حسگر حرکتی هستند)، و میزانی که در عمل به آن بر اساس یک مدل معکوس داخلی که ممکن است آموخته شود، پرداخته می‌شود. بدین ترتیب، این مقایسه‌ها درجه عملکرد/شایستگی عامل را مشخص می‌کنند و اگر مدل نیز آموخته شود، به طور معمول منجر به انگیزه تطبیقی می‌شود. این رویکرد مستقیما از تئوری‌های منطقی روانشناسی الهام گرفته شده است. این تئوری‌ها عبارتند از: تاثیرگذاری [۲۴]، علیت فردی [۳۲]، شایستگی، خودتصمیم‌گیری [۲۵] و غوطه‌وری [۳۳]. مرکز بحث در اینجا مفهوم «چالش» است. این مفهوم مرتبط با اندازه‌گیری دشواری به عنوان اندازه‌گیری بهره‌وری واقعی می‌باشد. «چالش» در اینجا هر پیکربندی حسگر حرکتی یا هر مجموعه از مشخصات حسگر حرکتی است که هر فرد خودش تنظیم کرده و تلاش می‌کند از طریق عمل به این هدف دست یابد.

بارتو[1] و همکاران [۳۷] با ارائه چارچوبی برای یادگیری تقویتی انگیزش درونی یکی از اساسی‌ترین و پایه‌ای‌ترین چارچوب‌ها را ارائه کردند. ایده اصلی آن‌ها بررسی یادگیری مبتنی بر انگیزش درونی در عامل‌های مصنوعی است تا بتوانند سلسله‌مراتبی از مهارت‌های قابل‌استفاده مجدد را بسازند و توسعه دهند که برای خودمختاری مؤثر لازم است. آن‌ها با الهام از مفهوم انگیزش درونی در روان‌شناسی، معماری را ارائه کردند که عامل رفتارهایش را به‌خاطر خودش انجام می‌دهد، نه به‌عنوان گامی در حل مشکلات عملی. عامل‌های مصنوعی می‌توانند مهارت‌های قابل‌استفاده مجددی را یاد بگیرند که برای طیف وسیعی از چالش‌ها مفید هستند، بدون نیاز به تنظیم دستی برای هر وظیفه خاص. البته چارچوب پیشنهادی برای ایجاد IR پیچیده‌تر که بتوانند انواع مختلفی از کاوش و دست‌کاری را پوشش دهند، کافی نبوده است.

در [۵۸] محققان تلاش کردند با استفاده از الگوریتم یادگیری تقویتی عامل ـ منتقد به همراه شبکه عصبی، یک مدل نوین یادگیری تقویتی انگیزش درونی را توسعه دهند که ربات‌ها قادر باشند به طور خودکار از IR استفاده کرده و مهارت‌های عمومی ساخته‌شده را به طور خودکار ترکیب کرده تا وظایف خاص متعددی را حل کنند. انعطاف‌پذیری و خودکارسازی در کشف تقویت‌کننده‌های درونی، قابلیت استفاده در محیط‌های رباتیک واقعی و توانایی مقابله با محیط‌های پیوسته و نویزی، عملکرد و استحکام بهتر نسبت به سایر معماری‌های استاندارد از مزایای این تحقیق است. معرفی الگوریتم تولید هدف خودتنظیمی توسط بارانس[2] و همکاران [۶۰] باعث شده است که فضای فعال یادگیری حرکتی هوشمند و قدرتمند به‌عنوان یک مکانیسم بررسی هدف برای ربات‌ها امکان یادگیری معکوس کینماتیکی را فراهم کند. همچنین امکان کاوش و یادگیری مهارت‌های پیچیده‌تر را فراهم می‌کند. چالشی که این الگوریتم با آن سروکار دارد مسئله مقیاس‌پذیری در محیط‌های حسگر ـ موتور متنوع است. ایده جالب دیگری که توسط هانگل[3] و همکاران [۶۱] ارائه شده است، یادگیری مهارت‌های جدید توسط ربات‌ها با استفاده از بازی فعال و خلاقیت است که باعث می‌شود ربات بتواند مهارت‌های جدید را برای محیط‌های خاصی یاد بگیرد و از آن‌ها برای حل وظایف مشابه استفاده کند. این ایده با استفاده از ترکیب الگوریتم یادگیری تقویتی مبتنی بر مدل و بدون مدل پیاده‌سازی شده است. مشکل این روش، طولانی‌شدن زمان یادگیری در مواجهه با تعداد زیادی مهارت است که برای رفع آن نیاز به مدل‌های پیچیده‌تری است. در جدول ۳ سایر پژوهش‌های اخیر مرتبط شامل ایده اصلی مقاله، مزایا، معایب یا چالش‌های پیشرو و روش یا الگوریتم پایه آن‌ها نشان‌داده‌شده است.

۵-۳- تازگی (نوآوری)

تشخیص تازگی عامل مهمی برای کارکرد مؤثر و درازمدت ربات‌های هوشمند است که برای کشف، تکامل و تصمیم‌گیری بر اساس دانش و تجربیات تجمعی، طراحی شده است. تشخیص تازگی و انگیزش درونی به‌شدت به هم مرتبط هستند و نقش مهمی در روند یادگیری تجمعی ایفا می‌کنند. تازگی به‌عنوان فرایند تشخیص محرک‌های جدید که «متفاوت از هر چیزی که تاکنون شناخته شده است، جدید، جالب و کمی عجیب به نظر می‌رسد» می‌توان در نظر گرفت [۶۵]. در دسته‌بندی تشخیص تازگی یکی از اولین پژوهش‌ها توسط مارسلند[4] و همکاران [۶۶] صورت گرفت. آن‌ها با استفاده از نقشه خودسازمانده، الگوریتمی برای شناسایی ویژگی‌های جدید یا غیرعادی در محیط به وسیله یک ربات متحرک را معرفی کردند که با استفاده از اسکن‌های سونار یک مدل داخلی از «عادی بودن» ایجاد می‌کند و با ارزیابی جدید بودن هر اسکن سونار، ویژگی‌های جدید را شناسایی می‌کند. توانایی شناسایی ویژگی‌های جدید در زمان واقعی، کاهش نیاز به پردازش بیش از حد داده‌های حسی و قابلیت استفاده در محیط‌های پویا و تغییرپذیر از مزایای این تحقیق است. اما شناسایی نادرست محرک‌های جدید به عنوان آشنا، و نیاز به بررسی و بهبود سیستم‌های حسی اضافی از معایب این تحقیق محسوب می شود.

[1] Barto
[2] Baranes
[3] Hangl
[4] Marsland

جدول ۳: مدل‌های محاسباتی IM مبتنی بر شایستگی

سال انتشار و مرجع	ایده اصلی	مزایا	معایب/چالش‌ها	روش یا الگوریتم
۲۰۱۸ [۶۲]	یادگیری مهارت‌های اجتماعی شبیه به انسان توسط ربات‌ها با استفاده از چارچوب IMRL	یادگیری مهارت‌های اجتماعی شبیه به انسان بدون نیاز به پاداش‌های مستقیم و توانایی تصمیم‌گیری انسانی‌تر توسط ربات	نیاز به داده‌های آموزشی فراوان برای یادگیری و محدودیت در فضای عملیاتی ربات به چهار عمل	یادگیری تقویتی عمیق (DRL)[1] با IM شامل شبکه پیش‌بینی کنش ـ شرطی(Pnet) و شبکه سیاستی (Qnet)
۲۰۱۹ [۶۴]	چارچوبی جدید برای برنامه‌ریزی حرکت آنلاین احتمالاتی با سازوکار تطبیق آنلاین کارآمد بر اساس شبکه عصبی بازگشتی تصادفی. ربات قادر می‌شود از تعاملات فیزیکی کم، یاد گرفته و با محیط‌های جدید به سرعت سازگار شوند.	تطبیق سریع آنلاین، یادگیری کارآمد از تعاملات کم، عدم نیاز به تعریف وظایف یادگیری توسط کارشناسان و قابلیت استفاده در محیط‌های واقعی	نیاز به تعاملات بیشتر برای یادگیری سیگنال‌های محلی، مشکلات مقیاس‌پذیری به فضاهای با ابعاد بالاتر و بهبود کدگذاری برای کاربردهای پیچیده‌تر	شبکه عصبی بازگشتی تصادفی با سیگنال‌های IM و استراتژی تکرار ذهنی ترکیب شده تا یادگیری و تطبیق آنلاین را انجام دهد.
۲۰۲۰ [۵۳]	ارائه روش‌های جدید برای کاهش پیچیدگی نمونه‌برداری در یادگیری مدل ربات‌ها به‌صورت آنلاین. این روش‌ها از طریق سیگنال IM جدید که عناصر مبتنی بر دانش و شایستگی را ترکیب می‌کند، اکتشاف را هدایت می‌کنند. همچنین از تکرار ذهنی اپیزودیک آنلاین برای تسریع یادگیری و افزایش کارایی استفاده می‌شود.	کاهش قابل‌توجه پیچیدگی نمونه‌برداری، تسریع یادگیری آنلاین، کاهش سایش و پارگی در ربات‌ها، و امکان یادگیری هم‌زمان چندین مدل با استفاده از یک مجموعه‌داده تجربی، کارایی بالا و قابلیت استفاده در دنیای واقعی.	نیاز به تنظیم دقیق پارامترها برای عملکرد بهینه و پیچیدگی در پیاده‌سازی روش‌ها در محیط‌های واقعی، تعمیم روش‌ها به ربات‌های با درجه آزادی بسیار بالا.	استفاده از تکنیک GB[2] که به ربات اجازه یادگیری مهارت‌های خود را به‌صورت آنلاین و از ابتدا می هد. از سیگنال IM جدید و تکرار ذهنی اپیزودیک آنلاین برای بهبود کارایی و سرعت یادگیری استفاده می‌کند.
۲۰۲۱ [۵۴]	ارائه شمای یادگیری ترکیبی IM ـ EM برای ربات‌های توسعه‌ای شامل ترکیب IM با یادگیری از مشاهده برای تسریع فرایند یادگیری ربات. چهار عنصر اصلی عبارت‌اند از: سیگنال IM احتمالاتی، سیگنال انگیزش بیرونی احتمالاتی، تشخیص نوآوری و اندازه‌گیری درجه نوآوری.	افزایش سرعت یادگیری، افزایش کارایی در کاهش تعداد نمونه‌های موردنیاز برای یادگیری مدل و توانایی تصمیم‌گیری مستقل ربات در مورد زمان و چگونگی اکتشاف است.	پیچیدگی در پیاده‌سازی شماهای یادگیری در محیط‌های واقعی، نیاز به تنظیم دقیق پارامترها و مدیریت تعادل بین اکتشاف و بهره‌برداری از تجربیات، نیاز به داده‌های انسانی دقیق برای یادگیری مشاهده‌ای	شمای یادگیری ترکیبی از IM و انگیزش که از تشخیص نوآوری، اندازه‌گیری درجه نوآوری، سیگنال IM احتمالاتی و سیگنال انگیزش بیرونی احتمالاتی برای هدایت اکتشاف و یادگیری ربات استفاده می‌کند.
۲۰۲۲ [۵۷]	عامل یادگیری تقویتی سلسله‌مراتبی (HRL)[3] جدید که مهارت‌های گسترش‌یافته موقتی را با یک مدل پیشرو در انتزاع نمادین حالت محیط برای برنامه‌ریزی، ترکیب می‌کند. مهارت‌های متنوعی را از طریق IM به‌صورت یادگیری بدون نظارت می‌آموزد و از این مهارت‌ها به‌عنوان اقدامات نمادین استفاده می‌کند.	توانایی حل وظایف پیچیده‌ای که نیاز به کنترل پیوسته و برنامه‌ریزی بلندمدت دارند. این روش در مقایسه با عامل RL مسطح و سلسله‌مراتبی پایه، عملکرد بهتری دارد.	پیچیدگی در تعریف انتزاع حالت، نیاز به مقیاس‌پذیری برنامه‌ریزی جستجوی اول عرض برای عمق‌های زیاد راه‌حل و دشواری در یادگیری تمام مهارت‌های ممکن در محیط‌های پیچیده	عامل HRL به نام SEADS به همراه انتزاع نمادین حالت و یک مدل پیشرو برای یادگیری مهارت‌های گوناگون
۲۰۲۳ [۵۹]	معرفی و بررسی معماری REAL-X که برای ایجاد سیستم‌های رباتیک باقابلیت یادگیری خودکار و پایان‌ناپذیر از طریق تعامل با محیط. این تعامل با استفاده از IM برای اکتشاف و رویکردهای برنامه‌ریزی که به طور پویا انتزاع را افزایش می‌دهند، صورت می‌گیرد.	توانایی یادگیری خودکار و پایان‌ناپذیر بدون وابستگی به وظایف یا پاداش‌های از پیش تعریف شده، افزایش سطح انتزاع به طور پویا، و قابلیت سازگاری با شرایط مختلف محیطی	یادگیری از سطح صفر بدون داشتن هرگونه دانش پیشین، پیچیدگی در تنظیم اهداف و وظایف خودمختار، مشکلات ناشی از نادر بودن پاداش‌ها در مراحل ابتدایی یادگیری، تعامل هم‌زمان با چندین شیء	معماری با طراحی ماژولار شامل سه جزء اصلی: انتزاع‌کننده، کاوشگر، و برنامه‌ریز. این اجزاء به ترتیب مسئول انتزاع ورودی‌های حسی، تولید تجربیات حرکتی برای یادگیری اهداف و اقدامات، و تدوین و اجرای برنامه‌های عملیاتی برای دستیابی به اهداف خارجی
۲۰۲۴ [۶۳]	استفاده از RL و یادگیری مبتنی بر برنامه درسی برای دستیابی به کنترل حرکتی در وظایف پیچیده دست‌کاری اشیاء. استفاده از شبیه‌سازهای عضلانی ـ اسکلتی و الگوریتم‌های یادگیری قدرتمند	ارائه رویکردی کارآمد برای شبیه‌سازی کنترل حرکتی پیچیده، توانایی تحلیل سیاست‌های یادگیری و ارائه بینش‌های جدید در مورد کنترل موتور بیولوژیکی.	هزینه محاسباتی بالا برای ترکیب شبیه‌سازهای عضلانی ـ اسکلتی با RL و مشکلات بهینه‌سازی مدل‌ها	الگوریتم RL خاصی به نام PPO[4] [۱۵۳] با معماری بازگشتی شامل لایه‌های LSTM[5]

[1] Deep Reinforcement Learning
[2] Goal Babbling
[3] Hierarchical Reinforcement Learning
[4] Proximal Policy Optimization
[5] Long short-term memory

۲۰۲۴ [۷۴]	رویکردی جدید برای کشف مهارت‌ها به نام SLIM که از منتقدان متعدد در چارچوب عامل ـ منتقد استفاده می‌کند. بهبود کشف مهارت‌های متغیر پنهان در زمینه‌ی دست‌کاری رباتیک	کشف مهارت‌های متنوع و ایمن، بهبود HRL و برنامه‌ریزی، و توانایی انجام آزمایش‌های گسترده برای نشان‌دادن مزایای رویکرد چند منتقد	فرض طراحی آسان پاداش‌های اضافی، نیاز به بررسی بیشتر تداخل بین پاداش‌ها و مشکلات بالقوه در پیاده‌سازی به‌صورت شبیه‌سازی تا واقعیت	RL سلسله‌مراتبی با استفاده از رویکرد چند منتقد و ترکیب پاداش‌ها بهبودیافته

جدول ۴: مدل‌های محاسباتی IM مبتنی بر تازگی

سال انتشار و مرجع	ایده اصلی	مزایا	معایب/چالش‌ها	روش یا الگوریتم
۲۰۱۷ [۶۸]	معرفی چارچوبی برای برنامه‌ریزی حرکتی آنلاین و یادگیری مبتنی بر شبکه‌های عصبی بازگشتی تصادفی بیولوژیکی است که از سیگنال‌های IM و یک استراتژی بازپخش ذهنی برای انطباق آنلاین و به طور کارآمد نمونه‌گیری استفاده می‌کند.	انطباق آنلاین و بهبودیافته، نمونه‌گیری کارآمد و بی‌درنگ، قابلیت استفاده از تعداد کمی از تعاملات فیزیکی برای یادگیری، و قابلیت استفاده مدل در محیط‌های مختلف	نیاز به تجربه کارشناسی، پایین آوردن دانش متخصص و مشکلات مرتبط با اکتساب مدل‌های جدید برای محیط‌های ناشناخته است.	شبکه‌های عصبی بازگشتی تصادفی بیولوژیکی به همراه سیگنال‌های IM.
۲۰۱۹ [۶۹]	ایده اصلی این مقاله معرفی یک روش تشخیص نوآوری مبتنی بر انتظار است که از روش‌های شبکه‌های عصبی بازگشتی آنلاین استفاده می‌کند تا ساختار شبکه را به طور پویا یاد بگیرد و با تغییرات در داده‌ها ساختار خود را تطبیق دهد.	توانایی یادگیری آنلاین و پویا، تطبیق ساختار با تغییرات در داده، و امکان شناسایی نوآوری‌های محلی همراه با انتظارات است.	محدودیت‌هایی در تشخیص داده‌های نوآور، ازدست‌دادن جزئیات کوچک در داده‌های حسی و نیاز به بهبود سیستم تشخیص نوآوری به‌منظور درنظرگرفتن ویژگی‌های محلی جالب است.	شبکه‌های عصبی بازگشتی آنلاین است که بر اساس انتظارات و پیش‌بینی داده‌های بعدی ساختار شبکه را پویا یاد می‌گیرد.
۲۰۲۰ [۹۰]	پیشنهاد یک روش جدید برای هدایت جستجوی نوآوری در RL است که با استفاده از رمزگذار خودکار به‌جای استفاده از روش‌های مبتنی بر فاصله برای ارزیابی نوآوری، به ارزیابی رفتارهای عامل‌ها می‌پردازد. این روش به دنبال بهبود مقیاس‌پذیری و قابلیت تعمیم روش‌های جستجوی نوآوری در مسائل پیچیده است.	بهبود مقیاس‌پذیری و قابلیت تعمیم، کاهش نیاز به ارزیابی‌های متعدد سیاست‌ها، و توانایی تشویق کاوش مؤثرتر در فضای رفتارهای کمتر دیده شده	پیچیدگی در مدل‌سازی یا یادگیری ویژگی‌های رفتاری مناسب و نیاز به تحقیقات بیشتر در یادگیری نمایش‌های رفتارهای کارآمد با استفاده از معماری‌های دنباله به دنباله است.	ترکیبی از روش‌های جستجوی نوآوری[1] و RL با استفاده از رمزگذار خودکار برای ارزیابی نوآوری رفتارهای عامل‌ها است.
۲۰۲۱ [۹۱]	بررسی رفتار انسان‌ها در مواجهه با محیط‌های با پاداش خلوت و تغییرات ناگهانی، نویسندگان نشان می‌دهند که نظریه‌های کلاسیک RL نمی‌توانند رفتار انسان‌ها را در غیاب پاداش خارجی یا در زمان تغییرات محیطی توضیح دهند. به همین دلیل، آن‌ها پیشنهاد می‌کنند که تئوری‌های RL باید شامل مفاهیم تعجب و نوآوری باشند، هر کدام با نقش‌های جداگانه.	توانایی تبیین بهتر رفتار اکتشافی و تطبیقی انسان‌ها است. این روش نشان می‌دهد که چگونه انسان‌ها از نوآوری برای اکتشاف و از تعجب برای به‌روزرسانی سریع مدل‌های داخلی و پاسخ‌های عادت شده استفاده می‌کنند.	پیچیدگی در تفکیک دقیق نقش‌های نوآوری و تعجب در یادگیری و تصمیم‌گیری است. همچنین، تعیین دقیق تأثیر هر کدام بر رفتار انسان و اندازه‌گیری آن‌ها در سیگنال‌های EEG[2] نیاز به تحقیق و توسعه بیشتر دارد.	مدل SurNoR[3] است که یک الگوریتم RL ترکیبی است که شامل تعجب، نوآوری و پاداش می‌شود. این مدل برای توضیح رفتار انسان‌ها و سیگنال‌های EEG در وظایف تصمیم‌گیری توالی عمیق طراحی شده است.
۲۰۲۱ [۱۱۷]	توسعه یک چارچوب جدید با استفاده از چندین فیلتر تشخیص نوآوری برای یادگیری یک مدل از دید ربات است. سپس از این مدل برای برجسته‌کردن درک‌های غیرمشابه هنگام کاوش ربات در یک محیط استفاده می‌شود. هدف اصلی ترکیب چندین فیلتر نوآوری این است که هر فیلتر در تشخیص انواع خاصی از نوآوری‌ها خوب عمل می‌کند؛ بنابراین، یک چارچوب جدید پیشنهاد می‌شود که نشان می‌دهد چگونه می‌توان چندین فیلتر تشخیص نوآوری مختلف	تشخیص نوآوری‌های قوی و قابل‌اعتماد، عملکرد بهتر در مقایسه با فیلترهای نوآوری مبتنی بر انتظار و ظاهر به‌صورت جداگانه. این سیستم کاملاً آنلاین و در زمان واقعی است که برای کاربردهای رباتیک موبایل بسیار مهم است.	فیلترهای نوآوری زیر سطح دارای ضعف‌هایی هستند که می‌تواند تأثیر منفی بر عملکرد کلی داشته باشد. سرعت پاسخ فیلتر نوآوری می‌تواند کندتر از زمانی باشد که فقط ویژگی‌های محلی به فیلتر نوآوری ارائه شود. پیچیدگی اجرای ترکیب چندین فیلتر تشخیص نوآوری.	استفاده از دو مدل تشخیص نوآوری: مدل تشخیص نوآوری مبتنی بر انتظار با استفاده از شبکه عمیق MobileNetV2 و مدل تشخیص نوآوری مبتنی بر ظاهر با استفاده از ویژگی‌های SURF[4]

[1] Novelty Search
[2] Electroencephalogram
[3] Surprise-Novelty- Reward
[4] Speeded Up Robust Feature

	را ترکیب کرد تا وضعیت کلی نوآوری را بر روی ویژگی‌های بصری قوی‌تر نماید.			
۲۰۲۳ [۱۱۸]	بررسی اینکه آیا انسان‌ها در مواجهه با محرک‌های تصادفی و بدون پاداش، مانند الگوریتم‌های IMRL، دچار حواس‌پرتی می‌شوند. همچنین بررسی کند که آیا این حواس‌پرتی به‌مرور زمان از بین می‌رود یا پایدار می‌ماند.	سادگی محاسباتی نسبت به روش‌های دیگر و توانایی مدل‌سازی دقیق‌تر رفتارهای کنجکاوی انسان، این روش می‌تواند به طور کیفی و کمی رفتار انسان‌ها را در مواجهه با محرک‌های تصادفی بدون پاداش پیش‌بینی کند.	رفتار انسان‌ها به طور کامل توسط مدل‌های موجود توضیح داده نمی‌شود و نیاز به مطالعات بیشتر برای درک بهتر نقش تازگی و پاداش در رفتارهای اکتشافی انسان‌ها وجود دارد.	الگوریتم‌های IMRL که بر اساس IM مانند تازگی، شگفتی و افزایش اطلاعات طراحی شده‌اند.
۲۰۲۴ [۱۱۹]	معرفی یک مدل محاسباتی جدید برای IM در RL است که بر محدودیت‌های کاوش‌های مبتنی بر شگفتی غلبه می‌کند. پاداش به‌جای شگفتی، از تازگی شگفتی محاسبه می‌شود.	کاهش حواس‌پرتی ناشی از مشاهدات غیرقابل‌پیش‌بینی یا نویزی، بهبود رفتارهای اکتشافی و افزایش عملکرد نهایی در محیط‌های با پاداش کم.	پیچیدگی محاسباتی مرتبط با پیاده‌سازی و آموزش شبکه حافظه و نیاز به تنظیم دقیق مدل‌های مختلف پیش‌بینی شگفتی است.	سیستم حافظه شگفتی[1] است که شامل حافظه انجمنی خود رمزگذار و حافظه اپیزودیک مبتنی بر توجه برای محاسبه تازگی شگفتی است.
۲۰۲۴ [۱۲۰]	معرفی روشی جدید و مؤثر برای کاوش هماهنگ چندعاملی در RL غیرمتمرکز که از طریق اشتراک تازگی محلی و استفاده از IR مبتنی بر تازگی و اطلاعات متقابل وزنی، به عامل‌ها کمک می‌کند تا به‌صورت هماهنگ کاوش کنند.	بهبود عملکرد در محیط‌های چندعاملی با پاداش‌های خلوت، کاهش کاوش‌های تکراری و افزایش هماهنگی بین عامل‌ها، عملکرد خوب باوجود محدودیت‌های ارتباطی	نیاز به شبکه ارتباطی کاملاً متصل برای اشتراک تازگی بین عامل‌ها و پیچیدگی محاسباتی مرتبط با اندازه‌گیری اطلاعات متقابل وزنی	سیستم MACE[2] که شامل اشتراک تازگی محلی و IR مبتنی بر اطلاعات متقابل وزنی برای هدایت کاوش هماهنگ عامل‌ها.

روش جدیدی برای شناسایی نوآوری در زمان واقعی با استفاده از تحلیل مؤلفه‌های اصلی افزایشی[3] [۶۷] ارائه شده است و عملکرد آن را با تکنیکی بر اساس شبکه عصبی GWR[4] مقایسه کرده‌اند. این روش باعث عملکرد بهتر در شناسایی نوآوری، کاهش ابعاد داخلی داده‌ها، توانایی بازسازی خوب و ارزیابی ویژگی‌های یاد گرفته شده توسط سیستم شده است. نیاز به کاهش ابعاد داده‌ها قبل از پردازش با شبکه GWR و ارزیابی شباهت بین ورودی‌ها در ابعاد کاهش‌یافته، مشکلاتی است که این الگوریتم با آن‌ها مواجه است. بررسی نقش شناسایی نوآوری به‌عنوان یک انگیزش درونی در یادگیری انباشته ربات‌ها توسط [۶۵] انجام شده است. همچنین روش‌های مختلف شناسایی نوآوری برای عملکرد مؤثر و بلندمدت ربات‌های هوشمند طراحی شده بر اساس دانش و تجربه‌ی انباشته بررسی شده است. این روش مبتنی بر «خوگیری[5]» از ساختارهای یادگیری پویا و افزایشی برای شناسایی نوآوری و انجام عملکردهای یادگیری در زمان بلندمدت استفاده می‌کند. مشکلات مربوط به تصمیم‌گیری در مورد آموزش و توسعه بیشتر، و پیچیدگی‌های مربوط به پردازش داده‌های نویزی از چالش‌های این روش است. در جدول ۴ سایر روش‌ها با جزئیات بیشتری ارائه شده‌اند.

۵-۴- توانمندسازی، قدرت، کنترل

توانمندسازی نشان‌دهنده درجه‌ی آزادی است که یک عامل بر محیط دارد. یک عامل توانمند ترجیح می‌دهد که وضعیت‌ها جایی باشند که او بیشترین کنترل را داشته باشد و همچنین بتواند این کنترل را حس کند. کلاییوبین[6] و همکاران [۷۰] معیاری جدید برای اندازه‌گیری کنترل در عامل‌های هوش مصنوعی که به آن «تقویت» یا «توانمندسازی» گفته می‌شود، ارائه کردند. این توانمندسازی بر مبنای تأثیرگذاری عامل‌ها در محیط خود تعریف می‌شوند. چالش اصلی، پیچیدگی مفهومی معیار توانمندسازی و نیاز به محاسبات پیچیده برای ارزیابی آن است. بیشینه‌سازی جریان اطلاعات بالقوه می‌تواند به‌عنوان یک تابع سودمندی سراسری برای هدایت و بهینه‌سازی رفتار جمعی در سیستم‌های چندعامله عمل کند. [۷۱] با تمرکز بر افزایش جریان اطلاعات درون‌گروهی، می‌تواند هماهنگی، تطبیق‌پذیری و عملکرد کلی بهتری داشته باشد. دشواری در تعریف و کمی‌سازی جریان اطلاعات در سناریوهای عملی، و مشکلات مقیاس‌پذیری بالقوه در زمان اعمال به سیستم‌های بسیار بزرگ چالش‌هایی است که این تحقیق با آن مواجه است.

[1] Surprise Memory
[2] Multi-Agent Coordinated Exploration
[3] IncrementalPCA
[4] Grow-When-Required
[5] habituation
[6] Klyubin

جدول ۵: مدل‌های محاسباتی IM: توانمندسازی، کنترل، قدرت

سال انتشار و مرجع	ایده اصلی	مزایا	معایب/چالش‌ها	روش یا الگوریتم
۲۰۱۹ [۷۵]	بررسی استفاده از توانمندسازی در حضور یک سیگنال پاداش بیرونی با این فرض که توانمندسازی می‌تواند عامل‌های RL را به‌وسیله تشویق به‌سوی وضعیت‌های توانمند هدایت کرده تا به راه‌حل‌های رفتاری اولیه خوبی برسند.	عمومی‌سازی اصل بهینگی[1] و گسترش‌های اخیر اطلاعاتی به آن.	نبود داده‌های تجربی برای تأیید اثربخشی و محدودیت در استفاده از این روش در دامنه‌های پیچیده و بزرگ‌تر.	هدایت RL بر اساس توانمندسازی
۲۰۲۰ [۱۲۱]	ارائه یک الگوریتم آنلاین و کارآمد برای محاسبه توانمندسازی به‌عنوان نوعی IM در RL. این روش با کاهش پیچیدگی نمونه و محاسبه، و بهبود پایداری آموزشی نسبت به روش‌های موجود، به آموزش سیاست‌های بهینه بدون نیاز به دانش خاص حوزه و دخالت دستی کمک می‌کند.	کاهش پیچیدگی محاسباتی و نمونه، بهبود پایداری آموزشی، عدم نیاز به دانش خاص حوزه و کاهش نیاز به تنظیمات دستی، امکان آموزش در شرایط واقعی بدون نیاز به بازنشانی‌های مکرر.	نیاز به تنظیم دقیق شبکه‌های عصبی برای محاسبه توانمندسازی و پیچیدگی‌های مرتبط با بهینه‌سازی کانال‌های گوسی، نیاز به بهبود بیشتر در محیط‌های با ابعاد بالا	RL مستقل از مدل
۲۰۲۱ [۱۲۲]	ارائه یک چارچوب به نام «تقویت یادگیری هدف‌محور متغیر» که یادگیری مهارت‌های خودنظارتی مبتنی بر توانمندسازی متغیر را با RL هدف‌محور ترکیب می‌کند. این چارچوب نشان می‌دهد که چگونه می‌توان از یادگیری نمایشی و الگوریتم‌های توانمندسازی برای کشف مهارت‌ها و رسیدن به اهداف استفاده کرد.	امکان یادگیری مهارت‌های متنوع بدون نیاز به طراحی پیچیده پاداش‌ها، بهبود کیفیت اهداف نهان کشف شده، پایداری بیشتر الگوریتم‌ها و امکان استفاده از تکنیک‌های بهینه‌سازی محبوب مانند بازپخش تجربه بازنگری شده	پیچیدگی محاسباتی و نیاز به تنظیم دقیق پارامترها برای بهبود کیفیت نمایشی و مهارت‌ها، نیاز به ارزیابی دقیق‌تر و گسترده‌تر در محیط‌های مختلف.	ترکیبی از RL هدف‌محور و توانمندسازی متغیر[2] است که از حداکثرسازی اطلاعات متقابل[3] و یادگیری نمایشی برای کشف مهارت‌ها و اهداف نهان استفاده می‌کند.
۲۰۲۲ [۱۲۳]	مطالعه رفتار کاوشی انسان‌ها در یک محیط پیچیده و ساختارمند. نویسندگان بررسی می‌کنند که چگونه بازیکنان بازی آنلاین «Little Alchemy 2» با ایجاد عناصری که امکان ایجاد عناصر بیشتر را فراهم می‌کنند، به دنبال توانمندسازی خود هستند. آنها نشان می‌دهند که این نوع کاوش به‌عنوان یک منبع IM می‌تواند در محیط‌های پیچیده که سیگنال‌های پاداش صریحی ندارند، مؤثر باشد.	فراهم‌کردن یک محیط پیچیده و ساختارمند برای مطالعه رفتار کاوشی انسان‌ها، ارائه دیدگاه‌های جدید درباره IM و امکان استفاده از داده‌های بزرگ برای تجزیه و تحلیل دقیق‌تر رفتار کاوشی	پیچیدگی محاسباتی مدل‌ها، نیاز به داده‌های بزرگ و متنوع و احتمال تأثیر ساختارهای از پیش تعریف شده بازی بر رفتار کاوشی بازیکنان	مدل کاوش به‌عنوان توانمندسازی که ترکیبی از استراتژی‌های کاوش مبتنی بر عدم قطعیت و ایجاد عناصری است که امکان ایجاد عناصر بیشتر را فراهم می‌کند.
۲۰۲۳ [۱۲۴]	ارائه یک رویکرد IM مبتنی بر توانمندسازی برای وظایف دست‌کاری رباتیک با پاداش‌های پراکنده است. این رویکرد با ادغام توانمندسازی و کنجکاوی، به ربات‌ها کمک می‌کند تا مهارت‌های مفید دست‌کاری را باوجود پاداش‌های خارجی محدود یاد بگیرند و در مقایسه با سایر روش‌های پیشرفته کاوش درونی، عملکرد بهتری نشان دهند.	امکان یادگیری مهارت‌های دست‌کاری باوجود پاداش‌های پراکنده، بهبود کارایی کاوش و نرخ موفقیت وظایف، و قابلیت ادغام آسان با هر الگوریتم یادگیری تقویتی، توانایی ترکیب با سایر استراتژی‌های کاوش در جهت عملکرد بهتر.	پیچیدگی محاسباتی ناشی از به حداکثر رساندن اطلاعات متقابل شرطی با متغیرهای تصادفی پیوسته و با ابعاد بالا، نیاز به ترکیب با روش‌های کاوش مبتنی بر کنجکاوی در ابتدای فرایند یادگیری برای ارائه تخمین‌های معقول از مقادیر توانمندسازی.	الگوریتم یادگیری تقویتی با رویکرد IM مبتنی بر توانمندسازی با استفاده از حداکثر رساندن وابستگی متقابل بین اعمال و وضعیت‌ها
۲۰۲۳ [۱۲۵]	رفتار اکتشافی انسان‌ها در محیط‌های پیچیده‌تر از آنچه که مطالعات قبلی نشان داده‌اند، غنی‌تر است. به‌خصوص، این مقاله بر IM انسان‌ها برای توانمندسازی و خلق اشیایی که به ایجاد اشیای جدیدتر منجر می‌شوند، تمرکز دارد.	مطالعه رفتارهای اکتشافی پیچیده در محیط‌های غنی و معنادار، ارائه مدل‌های بهتر از IM انسان، و امکان بررسی استراتژی‌های اکتشافی انسانی در سناریوهای واقعی‌تر.	پیچیدگی مدل‌سازی رفتارهای اکتشافی، محدودیت‌های ناشی از طراحی بازی توسط یک فرد و نیاز به بررسی بیشتر برای مقایسه مدل‌های مختلف اکتشافی در محیط‌های دیگر.	استفاده از مدل‌های اکتشافی هدایت‌شده توسط عدم قطعیت و مدل اکتشاف به‌عنوان توانمندسازی که ترکیبی از درک معنایی و IM برای ایجاد اشیای جدید است.

[1] Bellman
[2] variational empowerment
[3] Mutual information

۲۰۲۳ [۱۲۶]	معرفی یک معیار جدید برای اکتشاف در عامل‌های خودمختار با این هدف که عامل‌ها در غیاب یک وظیفه مشخص، اقداماتی را انجام دهند که بیشترین اطلاعات را برای بهبود برآورد توانمندی‌شان فراهم کند. این معیار به‌صورت ترکیبی از جستجوی نوآوری، به حداکثر رساندن تعجب و پیشرفت در یادگیری طراحی شده است.	توانایی ادغام چندین معیار اکتشافی به یک فرمول واحد، کمک به عامل‌ها در شناسایی توانمندی‌هایشان برای تعامل با محیط و جلوگیری از به دام افتادن در مناطق با میزان تصادفی بودن بالا یا محدودیت‌های کنترل	پیچیدگی محاسباتی معیار افزایش توانمندی و نیاز به تقریب‌های هوشمند برای مقیاس‌پذیری به سناریوهای پیچیده‌تر	افزایش توانمندسازی به‌عنوان معیاری برای اکتشاف در عامل‌های خودمختار
۲۰۲۴ [۱۲۷]	استفاده از مفهوم «توانمندسازی انتقال» در RL چندعاملی است. این مفهوم به دنبال اندازه‌گیری تأثیر بالقوه بین اقدامات عامل‌ها است تا فرایند یادگیری را به سمت استراتژی‌های واکنشی به رفتارهای دیگر عامل‌ها هدایت کند. هدف این است که استراتژی‌هایی ایجاد شود که انعطاف‌پذیری بیشتری در برابر تغییرات رفتارهای شرکا داشته باشند.	بهبود عملکرد کلی عامل‌های همکاری‌کننده در وظایف پیچیده‌تر، افزایش انعطاف‌پذیری در برابر تغییرات رفتار شرکا و عدم تداخل با بهره‌برداری از سیاست‌های خوب کشف شده.	پیچیدگی محاسباتی مفهوم توانمندسازی انتقال و نیاز به ارزیابی بیشتر برای تعمیم‌پذیری در سناریوهای مختلف، نیاز به بررسی بیشتر تأثیر این روش در سناریوهای رقابتی و یا مختلط.	روش «توانمندسازی انتقال» با استفاده از اندازه‌گیری ظرفیت کانال بین اقدامات عامل‌ها، به‌عنوان یک مکانیزم پاداش اضافی در فرایند آموزش.

[۷۲] با ترکیب تکنیک‌های استنباط واریاسیونی و یادگیری عمیق، روشی مقیاس‌پذیر برای بیشینه‌سازی اطلاعات متقابل به‌منظور پشتیبانی از یادگیری تقویتی انگیزش درونی معرفی کرده است. مقیاس‌پذیری به مسائل با ابعاد بالا، کاهش پیچیدگی محاسباتی نسبت به روش‌های سنتی، کاربرد در تنظیمات گسسته و پیوسته، و توانایی یادگیری مستقیم از ورودی‌های بصری بدون نیاز به مدل مولد محیط از مزایای این تحقیق به شمار می‌رود. احتمال دشواری در برآورد دقیق اطلاعات متقابل در محیط‌های بسیار پیچیده، اتکا به تقریب‌هایی که ممکن است همیشه تمامی جزئیات داده را به‌درستی بازتاب ندهند، و نیاز به منابع محاسباتی زیاد برای کاربردهای یادگیری عمیق، چالش‌های این تحقیق است. گرگور[1] و همکاران [۷۳] روش جدید یادگیری تقویتی بدون نظارت به‌منظور کشف مجموعه‌ای از گزینه‌های درونی موجود برای یک عامل را معرفی کردند که این مجموعه با حداکثرکردن تعداد حالت‌های مختلفی که یک عامل می‌تواند به طور قابل‌اعتماد به آن‌ها برسد، به دست می‌آید. این کار با اندازه‌گیری اطلاعات متقابل بین‌مجموعه‌ای از گزینه‌ها و حالت‌های پایان گزینه‌ها انجام می‌شود. ویژگی مهم این روش، مقیاس‌پذیری بالای آن است. با مراجعه به جدول ۵ می‌توان سایر روش‌های اخیر را به همراه بررسی آن‌ها مشاهده کرد.

۵-۵- تنهایی، وابستگی اجتماعی

شاید بتوان گفت نیاز به ارتباطات اجتماعی سخت‌ترین انگیزش درونی برای پیاده‌سازی به‌عنوان ارزیابی‌های محاسباتی باشد. [۸۱] رویکرد جدیدی مبتنی بر سائق‌ها را برای تولید و ترکیب احساسات مصنوعی در فرایند تصمیم‌گیری عامل‌های خودکار (فیزیکی و مجازی) ارائه کرده است. در این تحقیق هر احساس مصنوعی به طور جداگانه مورد بررسی قرار می‌گیرد و این اجازه را می‌دهد که سیستم تصمیم‌گیری بر مبنای تجربیات خود به یادگیری احساسات و رفتارهای بهینه بپردازد. البته باید اشاره نمود که تعیین روش‌های انتشار احساسات مصنوعی و نقش آن‌ها در سیستم تصمیم‌گیری ممکن است پیچیده باشد. دورنر[2] و همکاران در [۸۲] به معرفی نظریه PSI که یک معماری محاسباتی فرمال از فرایندهای روان‌شناختی انسانی است پرداخته‌اند. این نظریه مبتنی بر معماری شناختی DARPA، بر خلاف نظریات موجود دیگر، نه‌تنها فرایندهای شناختی را مدل می‌کند، بلکه فرایندهای انگیزشی و احساسی و تعاملات آن‌ها را نیز مدل‌سازی می‌کند.

بهبود مکانیزم‌های پاداش در عامل RL با استفاده از عامل اجتماعی ـ عاطفی الهام‌گرفته از احساسات انسانی و تعاملات اجتماعی در [۸۳] صورت‌گرفته است. این روش به‌منظور ایجاد عامل RL انعطاف‌پذیرتر و سازگارتر با عملکرد بهتر در محیط‌های پویا و غیرقابل‌پیش‌بینی ارائه شده است. همچنین به کاهش نیاز به تنظیم دقیق توابع پاداش برای محیط‌های خاص به‌عنوان دیگر مزیت آن می‌توان اشاره کرد. از طرفی پیچیدگی طراحی مکانیزم‌های پاداش اجتماعی ـ عاطفی، نیاز به منابع محاسباتی گسترده برای شبیه‌سازی‌ها و دشواری‌های احتمالی در تعمیم این روش به همه انواع مسائل RL چالش‌هایی است که این روش با آن‌ها مواجه است. در ادامه در جدول ۶ به معرفی برخی دیگر از روش‌های مرتبط اخیر با این دسته پرداخته شده است.

۵-۶- هدف

پژوهش‌های اخیر روی به‌کارگیری انگیزش درونی برای تولید خودمختار و/یا انتخاب اهدافی که می‌توانند اخذ مهارت را هدایت کنند، تمرکز کرده‌اند [۸۴،۸۵]. این دیدگاه مزایای متنوعی را فراهم می‌کند، برای مثال، امکان بهینه‌سازی فرآیندهای یادگیری در فضاهای عمل با ابعاد بالا با چندین ربات کنترل کننده [۵۴،۵۵] فراهم می‌شود. بارانز[3] و همکاران [۸۶] معماری «تولید هدف خود تطبیق دهنده هوشمند کنجکاو تطبیق پذیر با هدف توسعه مکانیسم فعال برای یادگیری

[1] Gregor
[2] Dörner
[3] Baranes

مدل‌های معکوس» در ربات‌ها که از انگیزش درونی الهام گرفته شده است، را ارائه کرده‌اند. این معماری اجازه می‌دهد تا ربات به طور فعال و بهینه توزیع مهارت‌ها/سیاست‌های پارامتری را که رویکردهای مختلف یکسانی را برای حل کردن توزیع وظایف/اهداف پارامتری مواجه می‌شوند یاد بگیرد. در نتیجه می‌تواند موثرتر و کارآمدتر از روش‌های سنتی یادگیری فعال در رباتیک باشد. البته زمان برای جمع‌آوری نمونه های یادگیری در دنیای واقعی محدود است و مدل‌های معکوس یادگیری باید به طور خودکار و به صورت تدریجی توسط ربات‌ها جمع آوری شود.

در [۵۶] معماری انگیزش درونی ارائه شده است که قادر است به طور خودکار ۱) تغییرات در محیط را کشف کند، ۲) نمایش اهداف متناظر با این تغییرات را شکل دهد، ۳) هدف را بر اساس انگیزش درونی انتخاب کند، ۴) منابع محاسباتی مناسب را برای دستیابی به هدف انتخاب شده انتخاب کند، ۵) پیگیری دستیابی به هدف انتخاب شده را انجام دهد و ۶) یک سیگنال یادگیری را خودکار تولید کند زمانی که هدف انتخاب شده با موفقیت دستیابی شود. باید اشاره کرد که نیاز به تنظیمات دقیق یکی از چالش‌های اصلی این معماری است. [۸۷] سیستمی را معرفی نموده است که با استفاده از انگیزش درونی، نتایج و مهارت‌های جدید را کشف کرده و سپس از مکانیسم‌های مبتنی بر اهداف برای دستیابی به اهداف بیرونی کاربر استفاده می‌کند. وابستگی به تکنولوژی تصویری و ماهیت ماژولار مهارت‌ها موانعی است که این سیستم باید برای عبور از آن‌ها برنامه‌ریزی کند. در [۴۰] چارچوب مهمی متناسب با یادگیری چندهدفه بدون نظارت و الگوریتم‌های مبتنی بر اکتشاف هدف‌محور انگیزش درونی ارائه نمود که بعداً مورد رجوع و استفاده بسیاری از محققان دیگر قرار گرفت و توسعه‌های مختلفی بر روی آن انجام شد. ایده اصلی این تحقیق، مدل‌سازی یادگیری خودکار و پیوسته در کودکان با استفاده از فرایندهای اکتشاف هدف‌محور انگیزش درونی و یادگیری برنامه آموزشی خودکار است. امکان کشف مهارت‌های پیچیده با کمک یک برنامه آموزشی خودکار و اکتشاف فضاهای اهداف درونی مختلف، مزیت‌های مهم این روش است. محدودیت‌های زمانی، وجود نویز پیچیده و نیاز به پیاده‌سازی واقعی در دنیای رباتیک مسائلی است که این روش باید روی آن‌ها کار کند. سایر روش‌های ارائه شده در این دسته‌بندی در جدول (۷) با جزئیات بیشتری مورد کندوکاو قرار گرفته است.

۵-۷- غافلگیری، کنجکاوی، (عدم) قطعیت

غافلگیری یعنی عامل از نتیجه‌هایی که برخلاف درک او از جهان باشد، هیجان‌زده می‌شود [۷۶]. این ارزیابی باعث می‌شود عامل به دنبال تجربیات جدید باشد. به‌عنوان مثال، عامل می‌تواند یک پاداش اضافه را بگیرد اگر وضعیت‌های دیده نشده قبلی را ببیند، یا خروجی عملی را مشاهده کند که پیش‌بینی نشده باشد، یعنی در تضاد با آنچه تابه‌حال فهمیده است، باشد. [۷۷] روشی را برای بهبود اکتشاف در RL با استفاده از پاداش‌های اکتشافی مشتق شده از یک مدل یاد گرفته شده از دینامیک سیستم پیشنهاد کرده است. این مدل که با یک شبکه عصبی پارامتریزه شده است، پاداش‌ها را بر اساس خطاهای پیش‌بینی اختصاص می‌دهد و عامل را به اکتشاف حالت‌های جدید که به طور قابل‌توجهی با پیش‌بینی‌هایش متفاوت هستند، تشویق می‌کند. اکتشاف مقیاس‌پذیر و کارآمد در وظایف پیچیده و با ابعاد بالا، عملکرد بهتر در حوزه‌های چالش‌برانگیزی مانند بازی‌های آتاری و سرعت یادگیری سریع‌تر، عملکرد بهتر نسبت به استراتژی‌های ساده‌تری مانند epsilon_greedy، کاربرد بهتر برای مسائل بزرگ‌مقیاس نسبت به رویکردهای بیزین از مزیت‌های این روش است. منتها این روش در مدیریت محیط‌هایی با دینامیک بسیار تصادفی جایی که خطای پیش‌بینی ممکن است به طور دقیق، تازگی را منعکس نکند، دچار چالش می‌شوند.

بلمر[1] و همکاران [۷۸] با استفاده از الگوریتم RL، روشی را معرفی کردند که اکتشاف مبتنی بر شمارش و انگیزش درونی در RL از طریق معرفی یک شمارش کاذب مشتق شده از یک مدل چگالی ادغام شده است. این روش اکتشاف مبتنی بر شمارش را به محیط‌های غیر جدولی تعمیم می‌دهد و بهره‌وری اکتشاف را در سناریوهای پیچیده؛ مانند بازی‌های Atari 2600 بهبود می‌بخشد. این روش، بهره‌وری اکتشاف در فضای حالت‌های بزرگ و پیچیده، را بهبود بخشیده است. پیچیدگی محاسباتی برآورد مدل چگالی، مشکلات احتمالی در انتخاب مدل‌های چگالی مناسب برای محیط‌های مختلف، و نیاز به تنظیم و آزمایش گسترده چالش‌هایی است که این روش با آن‌ها مواجه است.

ارائه روشی برای اکتشاف مبتنی بر کنجکاوی در RL با استفاده از پیش‌بینی خود نظارتی توسط پاتاک[2] و همکاران [۷۹] ارائه شده است. در این روش، کنجکاوی به‌عنوان خطا در پیش‌بینی نتیجه اقدامات عامل در یک فضای ویژگی‌های بصری یاد گرفته شده، فرمول‌بندی می‌شود. این رویکرد هدفش مدیریت فضای حالت‌های با ابعاد بالا مانند تصاویر است و با پاداش‌دادن به عامل برای خطاهای پیش‌بینی، آن را به اکتشاف تشویق می‌کند. اکتشاف کارآمد در فضای حالت‌های با ابعاد بالا، جلوگیری از نیاز به پیش‌بینی ورودی‌های حسی خام مانند پیکسل‌ها به طور مستقیم، کمک به عامل برای یادگیری مهارت‌های قابل‌تعمیم حتی در نبود پاداش‌های بیرونی باعث شده است که این تحقیق مورد استناد زیادی قرار گیرد. البته باید اشاره کرد که این روش ممکن است با سناریوهایی که نیاز به دنباله‌های بسیار خاصی از اعمال دارند و پیش-بینی آن‌ها دشوار است، دچار مشکل شود و اکتشاف را محدود کند.

مقاله [۷۶] پیشنهاد استفاده از شگفتی به‌عنوان شکلی از انگیزش درونی در DRL برای بهبود اکتشاف در محیط‌هایی با پاداش‌های کم را ارائه نموده است. این کار با محاسبه IR بر اساس KL-divergence بین احتمالات انتقال واقعی و یک مدل یاد گرفته شده انجام می‌شود که عامل را به اکتشاف حالت‌هایی که از پیش‌بینی‌هایش منحرف می‌شوند، تشویق می‌کند. اکتشاف کارآمد در وظایف کنترل با ابعاد بالا و پیوسته، مقیاس-پذیری و توانایی مدیریت محیط‌هایی با پاداش‌های بسیار کم، کاهش هزینه‌های محاسباتی در مقایسه با برخی از تکنیک‌های انگیزش درونی

[1] Bellemare

[2] Pathak

پیشرفته نکات مثبت این تحقیق به شمار می‌رود. منتها پیچیدگی یادگیری دقیق مدل انتقال و احتمال سربار محاسباتی در محیط‌هایی با دینامیک بسیار تصادفی چالش جدی است. همچنین ممکن است تقریب‌های استفاده شده برای KL-divergence همیشه به طور کامل شگفتی واقعی را ثبت نکنند. در جدول ۸ به معرفی برخی دیگر از روش‌های مرتبط اخیر با این دسته پرداخته شده است.

جدول (۶): مدل‌های محاسباتی IM: تنهایی، وابستگی اجتماعی

سال انتشار و مرجع	ایده اصلی	مزایا	معایب/چالش‌ها	روش یا الگوریتم
۲۰۱۸ [۱۲۹]	معرفی مدلی از احساسات مصنوعی برای تطبیق و هماهنگی ضمنی در سیستم‌های چند رباتی. احساسات مصنوعی به دو صورت مختلف عمل می‌کنند: به‌عنوان تنظیم‌کننده رفتار فردی و وسیله‌ای برای هماهنگی اجتماعی. پیاده‌سازی برای یک وظیفه ناوبری شده تا نشان دهد چگونه هماهنگی از طریق اضافه‌کردن احساسات مصنوعی به یک چارچوب ناوبری موجود به طور مؤثری به دست می‌آید.	بهبود عملکرد سیستم‌های چند رباتی در جلوگیری از بن‌بست‌ها، افزایش کارایی ناوبری در وظایف بحرانی زمانی، کمک به ربات‌های با سنسورهای خراب، عدم نیاز به تنظیمات دستی پیچیده یا تکنیک‌های مصرف‌کننده پهنای باند برای اشتراک اطلاعات	پیچیدگی در مدل‌سازی و پیاده‌سازی احساسات مصنوعی و نیاز به اعتبارسنجی در سناریوهای واقعی، تحقیقات بیشتر پیرامون ارزیابی قابلیت استفاده عمومی واژگان احساسی پیشنهاد شده	معماری کنترل سطح بالا بر اساس احساسات مصنوعی که شامل نمایندگی و دینامیک احساسات، تنظیم رفتار، و اشتراک اطلاعات برای هماهنگی است.
۲۰۲۰ [۱۳۰]	توسعه چارچوبی برای انگیزش اجتماعی در تعامل انسان ـ ربات که در آن یک عامل خودمختار نه‌تنها توسط محیط خود، بلکه بر اساس حالت ذهنی انسان که در همان محیط فعالیت می‌کند نیز پاداش می‌گیرد. این چارچوب به طور خاص در محیط‌های نیمه مشاهده شده پیاده‌سازی می‌شود که در آن تابع پاداش عامل به باور انسان بستگی دارد.	ایجاد رفتارهای انطباقی و هماهنگی مؤثر بین ربات و انسان، استفاده از انگیزش اجتماعی برای رسیدن به هماهنگی به‌جای استفاده از تکنیک‌های پیچیده و مصرف‌کننده پهنای باند برای اشتراک اطلاعات	مقیاس‌پذیری محدود مدل باور که برای دامنه‌های پیچیده‌تر نیاز به بهبود دارد. چالش نمایش دقیق‌تر باورها و استفاده از روش‌های برنامه‌ریزی برای بهبود مقیاس‌پذیری	چارچوب فرایند تصمیم‌گیری مارکوف نیمه مشاهده شده[1] عمومی که برای مسئله انگیزش اجتماعی در تعامل انسان ـ ربات استفاده می‌شود.
۲۰۲۱ [۱۳۱]	پیشنهاد یک مدل محاسباتی تصمیم‌گیری انسانی که رفتارهای ناشی از احساسات را در بر می‌گیرد. این مدل می‌تواند یک عمل منطقی یا غیرمنطقی را بر اساس توزیع احتمالاتی تعیین کند که از ترکیب یک سیاست بهینه از فرایند تصمیم‌گیری مارکوف نیمه مشاهده شده و توزیع احتمالاتی تکامل‌یافته توسط دینامیک جدید احساسات به دست می‌آید.	توانایی مدل‌سازی و پیش‌بینی رفتارهای انسانی ناشی از احساسات، امکان ارائه اقدامات پیشگیرانه برای جلوگیری از سناریوهای منفی مانند قتل، و بهبود سطح هوشمندی و باورپذیری شخصیت‌های هوش مصنوعی است.	نیاز به بهبود مقیاس‌پذیری مدل، ضرورت بهبود روش‌های تخمین و یادگیری پارامترهای مدل، و نیاز به اعتبارسنجی مدل با استفاده از سوابق مختلف فرایند تصمیم‌گیری انسانی	فرایند تصمیم‌گیری مارکوف نیمه مشاهده شده برای تصمیم‌گیری منطقی، معماری پویا برای محاسبه توزیع احتمالاتی ناشی از احساسات، و یک فرایند انتخاب عمل نهایی با استفاده از روش حد آستانه
۲۰۲۱ [۱۳۲]	احساسات می‌توانند به‌عنوان پدیده‌ای پدیدار از مکانیزم تنظیم انرژی عصبی ـ محاسباتی یک عامل شناختی در یک وظیفه‌ی تصمیم‌گیری در نظر گرفته شوند.	به عامل اجازه می‌دهد رفتار خود را به‌گونه‌ای تنظیم کند که انرژی عصبی ـ محاسباتی موردنیاز را برای انجام وظیفه‌ای به حداقل برساند.	نیاز به مدل‌سازی و اندازه‌گیری دقیق احساسات	RL بدون مدل برای انتخاب بهینه اعمال
۲۰۲۳ [۱۳۳]	نمایش احساسات مبتنی بر برچسب‌ها که احساسات و شدت آن‌ها را با استفاده از یک نمایش متغیر، ارائه کرده و تغییر می‌دهد. همچنین، فرایندی برای فازی‌سازی تعریف شده که بر اساس یک جفت از مقادیر لذت و برانگیختگی، برچسب احساسی را در یک زبان خاص تعیین می‌کند؛ بنابراین، با استفاده از این مدل، عامل می‌تواند احساسات خود را با همان اصطلاحات فازی انسان‌ها ابراز کند. ادغام این مدل در معماری عامل دارای عاطفه GenIA3 تا بهبود شبیه‌سازی فرایندهای عاطفی مختلف. بررسی اینکه آیا نمایش احساسات با معرفی بعد سوم (سلطه) برای شناسایی بهتر احساساتی مانند ترس یا عصبانیت و تجزیه و تحلیل تأثیر این بعد بر رفتار اجتماعی قابل‌بهبود است یا خیر.	احساسات به‌صورت یک فضای پیوسته چندبعدی نمایش داده می‌شوند که این امکان را فراهم می‌آورد که اطلاعات بیشتری در مورد احساسات، مانند شدت و نزدیکی بین احساسات، ذخیره شود.	نیاز به انجام آزمایش‌هایی در محیط‌ها و فرهنگ‌های مختلف برای تطبیق مدل‌های احساسات	دو روش مبتنی بر منطق فازی برای نمایش و بیان احساسات

[1] Partially Observable Markov Decision Process

جدول ۷: مدل‌های محاسباتی IM: هدف

سال انتشار و مرجع	ایده اصلی	مزایا	معایب/چالش‌ها	روش یا الگوریتم
۲۰۱۸ [۸۸]	معرفی الگوریتم‌های اکتشاف هدف‌محور الهام‌گرفته شده از IM به‌منظور توانمند کردن ماشین‌ها به کشف مجموعه‌ای از سیاست‌ها که تنوعی از اثرات در محیط‌های پیچیده ایجاد می‌کنند.	این روش به ماشین‌ها امکان می‌دهد تا سیاست‌هایی را کشف کنند که تنوعی از اثرات را در محیط‌های پیچیده ایجاد کنند.	نیاز به نمایش‌های از قبل طراحی شده برای اکتشاف هدف‌محور	معماری اکتشاف هدفمند با IM با یادگیری بدون نظارت فضای هدف
۲۰۱۹ [۸۹]	با الهام از تنوع در تجربیات هدف در انسان‌ها، در RL چندهدفه، عامل یاد می‌گیرد چگونه با یک سیاست شرطی به هدف‌های مختلف برسد. معیار پاداش شامل بازگشت مورد انتظار و دستیابی به اهداف متنوع بیشتر است.	کمک به یافتن تنوع بیشتر در اهداف عامل هم‌زمان با بیشینه‌کردن بازگشت مورد انتظار	در حین تکرار تجربه، مسیرهای نمونه‌برداری شده نسبت به سیاست‌های رفتاری، جهت‌گیری دارند.	شبکه‌های عصبی عمیق و الگوریتم DDPG[1] [۱۰۶]
۲۰۱۹ [۱۳۴]	معرفی الگوریتمی نوآورانه به نام «تولید هدف از دیدگاه واپس‌نگری» که اهداف واپس‌نگری ارزشمندی را تولید می‌کند که برای یک عامل در مدت کوتاه قابل‌دستیابی هستند و همچنین در طولانی‌مدت برای راهنمایی عامل برای رسیدن به هدف واقعی مناسب هستند.	افزایش قابلیت اطمینان و کارایی در استفاده از الگوریتم‌های RL برای مسائل هدف‌گرا	پیچیدگی در یافتن اهدافی مناسب برای ایجاد سیاست برای عامل	بازپخش تجربه واپس‌نگر [2]
۲۰۲۰ [۱۳۵]	معرفی الگوریتم جدیدی به نام «تولید هدف از دیدگاه واپس‌نگری مبتنی بر گراف» که اهداف واپس‌نگری را بر اساس کوتاه‌ترین فاصله‌ها در یک گراف جلوگیری از موانع انتخاب می‌کند که نمایانگر نمایش گسسته محیط است.	افزایش قابلیت اطمینان و کارایی در استفاده از الگوریتم‌های RL برای مسائل کنترل رباتیکی با موانع	پیچیدگی در ارزیابی عملکرد الگوریتم در محیط‌های واقعی	الگوریتم تولید هدف از دیدگاه واپس‌نگری مبتنی بر گراف (RL)
۲۰۲۱ [۱۳۶]	ارائه چارچوبی جدید برای آموزش یک سیاست سطح بالا با فضای عمل کاهش‌یافته توسط نقاط علامت زده، به نام «HRL راهنمایی شده توسط علامت‌ها»	افزایش کارایی در اکتشاف و ایجاد زیرهدف‌های معنادار برای آموزش در RL	افزایش زمان موردنیاز برای برنامه‌ریزی در هر مرحله آموزش	HRL هدف‌محور[3]
۲۰۲۲ [۴۰]	ارائه رویکرد الگوریتمی به نام فرایندهای اکتشاف هدف IM در یادگیری خودکار برنامه‌ریزی شده	ایجاد یک سیستم یادگیری خودکار برنامه‌ریزی شده است که توانایی کاوش و یادگیری خودکار را به ماشین‌ها می‌دهد.	نیاز به سیستم درک و دریافت حسی مؤثر برای ماشین‌ها	چارچوب رسمی برای معماری الگوریتمی به نام فرایندهای اکتشاف هدفمند با RL (IMGEP[4])
۲۰۲۳ [۱۳۷]	این مقاله به دنبال حل چالش اکتشاف در RL در دامنه‌های بسیار بزرگ است و از رویکرد مبتنی بر هدف برای کاوش استفاده می‌کند.	امکان تعیین یک فضای پیش‌هدف بزرگ اما معنی‌دار و تبدیل آن به یک مجموعه کوچک‌تر از اهداف مفید	چالش امکان ادغام دو عنصر اصلی فضای پیش‌هدف و ارزیاب پیش‌هدف به‌صورت مؤثر	RL هدف‌محور
۲۰۲۳ [۱۳۸]	توسعه یک روش مؤثر برای یادگیری هدفمند در محیط‌هایی با تعداد بالایی از داده‌های آموزشی بدون پاداش	امکان حل وظایف با زمان طولانی، استفاده راحت از داده‌های بدون برچسب	محاسبه نادرست تابع ارزش ـ عمل برای وظایف با پویایی نامشخص محیط	RL هدف‌محور آفلاین
۲۰۲۴ [۱۳۹]	معرفی چارچوب نوآورانه‌ای به نام GEASD[5] که طراحی شده است تا الگوهای ساختاری محیط را از طریق توزیع مهارت‌های تطبیقی در طول فرایند یادگیری گرفته و بررسی کند.	بهبود کیفیت اکتشاف و پیشبرد در وظایفی که شامل پاداش‌های کم و زمان‌های طولانی است.	انعطاف‌پذیری در استفاده از مهارت‌ها در وضعیت‌های ناشناخته	توزیع مهارت‌های تطبیقی بر اساس توزیع بولتزمن از توابع ارزیابی مهارت‌ها برای تسهیل اکتشاف عمیق

[1] Deep Deterministic Policy Gradient
[2] Hindsight Experience Replay (HER)
[3] Goal-conditioned hierarchical reinforcement learning
[4] Intrinsically Motivated Goal Exploration Processes
[5] Goal Exploration via Adaptive Skill Distribution

۶- یادگیری ماشین و انگیزش درونی

انگیزش درونی در یادگیری ماشین مفهومی است که از کنجکاوی انسان‌گونه الهام‌گرفته شده است، جایی که سیستم‌ها توسط پاداش‌های داخلی برای کاوش، یادگیری و تطبیق هدایت می‌شوند، بدون اینکه صرفاً به انگیزش‌های بیرونی متکی باشند. این رویکرد الگوریتم‌های یادگیری ماشین را، به‌ویژه در محیط‌هایی که پاداش‌های بیرونی خلوت یا تعریف آن‌ها دشوار است، تقویت می‌کند. با تعبیه مکانیسم‌هایی مانند تشخیص نوآوری، خطاهای پیش‌بینی یا اکتشاف مبتنی بر تعجب، مدل‌های یادگیری ماشین قادر به کاوش خودکار داده‌ها و یادگیری ویژگی می‌شوند. به‌عنوان مثال همان‌طور که اشاره شد، [۷۹] ماژول کنجکاوی ذاتی را معرفی می‌کند که اکتشاف را از طریق خطای پیش‌بینی تشویق می‌کند و به عامل‌ها اجازه می‌دهد تا استراتژی‌های جدیدی را در تنظیمات پاداش خلوت کشف کنند. این پیشرفت‌ها انگیزش درونی را به یک جنبه حیاتی در ساخت سیستم‌های یادگیری ماشین تطبیقی و قوی تبدیل می‌کند که می‌توانند در بین وظایف و محیط‌ها بهتر تعمیم یابند.

یادگیری با ناظر: نوعی از یادگیری ماشین است که در آن به الگوریتم، داده‌های ورودی و برچسب‌های خروجی صحیح مرتبط با آن داده می‌شود. هدف این روش، آموزش الگوریتم برای پیش‌بینی برچسب‌های خروجی برای داده‌های جدید و دیده نشده است. در این دسته، از انگیزش درونی برای ایجاد سیستم‌هایی استفاده می‌شود که می‌توانند به طور خودکار ویژگی‌های داده‌ای را شناسایی کنند که منجر به تعمیم و کارایی بهتر می‌شوند، بدون اینکه صرفاً به داده‌های برچسب‌دار متکی باشند. این رویکرد مدل‌ها را تشویق می‌کند تا بر یادگیری نمایش‌هایی تمرکز کنند که می‌توانند عملکرد را در وظایف مختلف بهبود بخشند، با تحریک اکتشاف مبتنی بر کنجکاوی در خود داده‌ها. به‌عنوان مثال، در طبقه‌بندی تصویر، انگیزش درونی به مدل‌ها کمک می‌کند با پیش‌بینی تحولات یا مناطق ماسک شده، همان‌طور که در مدل‌هایی مانندSimCLR [۱۵۵] دیده می‌شود، یاد بگیرند.

یادگیری بدون ناظر: در آن به الگوریتم تنها داده‌های ورودی داده می‌شود و الگوریتم باید به‌تنهایی الگوها و ساختارهای نهفته در داده‌ها را کشف کند. این نوع یادگیری برای کشف گروه‌های طبیعی داده‌ها (خوشه‌بندی)، کاهش ابعاد داده‌ها و شناسایی ناهنجاری‌ها کاربرد دارد. در یادگیری بدون ناظر، انگیزش درونی به مدل‌ها کمک می‌کند تا بدون نظارت خارجی داده‌ها را کاوش کنند و به آن‌ها اجازه می‌دهد تا ساختارها یا الگوهای پنهان را کشف کنند. این نوع انگیزه مبتنی بر پاداش‌های داخلی مانند نوآوری یا خطاهای پیش‌بینی است که مدل را هدایت می‌کند تا نمایش‌هایی را بیاموزد که با برچسب‌های از پیش تعریف‌شده مغرضانه نیستند؛ مانند کاربرد خودرمزگذارها یا شبکه‌های مولد متخاصم (GAN) . این مدل‌ها با هدف تولید یا بازسازی داده‌ها و پالایش درک خود از توزیع‌های پیچیده داده از طریق اکتشاف مبتنی بر کنجکاوی هدایت می‌شوند. تکنیک‌هایی مانند خودرمزگذارهای تغییری از انگیزش درونی برای تشویق یادگیری ویژگی بهتر و قابلیت‌های مولد استفاده کرده‌اند. به‌عنوان نمونه رمزگذاری پاداش عملکردی[1]، از یک خودرمزگذار تغییرپذیر مبتنی بر ترنسفورمر[2] برای یادگیری نمایش‌های عملکردی از وظایف دلخواه از طریق رمزگذاری نمونه‌های حالت - پاداش استفاده می‌کند. این روش به عامل اجازه می‌دهد با استفاده از مجموعه‌ای از توابع پاداش بدون نظارت پیش‌آموزش شود و با حداقل داده‌های دارای برچسب پاداش به وظایف جدید انطباق پیدا کند [۱۵۶].

یادگیری خودنظارت شده: دسته‌ای از یادگیری ماشین است که در آن الگوریتم با ایجاد وظایف نظارتی از خود داده‌ها، خود را آموزش می‌دهد. به‌عنوان مثال، پیش‌بینی بخش گم‌شده یک تصویر یا پر کردن کلمات گم‌شده در یک جمله، از جمله وظایف خودنظارتی هستند. الگوریتم‌های این دسته، با استفاده از انگیزش درونی مدل‌ها را برای پیش‌بینی بخش‌های خاصی از داده‌ها از بخش‌های دیگر آموزش می‌دهد و از خود داده‌ها به‌عنوان سیگنال نظارتی استفاده می‌کند. این نوع یادگیری با انگیزش درونی مدل برای حل چالش‌های خودتنظیم هدایت می‌شود که منجر به درک عمیق‌تر از داده‌ها می‌شود. یادگیری خودنظارت‌شده در NLP با مدل‌هایی مانندBERT [۱۵۴] و GPT [۱۵۷] محبوبیت پیدا کرده است، جایی که مدل برای پیش‌بینی کلمات ماسک‌شده یا جملات بعدی آموزش داده می‌شود. این شکل از یادگیری به طور قابل‌توجهی از انگیزش درونی بهره‌مند می‌شود؛ زیرا به مدل‌ها اجازه می‌دهد وابستگی‌های داده‌ای و روابط معنایی را بدون حاشیه‌نویسی انسانی کاوش کنند.

یادگیری نیمه‌نظارت‌شده: ترکیبی از یادگیری با ناظر و بدون ناظر است. در این روش، الگوریتم هم به داده‌های برچسب‌گذاری شده و هم به داده‌های بدون برچسب دسترسی دارد. هدف این روش، استفاده از هر دو نوع داده برای بهبود عملکرد مدل است. در این روش، انگیزش درونی به مدل‌ها کمک می‌کند تا از داده‌های برچسب‌دار و بدون برچسب استفاده بهتری کنند. مکانیسم انگیزشی مدل‌ها را تشویق می‌کند تا با جستجوی الگوها یا ساختارهایی که آموزنده هستند و می‌توانند با داده‌های برچسب‌دار موجود همسو شوند، داده‌های بدون برچسب را کاوش و یاد بگیرند .یکی از کاربردها مدل Mean Teacher است که در آن یک مدل دانش‌آموز از یک مدل معلم که از طریق میانگین متحرک‌نمایی وزن‌های خود به‌روزرسانی می‌شود، یاد می‌گیرد. دانش‌آموز برای تطبیق پیش‌بینی‌های معلم انگیزه دارد و این امر باعث بهبود خود و استفاده بهتر از داده‌های بدون برچسب برای افزایش کارایی یادگیری می‌شود [۱۵۸].

یادگیری تقویتی: یکی از شاخه‌های مهم یادگیری ماشین است که به عامل‌ها می‌آموزد تا در تعامل با محیط، تصمیماتی بگیرند که منجر به حداکثرکردن پاداش تجمعی شوند. با توجه به رشد چشمگیر تحقیقات و توسعه‌های صورت‌گرفته در حوزه یادگیری تقویتی به همراه انگیزش

[1] Functional Reward Encoding FRE
[2] transformer-based variational auto-encoder

درونی و کاربردهای متنوع آن در زمینه‌هایی مانند بازی‌های رایانه‌ای، رباتیک و کنترل فرایندها، نیاز به بررسی جداگانه و تخصصی این حوزه امری ضروری است. بدین منظور، در بخش‌های آتی به طور مفصل به تبیین مبانی و کاربردهای متنوع یادگیری تقویتی پرداخته خواهد شد.

۶-۱- یادگیری تقویتی و انگیزش درونی

یکی از اهداف اصلی حوزه هوش مصنوعی تولید عامل‌هایی کاملاً خودمختار است که با محیط به‌منظور یادگیری رفتارهای بهینه تعامل داشته و از طریق آزمون‌وخطا و در طول زمان، آن را بهبود بخشند [۹۲]. برای رسیدن به این هدف، الگوریتم RL عملکردی موفق داشته است [۹۳،۹۴]. تصور رایج این است که چارچوب محاسباتی RL تنها می‌تواند با انگیزش بیرونی مرتبط باشد، زیرا عامل RL دارای یک کانال ورودی متمایز است که سیگنال پاداش را از محیط خارجی آن دریافت می‌کند. بر خلاف این دیدگاه، این ادراک نتیجه عدم درک کاملِ ماهیت RL است که در حقیقت برای متحد شدن با انگیزش درونی مناسب است. بعلاوه ترکیب عناصر محاسباتی با سیستم‌های RL راه را برای توسعه بیشتر سیستم‌های یادگیری ماشین باز می‌کند [۲۲].

تقویت پلی بین هوش مصنوعی و روان‌شناسی است و مشارکتی برای متحد کردن هر دو دیدگاه انگیزش درونی و انگیزش بیرونی است [۲۲،۲۶]. محیط شامل منتقدی است که عامل را در هر گام با ارزیابی (نمره عددی) رفتار فعلی عامل، تجهیز کرده است. منتقد وضعیت‌های محیط (یا احتمالا جفت وضعیت-عمل یا حتی سه تایی وضعیت-عمل-وضعیت بعدی) را به سیگنال‌های پاداش عددی نگاشت می‌کند [۲۲]. مسئله‌ی پیش روی RL این است که، این روش یادگیری در محیطی که پاداش خلوت است و زمانی که کسب آن‌ها نیاز به هماهنگی توالی-های طولانی مدت دارد، دچار چالش جدی می‌شود. از طرفی اکثر کاربردهای دنیای واقعی یا نزدیک به آن، دارای این چنین شرایطی هستند. مشکل دیگر، مسئله تعمیم است [۹۶]. بعلاوه باید اشاره نمود چالش فضای حالت بی‌انتها/بدون محدوده باعث می‌شود اکتشاف به یک مسئله بسیار سختی مبدل گردد. حتی اگر جهان گسسته شود، اما با تعداد بسیار زیادی از وضعیت مواجه خواهیم بود [۴۱].

۶-۲- یادگیری تقویتی

عامل یادگیری همواره به دنبال انجام عملی است که بر طبق رفتارهای گذشته، انتظار عکس‌العمل مطلوب‌تر را از محیط داشته باشد [۹۵]. در یک مسئله یادگیری تقویتی با عاملی روبرو هستیم که از طریق سعی و خطا با محیط تعامل کرده و یاد می‌گیرد تا عملی را برای رسیدن به هدف انتخاب نماید. این الگوی رفتاری، به‌صورت محاسباتی فرموله شده و برای کنترل فرایندها و سیستم‌ها به کار برده می‌شود. به بیان ریاضی، یادگیری تقویتی، نگاشتی از وضعیتی که محیط در آن قرار دارد به رفتاری است که عامل باید انجام دهد، با هدف بیشینه‌کردن درجه رضایتمندی از محیط که می‌تواند به‌صورت تابعی از وضعیتی که محیط در آن قرار دارد تعریف شود. تعریف فوق در بیان کنترلی به‌صورت زیر تغییر می‌یابد: یادگیری تقویتی نگاشتی است از بردار حالت سیستم تحت کنترل به خروجی کنترل‌کننده با هدف کمینه‌کردن تابع هزینه [۹۵]. در مدل یادگیری تقویتی استاندارد، یک عامل از طریق دریافت‌ها و اعمال با محیط ارتباط برقرار می‌نماید. در اینجا سعی می‌شود، نگاه دقیق‌تری به ساختار تعاملی عامل با محیط داشت. این ساختار، بر اساس مدل پیشنهادی [۳۷] می‌باشد که نقش منتقد را به محیط اضافه نموده است که در شکل ۲ قابل مشاهده است.

۶-۳- یادگیری تقویتی انگیزش درونی

ایده آغازین برای بررسی انگیزش درونی با استفاده از چارچوب یادگیری تقویتی در این است که فرایندهای یادگیری و تولید رفتار، به درونی یا بیرونی بودن سیگنال‌های پاداش اهمیتی نمی‌دهند و می‌توان فرایندهای مشابهی را برای هر دو در نظر گرفت [۲۲]. وقتی معیارهای بیان شده در روانشناسی در قالب مدل‌های محاسباتی انگیزش درونی ظهور و بروز پیدا کرده‌اند، زمانی که در چارچوب یادگیری تقویتی مورد استفاده قرار می‌گیرند، نام انگیزش درونی خواهند گرفت [۸۸]. می‌توان گفت یادگیری تقویتی انگیزش درونی به نوعی همان یادگیری تقویتی سنتی است که به جای پاداش که مسئله عملی داده شده‌ی مشخصی باشد، از معیار اندازه‌گیری جذابیت استفاده می‌کند [۴۱]. منظور از جذابیت، عمل یا وضعیتی است که اکتشاف نشده است [۸۸]. سیستمی می‌تواند ساخته شود که یاد بگیرد که چگونه ترتیبی از اعمال را بدست آورد که مجموع پاداش تنزلی آینده را بیشینه کند [۴۱]. در محیط‌های واقعی و یا نزدیک به آن، کاربردی بودن در صورت عدم وجود ویژگی مارکوف[1] یا فقدان نمایش جدولی، از جذابیت‌های انگیزش درونی محسوب می‌شود [۷۸]. محیط، صرفاً دنیای خارج از ربات یا حیوان نیست و شامل اجزای درونی آن‌ها نیز می‌شود. شکل ۳ بهبودیافته‌ی شکل ۲ است که محیط را به دو بخش خارجی و داخلی تقسیم کرده است. محیط خارجی نشان‌دهنده چیزی است که خارج از حیوان یا ربات است (ارگانیسم یا موجود زنده)، درحالی‌که محیط داخلی شامل اجزایی است که درون موجود زنده قرار دارد. هر دو مؤلفه با هم محیط عامل یادگیری تقویتی را شامل می‌شوند. این بهبودِ چارچوب یادگیری تقویتی روشن می‌کند که تمام سیگنال پاداش، درون موجود زنده تولید می‌شود. سیگنال‌های پاداش می‌توانند منشأهای مختلفی برای راه‌اندازی داشته باشند از قبیل: احساساتی که توسط اشیا و رخدادها در محیط خارجی تولید می‌شوند، ترکیب تحریک خارجی و شرایط داخلی محیطی یا سیگنال‌های تولید شده توسط فعالیت‌های درونی محیط. همه این امکان‌ها می‌تواند با چارچوب RL سازگار باشد [۲۲]. اشمیدهوبر [۴۳،۹۷] روش‌هایی برای پیاده‌سازی کنجکاوی به وسیله‌ی پاداش یک عامل RL به‌منظور بهبود توانایی پیش‌بینی‌اش معرفی کرده است.

[1] Markov

جدول ۸: مدل‌های محاسباتی IM: غافلگیری، کنجکاوی، (عدم) قطعیت

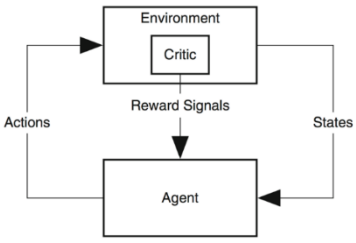
سال انتشار و مرجع	ایده اصلی	مزایا	معایب/چالش‌ها	روش یا الگوریتم
۲۰۱۹ [۸۰]	اصلی‌ترین ایده این مقاله استفاده از تخمین جریان نوری از حوزه بینایی ماشین برای مواجهه با مشکل تعادل بین کاوش و بهره‌برداری از تجربیات در وظایف با ابعاد بالا و پاداش‌های پراکنده در DRL است.	ارائه روشی برای مواجهه با مشکل فراموشی فاجعه‌بار[1] در کاوش مبتنی بر کنجکاوی	نیاز به استفاده از تخمین جریان نوری برای اندازه‌گیری جدید بودن حالت‌ها که ممکن است پیچیدگی بیشتری را به همراه داشته باشد.	تخمین جریان نوری از حوزه بینایی رایانه و استفاده از آن برای ایجاد جوایز درونی برای انگیزه‌بخشی به کاوش عمیق
۲۰۲۰ [۱۴۰]	ترکیب RL مبتنی بر کنجکاوی با کنترل زمان‌بندی در یک سلول تولیدی روباتیک انعطاف‌پذیر. هدف از این کار، کاهش نیاز به تنظیمات دستی پاداش‌ها در مسائل پیچیده مهندسی مانند بهینه‌سازی زمان‌بندی است.	حذف بایاس استقرایی ناشی از تنظیمات دستی پاداش، کاهش پیچیدگی تنظیم پاداش‌ها، امکان انتقال مستقیم تابع پاداش از یک محیط تولیدی به محیط دیگر بدون نیاز به تنظیم مجدد	پیچیدگی در پایداری آموزش، نیاز به بررسی بیشتر برای بهبود عملکرد در محیط‌های پیچیده‌تر، کافی‌نبودن امکان استفاده از کنجکاوی به‌تنهایی در برخی موارد	RL مبتنی بر کنجکاوی است که از IM برای ایجاد پاداش‌ها استفاده می‌کند.
۲۰۲۱ [۱۴۱]	معرفی مفهوم «کنجکاوی سریع و کند» برای تشویق به کاوش در بازه‌های زمانی طولانی‌تر در RL. این روش با تجزیه پاداش کنجکاوی به دو نوع پاداش سریع برای کاوش محلی و پاداش کند برای کاوش سراسری، به دنبال افزایش بهره‌وری کاوش و بهبود عملکرد عامل‌ها در محیط‌های با پاداش‌های کم یا تأخیردار است.	بهبود بهره‌وری کاوش، افزایش انعطاف‌پذیری عامل‌ها در مواجهه با محیط‌هایی با پاداش‌های کم و تأخیردار، عملکرد بهتر نسبت به روش‌های قبلی در اکثر وظایف آزمایشی	پیچیدگی محاسباتی و نیاز به ارزیابی بیشتر برای تعمیم‌پذیری در محیط‌های مختلف	ترکیبی از پاداش‌های کنجکاوی سریع و کند است که بر اساس خطا در بازسازی مشاهدات توسط شبکه‌های بازسازی فرموله شده‌اند. این روش با الگوریتم‌های RL سیاستی ترکیب می‌شود.
۲۰۲۲ [۱۴۲]	استفاده از کنجکاوی مبتنی بر شگفتی بیزین در فضای پنهان برای هدایت کاوش در وظایف RL با پاداش کم.	کاهش هزینه‌های محاسباتی، بهبود عملکرد کاوش در محیط‌های مختلف، مقاومت در برابر تغییرات تصادفی در دینامیک محیط	پیچیدگی در پیاده‌سازی مدل‌های فضای پنهان، نیاز به تنظیم دقیق پارامترها	روش شگفتی بیزین در فضای پنهان[2] که با استفاده از تفاوت بین باورهای قبلی و پسینی عامل در فضای پنهان، کاوش را هدایت می‌کند.
۲۰۲۳ [۱۴۳]	بررسی و حل مشکل کاوش در محیط‌های دارای پاداش کم یا بدون پاداش با استفاده از کنجکاوی درونی. این روش از مفهوم «کنجکاوی در واپس‌نگری»[3] استفاده می‌کند تا تفاوت بین جنبه‌های پیش‌بینی‌پذیر و غیرقابل‌پیش‌بینی نتایج را با استفاده از مدل‌های علّی[4] ساختاری یاد بگیرد.	قابلیت مقیاس‌پذیری ساده، مقاومت در برابر تصادفی بودن محیط، بهبود عملکرد کاوش در بازی‌های سخت مانند Montezuma's Revenge. این روش باعث می‌شود که عامل در مواجهه با نویز و تغییرات تصادفی گیر نکند.	پیچیدگی پیاده‌سازی مدل‌های ساختاری علّی، نیاز به تنظیم دقیق پارامترها، بررسی کامل تأثیر تغییرات تصادفی محیط بر عملکرد کاوش	روش «کنجکاوی در واپس‌نگری» که با استفاده از مدل‌های ساختاری علّی، تفاوت بین جنبه‌های پیش‌بینی‌پذیر و غیرقابل‌پیش‌بینی نتایج را یاد می‌گیرد.
۲۰۲۳ [۱۴۴]	معرفی یک تابع پاداش جدید برای IMRL که رفتار عامل‌های بازی را به‌گونه‌ای طراحی می‌کند که شبیه به رفتار انسان‌ها باشد. این روش بر اساس نظریه‌ای است که کنجکاوی را به‌عنوان شکاف اطلاعات تعریف می‌کند بنا شده است.	شباهت بیشتر به رفتار انسان، حفظ رقابت‌پذیری عامل، باورپذیری بیشتر آن در بازی‌های رایانه‌ای، کمک به توسعه‌دهندگان بازی جهت تست شبه‌انسانی	نیاز به داده‌های بیشتر برای تعمیم به سطوح و بازی‌های دیگر، پیچیدگی تنظیم تابع پاداش برای دستیابی به رفتار مطلوب، محدودیت آزمون‌ها محدود به یک بازی خاص	روش «کنجکاوی محتاطانه» است که بر اساس نظریه شکاف اطلاعات و منحنی U وارونه طراحی شده است. این روش با مدل کنجکاوی درونی ترکیب شده تا عامل‌ها را شبیه به انسان‌ها کند.
۲۰۲۴ [۱۴۵]	معرفی روشی برای جمع‌آوری داده‌های بدون نظارت با استفاده از کنجکاوی و فواصل زمانی تطبیقی برای RL آفلاین در چندین وظیفه. این روش به‌جای تمرکز صرف بر الگوریتم‌های یادگیری، بر بهبود فرایند جمع‌آوری داده‌ها	بهبود بازده محاسباتی و نمونه، جمع‌آوری داده‌های باکیفیت بالاتر برای وظایف پایین‌دستی مختلف، و توانایی تطبیق فواصل زمانی برای گسترش فضای	پیچیدگی پیاده‌سازی مکانیزم دسترسی و تطبیق فواصل زمانی اشاره، نیاز به ارزیابی جامع‌تری برای اطمینان از تعمیم‌پذیری نتایج به وظایف و محیط‌های مختلف.	روش «جمع‌آوری داده‌های بدون نظارت با استفاده از کنجکاوی» که از یک ماژول دسترسی برای تخمین احتمال دسترسی به حالت آینده k مرحله‌ای از حالت فعلی

[1] Catastrophic Forgetting
[2] Latent Bayesian Surprise
[3] Curiosity in Hindsight
[4] causal

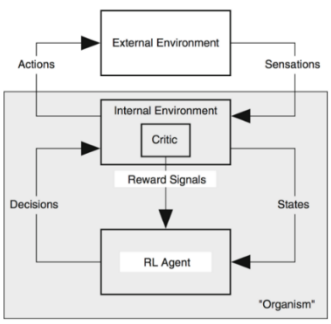
	تأکید دارد تا فضای ویژگی‌های عامل‌ها گسترش یابد و داده‌های باکیفیت‌تری جمع‌آوری شود.	ویژگی‌ها، برتری بیشتری نسبت به روش‌های بدون نظارت.		استفاده می‌کند و فواصل زمانی را به طور تطبیقی تعیین می‌کند.
۲۰۲۴ [۱۴۶]	معرفی یک روش RL بدون نظارت است که بر اساس کنجکاوی و حداکثرسازی آنتروپی در چندین محیط مختلف عمل می‌کند. این روش به دنبال بهبود عملکرد با استفاده از تغییراتی مانند اکتشاف مبتنی بر کنجکاوی، آلفای دینامیک، و افزایش آستانه KL Divergence است.	بهبود عملکرد در محیط‌های پیچیده‌تر مانند Ant، افزایش بازده یادگیری با استفاده از آستانه بالاتر KL Divergence و قابلیت تطبیق آلفا به‌صورت دینامیک برای یادگیری از مسیرهای ضعیف‌تر.	محدودیت در بهره‌وری اکتشاف مبتنی بر کنجکاوی در محیط‌های ساده‌تر مانند Grid World، نیاز به ارزیابی جامع‌تری برای تطبیق و ارزیابی روش در محیط‌های پیچیده‌تر	الگوریتم αMEPOL است که هدف آن پیش آموزش یک سیاست اکتشاف بدون نظارت و سپس تنظیم دقیق آن برای وظایف مختلف با استفاده از نظارت است. این مقاله با افزودن تغییراتی مانند اکتشاف مبتنی بر کنجکاوی و حداکثرسازی آنتروپی بهبودهای بیشتری را ایجاد کرده است.

ساتن[1] با تعریف «جایزه (انعام) اکتشافی» یک عامل برای مشاهده وضعیت‌ها به نسبت مدت زمانی که از مشاهده قبلی گذشته است، پاداش می‌دهد، در نتیجه عامل تشویق می‌شود که فضای حالت گسترده‌ای را پوشش دهد [۲۲]. از طرفی رویکردهای ماژولار و سلسله‌مراتبی، به‌عنوان حوزه‌ای از هوش مصنوعی به دنبال راه‌حل‌های عمومی که معمولاً روی خلاقیت در انسان‌ها و سایر حیوانات تکیه می‌کنند، محسوب می‌شوند. می‌توان یک توازی بین مشخصات بینش باتوجه به روان‌شناسی و مشخصات سیستم‌های HRL برای عامل‌های مجسم ایجاد کرد. تکنیک‌های HRL با توصیف بینش در انسان‌ها که شامل پیشرفت تحلیلی و توانایی بازسازی فضای جستجو است، تطابق دارند [۵۰]. در نتیجه، یکی از راه‌حل‌ها برای مشکل پاداش خلوت، استفاده از HRL است. راه‌حل آن، استفاده از مجموعه‌ای از اعمال موقتاً توسعه‌یافته یا گزینه‌هایی که هر کدام زیرهدف خودش را دارد. این زیراهداف، به‌صورت طبیعی برای وظایف خاصی ساخته می‌شوند [۹۶]. یادگیری در این تنظیمات نیازمند این است که عامل دانش را نمایش داده و در چندین سطح انتزاعی وابسته به فضا و زمان به‌منظور اکتشاف محیط به صورتی مؤثرتر عمل کند. اخیراً توابع غیرخطی تخمین‌زننده با RL ترکیب شده‌اند که یادگیری انتزاع‌ها روی فضای وضعیت با ابعاد بالای خلوت ممکن شده است. اما اکتشاف با بازخورد چالش اصلی است. عامل‌های انگیزش درونی می‌توانند رفتارهای جدید را برای خودشان اکتشاف کنند تا اهداف خارجی را مستقیماً حل نمایند [۹۸].

[۴۲] مدلی از انگیزش درونی را ارائه می‌کند که با استفاده از اندازه‌گیری نرخ بهبود اخذ مهارت بر اساس سیگنال تفاضل موقتی رفتار می‌کند. خبره از این مقدار از خطای تفاضل موقتی، در مدل HRL استفاده می‌کند. بدین صورت که از این خطا به‌عنوان تقویت داخلی برای انتخابگرها استفاده می‌شود. این روش باعث می‌شود که انتخابگر در هر وضعیت یاد بگیرد که خبره با بالاترین نرخ شایستگی را انتخاب کند [۴۲]. چالش الگوریتم‌های HRL این است که باوجود عملکرد مقاوم در فضای پیوسته و حتی باوجود محدود بودن فضا، در تشخیص مناطق غیرقابل‌یادگیری که در آن دانش یا شایستگی وجود ندارد، در ابعاد بالا غیرمؤثر هستند [۴۱]. چالش بعدی این است که اکثر روش‌های HRL ، نیازمند طراحی وظیفه‌ی خاص و آموزش مبتنی بر سیاست[2] (سیاست روشن) هستند. این موضوع باعث می‌شود که در سناریوهای دنیای واقعی به‌سختی استفاده شوند [۹۹]. بعلاوه رویکردی وجود ندارد که وظایف را به اجزای کوچک‌تری تقسیم نموده به‌گونه‌ای که شامل فضای عمل پیوسته‌ای باشد و سیاست کوتاه‌تری را در هر سطح انتزاع تضمین نماید [۱۰۰]. در اکثر الگوریتم‌های موجود، تنها عامل قادر است در سطوح بالاتری یاد بگیرد چنانچه فضای عمل، گسسته باشد [۱۰۱]. جدول ۹ به مقایسه برخی از الگوریتم‌های یادگیری تقویتی پرداخته است.



شکل ۲: تعامل عامل با محیط در RL [۳۷]



شکل ۳: تعامل عامل با محیط در RL-بهبودیافته شکل ۲ [۲۲]

[1] Sutton
[2] On-policy

۶-۴- یادگیری تقویتی عمیق

یادگیری عمیق به طور قابل‌توجهی یادگیری ماشین را پیشرفت داده است، به‌ویژه در کارهایی مانند تشخیص تصویر، تشخیص گفتار و ترجمه زبان. این روش در استخراج ویژگی‌های معنی‌دار از داده‌های با ابعاد بالا، کاهش مشکل ابعاد بالا، بسیار موفق است. این پیشرفت به یادگیری تقویتی نیز گسترش‌یافته است و منجر به حوزه DRLشده است. DRL از شبکه‌های عصبی برای توانمندسازی عامل‌ها برای یادگیری رفتارهای پیچیده و تصمیم‌گیری در محیط‌های پیچیده استفاده می‌کند [۹۲]. در ادامه روش‌های مختلف یادگیری عمیق مورداستفاده در ارتباط با RL موردبحث قرار می‌گیرند که هر کدام مزایای منحصربه‌فردی در بهبود عملکرد و کاربرد در کارهای دنیای واقعی دارند.

شبکه‌های عمیق Q[1] (DQN): با معرفی یادگیری عمیق برای تقریب مقادیر Q، انقلابی در این زمینه ایجاد کردند و به عامل‌ها اجازه دادند تا از داده‌های بصری خام بیاموزند [۱۰۴]. DQN از یک شبکه عصبی کانولوشنال (CNN) برای پردازش ورودی‌های تصویر و پیش‌بینی بهترین عمل در یک حالت داده شده استفاده می‌کند. تکنیک‌های کلیدی مانند پخش تجربه و شبکه‌های هدف برای تثبیت آموزش و جلوگیری از واگرایی پیاده‌سازی شدند DQN .اثربخشی خود را با شکست‌دادن بازیکنان انسانی در بازی‌های کلاسیک آتاری نشان داد و نقطه عطفی مهم در یادگیری تقویتی عمیق را رقم زد.

روش‌های گرادیان سیاست: گرادیان‌ها می‌توانند یک سیگنال یادگیری قوی در مورد چگونگی بهبود یک خط‌مشی پارامتری شده را ارائه دهند. از جمله این روش‌ها می‌توان به بهینه‌سازی سیاست مجاور (PPO)[۱۵۳] اشاره کرد که سیاست‌ها را مستقیماً با پارامترگذاری آن‌ها با شبکه‌های عصبی بهینه می‌کنند. PPO به‌ویژه برای ایجاد تعادل بین اکتشاف و بهره‌برداری با استفاده از یک تابع هدف محدود شده که پایداری آموزش را بهبود می‌بخشد، قابل‌توجه است. این روش‌ها برای کارهایی که شامل فضاهای عملی پیوسته و محیط‌های پیچیده هستند، جایی که روش‌های گسسته مانند DQN کوتاه می‌آیند، بسیار مناسب هستند.

روش‌های عامل - منتقد: روش‌های بازیگر - منتقد مزایای رویکردهای مبتنی بر سیاست و مبتنی بر ارزش را ترکیب می‌کنند و منجر به یادگیری کارآمدتر می‌شوند. گرادیان سیاست قطعی عمیق (DDPG) [۱۰۶] و بازیگر-منتقد مزیت ناهمگام (A3C) [۹۳] نمونه‌های قابل توجهی هستند. DDPG چارچوب عامل-منتقد را با استفاده از شبکه‌های عمیق برای نمایش هم عامل (سیاست) و هم منتقد (تابع ارزش) به وظایف کنترل پیوسته گسترش می‌دهد. A3C از موازی‌سازی برای به‌روزرسانی همزمان سیاست و تابع ارزش استفاده می‌کند و منجر به یادگیری سریع‌تر و پایدارتر می‌شود.

یادگیری تقویتی سلسه‌مراتبی: HRL با ادغام یادگیری عمیق، مدیریت وظایف پیچیده را با تجزیه آن‌ها به زیر وظایف که هر کدام توسط یک شبکه عصبی متفاوت مدیریت می‌شوند، امکان‌پذیر می‌کند. این روش به عامل‌ها اجازه می‌دهد تا هم استراتژی‌های سطح بالا و هم اقدامات سطح پایین را هم‌زمان بیاموزند. [۹۸] رویکردی را معرفی کردند که در آن از یادگیری عمیق برای پیاده‌سازی ساختارهای وظیفه‌ای سلسله‌مراتبی استفاده می‌شد که ترویج اکتشاف و یادگیری استراتژیک‌تر را تسهیل می‌کرد.

۶-۵- یادگیری تقویتی عمیق انگیزش درونی

همان‌طور که اشاره شد رویکردهای پیشین، فاقد تعمیم‌پذیری بوده و در مسائل با ابعاد پایین کارایی دارند. بعلاوه نیاز دارند تا فضای عمل نیز گسسته باشد. این محدودیت‌ها باعث شده است که این الگوریتم‌ها با پیچیدگی حافظه، پیچیدگی محاسباتی و در یادگیری ماشین با پیچیدگی نمونه مواجه شوند [۹۲]. مشکل اصلی فضای پیوسته حسی حرکتی، ابعاد بسیار بالای آن است. اخیرا رویکردهایی که می‌توانند به صورت طبیعی در فضاهای عمل و/یا وضعیت پیوسته عملکردی امکان-پذیر داشته باشند، معرفی شده‌اند [۱۰۲،۱۰۳]. یادگیری عمیق، که مهم‌ترین مشخصه آن، شبکه‌های عصبی عمیق است می‌تواند به طور خودکار، بازنمایی (ویژگی‌ها) ابعاد پایین فشرده‌ای را از داده‌های با ابعاد بالا بیابد. این قابلیت می‌تواند در حوزه RL نیز تحولی ایجاد کند. یادگیری عمیق، RL را قادر می‌سازد تا به مسائل تصمیم‌گیری تعمیم یابد، مسائلی که قبلا قابل کنترل نبودند یعنی مسائلی با ابعاد بالای فضای حالت و عمل [۱۰۴].

DRL در چگونگی بازی آتاری از پیکسل‌های خام ورودی [۹۳،۱۰۴،۱۰۵] به موفقیتی در موضوع یادگیری رسید، که از آن می-توان به نقطه عطف تعبیر نمود. بعلاوه، DRL، پیشرفت شگرفی نیز بر حوزه‌ی وظایف کنترلی پیوسته داشته است. به طور نمونه، مهارت‌های جابه‌جایی [۱۰۶،۱۰۷]، یادگیری رفتارهای دستکاری ماهرانه [۱۰۸] و آموزش بازوهای ربات برای وظایف دستکاری ساده. اخیرا از DRL در حل مسائل سخت با پاداش‌های خلوت نیز استفاده شده است [۱۰۹]. اودیر[2] و همکاران استفاده از الگوریتم‌های یادگیری نمایش عمیق را برای یادگیری فضای هدف مناسب پیشنهاد داده‌اند. رویکرد پیشنهادی دارای دو فاز است: ۱) فاز یادگیری ادراکی (ترکیب یادگیری بازنمایی عمیق): الگوریتم‌های یادگیری عمیق با استفاده از تغییرات مشاهده حسگر خام غیر فعال[3]، اقدام به یادگیری فضای نهفته مورد نظر می‌کنند. ۲) فرآیند اکتشاف هدف به وسیله‌ی اهداف نمونه گیری در فضای پنهان. در این مقاله چارچوبی محاسباتی برای اشکال سطح بالای اکتشاف که فرآیندهای اکتشاف هدف RL می‌باشد، ارائه شده است [۴۱].

مقاله [۹۶] برای توانایی کنترل ویژگی‌های این چنین محیط‌هایی از الگوریتمی بهره برده است، که در آن عاملی که به صورت انگیزش درونی، جنبه‌های محیط خود را از طریق مجموعه‌ی گزینه‌ها کنترل می‌کند، طراحی شده است. معماری عامل از RL فئودال[4] و DRL الهام گرفته

[1] Deep Q-Network
[2] Oudeyer
[3] Passive
[4] Feudal

شده است. عامل دارای دو بخش است: ۱) فراکنترل کننده که پاداش بیرونی را بیشینه می‌کند که شامل مجموعه‌ای گسسته از زیر اهداف است. ۲) زیرکنترل کننده[1] که IR را شامل می‌شود. در صورت وجود انگیزش درونی، عملکرد عامل در محیط خلوت بهتر می‌شود. در [۱۱۰] از فضای وضعیت برای ساخت فضای هدف استفاده می‌شود (وضعیت‌هایی با پتانسیل هدف بودن). مستقیما با استفاده از فضای وضعیت به عنوان فضای هدف، سه سطح از گزینه‌ها را یاد می‌گیرد و فاصله بین هدف و وضعیت پایانی به عنوان پاداش در نظر گرفته می‌شود. [۹۹] متناسب با همین کار، به عنوان هدف، فاصله بین وضعیت ابتدایی و وضعیت پایان گزینه را در نظر می‌گیرد. IR مهارت‌ها را به سوی مناطق خاص فضایی (مکانی) هدایت می‌کند.

مقاله [۱۱۱] از خودرمزنگار متغیر[2] برای ساخت فضای ویژگی به عنوان هدف استفاده می‌کند که اطلاعات مفید و لازم در طول فشرده‌سازی وضعیت‌ها به بازنمایی جدید از دست نروند. اخیرا [۱۱۲] بهبودی را روی [۱۱۱] انجام داده است و به وضعیت‌های نادر، وزن اختصاص داده است تا منجر به سیاست‌های متنوع‌تری شود. در [۱۱۳] سعی شده است تا از قدرت انگیزش درونی و روش‌های سلسله مراتبی توأمان استفاده شود. بعلاوه شبکه‌های عصبی تصادفی با مفاهیم تئوری اطلاعات ترکیب شده است. این پژوهش یادگیری مهارت‌ها را به وسیله‌ی بیشینه کردن معادله ۱ که IR است، انجام می‌دهد. I اطلاعات متقابل، τ مسیر[3] در طول گزینه، s_i وضعیت ابتدایی، f تابعی که بخشی از مسیر را انتخاب می‌کند و g_t هدفی است که توسط سیاست درون گزینه‌ای یا نمونه‌برداری یکنواخت تهیه شده است.

$$\boldsymbol{R_{int}(s_t, g_t) = I(g_t, f(\tau)|s_i)} \qquad (۱)$$

البته باید توجه داشت که فضای هدف در اینجا گسسته است.

[۷۶] و [۱۱۴] ایده مشابه کار قبلی را ارائه می‌دهند با این تفاوت که در گسسته سازی (بهره‌گیری از شبکه عصبی) و انتخاب f به عنوان بخشی از مسیر مهارت در محیط با کار قبلی متفاوت هستند. [۱۱۴] $f(\tau)$ را به عنوان وضعیت مسیر انتخاب کرده و انگیزش درونی را در هر تکرارِ مسیر محاسبه می‌کند. [۷۶] با در نظر گرفتن $f(\tau)$ به عنوان مجموع همه وضعیت‌ها در مسیر و با تعیین پاداش در هر مسیر، روش متمایزی دارد. [۱۱۵] فضای هدف را به عنوان فضای وضعیت در نظر می‌گیرد، سپس با تلاش برای وضعیت پایانی مسیر به عنوان هدف صحیح در میان سایر اهداف انتخاب شده از توزیع مشابه با هدف واقعی، تقریب انجام می‌دهد. این مانند یادگیری برای یافتن نزدیک‌ترین هدف به وضعیت نهایی مجموعه‌ای از اهداف است. در [۱۱۶] عامل، رمز کردن مسیرها را به یک فضای پنهان یاد می‌گیرد. سپس به روشی مشابه، با استفاده از خودرمزنگار متغیر، آن‌ها را رمز گشایی می‌کند. بعلاوه مسیرها به وسیله سیاست نهفته-مشروط تولید شده و رمزگشاها سازگاری با هم را فرا می‌گیرند. این مقاله نتایج جذابی روی محیط‌های ساده با استفاده از روش‌های برنامه‌ریزی بدست آورده است.

الگوریتمی به نام Hierarchical-DQN ارائه شده است به‌گونه‌ای که مدل، تصمیمات را بر روی دو سطح سلسله‌مراتبی اتخاذ می‌کند [۹۸]: یکی ماژول سطح بالا (فرا کنترل کننده[4]): در وضعیت‌ها رخ می‌دهد و یک هدف جدید را انتخاب می‌کند (یادگیری یک سیاست بر روی اهداف ذاتی) و دیگری ماژول سطح پایین (کنترل کننده): با استفاده از هر دو مورد وضعیت و اهداف انتخابی برای انتخاب اعمال استفاده کرده تا به هدف برسد یا اپیزود خاتمه یابد (سیاست انجام اهداف هر گزینه را یاد می‌گیرد). سپس فرا کنترل کننده هدف دیگری را انتخاب کرده و گام‌های قبلی را تکرار می‌کند.

۷- محدودیت‌ها، کاربردها و کارهای آینده

۷-۱- محدودیت‌ها

پیچیدگی محاسباتی و نیاز به منابع زیاد: پیاده‌سازی انگیزش درونی اغلب شامل مدل‌های پیش‌بینی پیچیده و وظایف کمکی است که نیاز به منابع محاسباتی قابل‌توجهی دارند. برای مثال، آموزش مدل‌های یادگیری عمیق با انگیزش درونی نیازمند شبکه‌های اضافی برای پیش‌بینی یا اندازه‌گیری تازگی است که می‌تواند زمان آموزش و نیازهای سخت‌افزاری را به‌شدت افزایش دهد.

زمان آموزش طولانی: آموزش عامل‌های هوش مصنوعی با انگیزش درونی می‌تواند زمان‌بر باشد، به‌ویژه زمانی که با محیط‌های مقیاس بزرگ و کارهای پیچیده سروکار داشته باشیم.

پتانسیل برای بیش‌برازش به تازگی: عامل‌های هدایت‌شده توسط انگیزش درونی ممکن است به جنبه جستجوی تازگی بیش‌برازش[5] کنند و استراتژی‌هایی را توسعه دهند که IR را به حداکثر برسانند بدون این که به یادگیری بلندمدت کمک کند. این می‌تواند به رفتاری منجر شود که عامل به دنبال حالت‌های نامرتبط یا غیرمفید برود، صرفاً به دلیل این که آن‌ها جدید هستند، بدون این که به تکمیل وظیفه یا اهداف یادگیری گسترده‌تر کمک کند.

چالش‌های تعمیم‌پذیری: عامل‌هایی که توسط انگیزش درونی هدایت می‌شوند اغلب در تعمیم رفتارهای آموخته‌شده به وظایف یا محیط‌های مختلف دچار مشکل می‌شوند. مکانیزم‌های IR معمولاً به شرایط بستگی دارند و ممکن است زمانی که عامل در محیطی جدید با پویایی‌های متفاوت قرار می‌گیرد، به‌خوبی انتقال نیابند. در نتیجه، مدل‌های انگیزش درونی می‌توانند در مواجهه با کاربردهای پیچیده و واقعی که به یادگیری انتقالی قوی نیاز دارند، محدودیت داشته باشند.

۷-۲- کاربردها

چالش اصلی برای انگیزش درونی، کاربردپذیری آن به دلیل پیچیدگی نمونه بالای روش‌های پیشنهادی است؛

[1] Sub-controller
[2] Variational auto-encoder (VAE)
[3] Trajectory
[4] Meta controller
[5] Overfitting

جدول ۹: مقایسه برخی از مهم‌ترین الگوریتم‌های یادگیری تقویتی

سال انتشار و مرجع	روش	مؤلفه‌های کلیدی	کاربردها	مناسب برای محیط (گسسته/پیوسته)	نوع ورودی
۲۰۱۵ [۱۰۴]	Q-Learning عمیق (DQN)	شبکه عصبی پیچشی(CNN) ، بازپخش تجربه، شبکه هدف	بازی‌ها و شبیه‌سازی‌های تصویری	گسسته	تصویری
۲۰۱۵ [۱۰۶]	بازیگر - منتقد پیوسته (DDPG)	شبکه بازیگر و منتقد، تجربه بازپخش	رباتیک پیوسته	پیوسته	برداری
۲۰۱۶ [۷۸]	اکتشاف مبتنی بر شمارش	شبه‌شمارش‌ها، مدل‌های احتمالی	اکتشافات در بازی‌ها	گسسته	برداری، تصویری
۲۰۱۷ [۱۵۳]	روش‌های سیاست‌محور (PPO)	گرادیان سیاست، تابع کلیپ شده برای بهینه‌سازی	رباتیک، بازی‌ها	هر دو	برداری، تصویری
۲۰۱۷ [۷۹]	روش‌های کنجکاوی‌محور (ICM)	ماژول کنجکاوی، خطای پیش‌بینی	شبیه‌سازی‌های رباتیک، بازی‌ها	هر دو	تصویری، برداری
۲۰۱۸ [۱۵۹]	تقطیر شبکه تصادفی (RND)	شبکه پیش‌بینی، شبکه هدف تصادفی	بازی‌ها، شبیه‌سازی‌ها	هر دو	تصویری
۲۰۱۸ [۱۶۰]	SAC (Soft Actor-Critic)	سیاست نرم، شبکه بازیگر و منتقد	کنترل رباتیک	پیوسته	برداری
۲۰۲۰ [۱۶۱]	DreamerV2	مدل‌سازی محیط، پیش‌بینی آینده	شبیه‌سازی‌های سه‌بعدی	هر دو	برداری، تصویری
۲۰۲۰ [۱۶۲]	Muzero	ترکیب یادگیری مدل‌محور و سیاست‌محور	بازی‌های شطرنج، گو	هر دو	تصویری
۲۰۲۱ [۱۶۳]	Asymmetric Self-Play for Automatic Goal Discovery	بازی خودکار نامتقارن، سیاست مبتنی بر هدف، آموزش با پاداش‌های پراکنده	رباتیک و کنترل بازو، وظایف پیچیده مانند چیدن میز، روی هم چیدن بلوک‌ها، حل پازل	پیوسته	تصویری، برداری
۲۰۲۳ [۱۵۶]	Intrinsic Motivation with FRE	رمزگذاری پاداش عملکردی، خودرمزگذار تغییرپذیر	رباتیک و کنترل پیوسته	هر دو	برداری
۲۰۲۴ [۱۶۴]	Hierarchical Multi-Agent RL	معماری چندعاملی سلسله‌مراتبی، عامل متا، سیاست عامل - منتقد، دو جریان پردازشی (اطلاعات مکانی و زمانی)	وظایف همکاری در سیستم‌های چندعاملی پیچیده	پیوسته	برداری

بنابراین، تنها تعداد محدودی از روش‌ها بر روی ربات‌های واقعی به نمایش گذاشته شده‌اند [۴۰،۶۴، ۱۵۱-۱۴۷]. با این حال هیچ یک از این روش‌ها یادگیری موثری را در اکتشاف مبتنی بر هدف ارائه نکرده‌اند. به طور نمونه در [۸۷،۴۰] ربات در طول اکتشاف حرکات تصادفی انجام می‌دهد که ناکارآمد و با حرکت انسان ناسازگار است. در مقابل [۵۳،۵۴،۶۴،۱۵۲]، کارایی نمونه بالا و حرکت هدفمند را نشان داده‌اند. تنها روش‌هایی که با کارایی نمونه بالا قابل توجه هستند، روش‌هایی هستند که انگیزش درونی را با روش‌های بازپخش ذهنی ادغام کرده‌اند [۱۲۸،۱۵۱].

۷-۳- کارهای آینده

بهبود هم‌راستایی پاداش‌های درونی و بیرونی: یکی از حوزه‌های کلیدی برای تحقیقات آینده، توسعه روش‌های بهتر برای هم‌راستایی انگیزش درونی با اهداف وظایف بیرونی است. این امر به سیستم‌های هوش مصنوعی اجازه می‌دهد تا از انگیزش درونی برای اکتشاف و یادگیری استفاده کنند، درحالی‌که همچنان بر دستیابی به نتایج خاص تمرکز دارند. تکنیک‌هایی که به طور پویا پاداش‌های درونی و بیرونی را متعادل می‌کنند، می‌توانند به حفظ تعادل بین اکتشاف و بهره‌برداری کمک کرده و از بروز تعارضات بین انواع مختلف انگیزش جلوگیری کنند.

ادغام با یادگیری انتقالی و متا - یادگیری: یک جهت‌گیری آینده طراحی سیستم‌های انگیزش درونی است که به تعمیم بهتر در میان وظایف و حوزه‌ها کمک می‌کند. ادغام انگیزش درونی با یادگیری انتقالی [1] و متا - یادگیری [2] باعث می‌شود مدل‌های هوش مصنوعی بتوانند استراتژی‌های اکتشافی آموخته‌شده را با حداقل بازآموزی در محیط‌های جدید به کار گیرند. این پیشرفت برای کاربردهایی مانند رباتیک انطباقی که هوش مصنوعی باید در محیط‌های متنوع و غیرقابل‌پیش‌بینی عمل کند، ضروری خواهد بود.

ترکیب انگیزش درونی **با معماری‌های شناختی مشابه انسان:** ادغام انگیزش درونی در سیستم‌های هوش مصنوعی که برای شبیه‌سازی فرایندهای شناختی انسانی طراحی شده‌اند، مانند معماری‌های شناختی،

[1] Transfer Learning

[2] Meta-Learning

می‌تواند سیستم‌هایی را به وجود آورد که به روش‌هایی مشابه با انسان یاد بگیرند. با قراردادن محرک‌های کنجکاوی و اکتشاف در چارچوب‌هایی مانند ACT-R یاSOAR ، عامل‌های هوش مصنوعی می‌توانند جنبه‌هایی از تصمیم‌گیری انسانی مانند حل مسئله مبتنی بر کنجکاوی و یادگیری تطبیقی را شبیه‌سازی کنند.

توسعه انگیزش درونی اجتماعی: سیستم‌های هوش مصنوعی آینده می‌توانند از انگیزش درونی اجتماعی بهره‌مند شوند، جایی که عامل‌ها نه‌تنها توسط کنجکاوی شخصی بلکه توسط عوامل اجتماعی مانند یادگیری یا همکاری با دیگر عامل‌ها ترغیب می‌شوند. این رویکرد به‌ویژه در محیط‌های چندعاملی و وظایف تعاونی که در آن‌ها هوش مصنوعی نیاز به همکاری با دیگر سیستم‌ها دارد، مهم خواهد بود.

ادغام با یادگیری خودنظارتی و هوش مصنوعی چندوجهی: یادگیری خودنظارتی به دلیل توانایی‌اش در استفاده از داده‌های بزرگ بدون برچسب برجسته شده است. مدل‌های انگیزش درونی آینده می‌توانند با این نوع از الگوریتم‌ها ادغام شوند تا کشف ویژگی‌های معنادار و یادگیری نمایش را تحریک کنند. علاوه بر این، ترکیب انگیزش درونی با یادگیری چندوجهی که در آن هوش مصنوعی می‌تواند هم‌زمان از متن، تصویر، صدا و ویدئو یاد بگیرد، امکان درک جامع‌تر و سازگاری بیشتری را فراهم می‌کند.

۸- نتیجه

توانایی شناختی و یادگیری مادام‌العمر در ارگانیسم‌های زنده به طور قابل‌توجهی از انگیزش درونی ناشی می‌شود. این انگیزش، بدون نیاز به پاداش‌های بیرونی، به موجودات زنده امکان می‌دهد تا به طور خودانگیخته به جستجو و اکتشاف بپردازند. این مقاله نشان می‌دهد که انگیزش درونی می‌تواند به طور مؤثر در سیستم‌های مصنوعی و ربات‌ها به کار گرفته شود تا یادگیری و توسعه خودمختار را بهبود بخشد. ترکیب الگوریتم‌های یادگیری تقویتی با مدل‌های انگیزش درونی، راهی امیدوارکننده برای ایجاد ربات‌هایی با توانایی‌های شناختی و ادراکی پیشرفته‌تر فراهم می‌کند. بااین‌حال، چالش‌های عملی همچنان وجود دارند و نیازمند پژوهش‌های بیشتر و بهبود روش‌های کنونی هستند.

مراجع

[1] M. Begum and F. Karray, "Computational intelligence techniques in bio-inspired robotics," in *Design and Control of Intelligent Robotic Systems*: Springer, 2009, pp. 1-28, doi: http://dx.doi.org/10.1007/978-3-540-89933-4_1.

[2] Lieto and D. P. Radicioni, "From human to artificial cognition and back: New perspectives on cognitively inspired ai systems," ed: Elsevier, 2016, doi: http://dx.doi.org/10.1016/j.cogsys.2014.11.001.

[3] G. Baldassarre, T. Stafford, M. Mirolli, P. Redgrave, R. M. Ryan, and A. Barto, "Intrinsic motivations and open-ended development in animals, humans, and robots: an overview," *Frontiers in psychology,* vol. 5, p. 985, 2014, doi: http://dx.doi.org/10.3389/fpsyg.2014.00985.

[4] Cangelosi and M. Schlesinger, Developmental robotics: From babies to robots. MIT Press, 2015, doi: http://dx.doi.org/10.7551/mitpress/9320.001.0001.

[5] K. Merrick, "Value systems for developmental cognitive robotics: A survey," *Cognitive Systems Research,* vol. 41, pp. 38-55, 2017, doi: http://dx.doi.org/10.1016/j.cogsys.2016.08.001.

[6] M. Asada *et al.*, "Cognitive developmental robotics: A survey," *IEEE transactions on autonomous mental development,* vol. 1, no. 1, pp. 12-34, 2009, doi: https://dx.doi.org/10.1109/TAMD.2009.2021702.

[7] J. Reeve, *Understanding motivation and emotion*. John Wiley & Sons, 2014.

[8] E. L. Deci, "Article commentary: on the nature and functions of motivation theories," *Psychological Science,* vol. 3, no. 3, pp. 167-171, 1992, doi: http://dx.doi.org/10.1111/j.1467-9280.1992.tb00020.x.

[9] R. M. Ryan and E. L. Deci, "Intrinsic and extrinsic motivations: Classic definitions and new directions," *Contemporary educational psychology,* vol. 25, no. 1, pp. 54-67, 2000, doi: http://dx.doi.org/10.1006/ceps.1999.1020.

[10] C. Darwin, *On the origin of species, 1859*. Routledge, 2004, doi: http://dx.doi.org/10.9783/9780812200515.

[11] R. S. Woodworth, "Columbia University lectures: Dynamic psychology," 1918, doi: http://dx.doi.org/10.1037/10015-000.

[12] C. L. Hull, "Principles of behavior :An introduction to behavior theory," 1943.

[13] J. W. Atkinson and N. T. Feather, *A theory of achievement motivation*. Wiley New York, 1966.

[14] L. Festinger, *A theory of cognitive dissonance*. Stanford university press, 1962, doi: http://dx.doi.org/10.1515/9781503620766.

[15] S. Harter, "Effectance motivation reconsidered. Toward a developmental model," *Human development,* vol. 21, no. 1, pp. 34-64, 1978, doi: http://dx.doi.org/10.1159/000271574.

[16] M. Csikszentmihalyi, *Beyond boredom and anxiety*. Jossey-Bass, 2000.

[17] E. A. Locke, "Motivation through conscious goal setting," *Applied and preventive psychology,* vol. 5, no. 2, pp. 117-124, 1996, doi: http://dx.doi.org/10.1016/S0962-1849(96)80005-9.

[18] M. E. Seligman, M. E. Seligman, and M. E. Seligman, "Helplessness: On depression, development, and death," 1975.

[19] Bandura, "Self-efficacy: toward a unifying theory of behavioral change ",Psychological review, vol. 84, no. 2, p. 191, 1977, doi: http://dx.doi.org/10.1037/0033-295X.84.2.191.

[20] H. Markus, "Self-schemata and processing information about the self," *Journal of personality and social psychology,* vol. 35, no. 2, p. 63, 1977, doi: http://dx.doi.org/10.1037/0022-3514.35.2.63.

[21] R. M. Ryan and E. L. Deci, "Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being," *American psychologist,* vol. 55, no. 1, p. 68, 2000, doi: http://dx.doi.org/10.1037/0003-066X.55.1.68.

[22] G. Barto, "Intrinsic motivation and reinforcement learning," in Intrinsically motivated learning in natural and artificial systems: Springer, 2013, pp. 17-47, doi: http://dx.doi.org/10.1007/978-3-642-32375-1_2.

[23] E. L. Deci and R. M. Ryan, "The general causality orientations scale: Self-determination in personality," *Journal of research in personality,* vol. 19, no. 2, pp.

Motivated Learning in Natural and Artificial Systems, pp. 49-72, 2013.
[40] S. Forestier, Y. Mollard, and P.-Y. Oudeyer, "Intrinsically motivated goal exploration processes with automatic curriculum learning," *arXiv preprint arXiv:1708.02190,* 2017, doi: https://dx.doi.org/10.48550/arXiv.1708.02190.
[41] P.-Y. Oudeyer, A. Baranes, and F. Kaplan, "Intrinsically motivated learning of real-world sensorimotor skills with developmental constraints," in *Intrinsically motivated learning in natural and artificial systems*: Springer, 2013, pp. 303-365, doi: http://dx.doi.org/10.1007/978-3-642-32375-1_13.
[42] G. Baldassarre and M. Mirolli, "Deciding which skill to learn when: temporal-difference competence-based intrinsic motivation (TD-CB-IM)," in *Intrinsically Motivated Learning in Natural and Artificial Systems*: Springer, 2013, pp. 257-278, doi: http://dx.doi.org/10.1007/978-3-642-32375-1_11.
[43] J. Schmidhuber, "A possibility for implementing curiosity and boredom in model-building neural controllers," in *Proc. of the international conference on simulation of adaptive behavior: From animals to animats*, 1991, pp. 222-227, doi: http://dx.doi.org/10.7551/mitpress/3115.003.0030.
[44] J. Schmidhuber, "Curious model-building control systems," in *Proc. international joint conference on neural networks*, 1991, pp. 1458-1463, doi: http://dx.doi.org/10.1109/IJCNN.1991.170605.
[45] N. Roy and A. McCallum, "Toward optimal active learning through monte carlo estimation of error reduction," *ICML, Williamstown,* pp. 441-448, 2001.
[46] M. Mutti, R. De Santi, M. Restelli, "The importance of non-markovianity in maximum state entropy exploration," in *International Conference on Machine Learning*, 2022, pp. 16223–16239, doi: https://dx.doi.org/10.48550/arXiv.2202.03060.
[47] P.-Y. Oudeyer, "Intelligent adaptive curiosity: a source of self-development," 2004.
[48] Kim, D., et al. "Accelerating reinforcement learning with value-conditional state entropy exploration," in *Advances in Neural Information Processing Systems*, vol. 36, 2024.
[49] Q. Yang, M. Spaan, "Cem: Constrained entropy maximization for task-agnostic safe exploration," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023, pp. 10798–10806, doi: http://dx.doi.org/10.1609/aaai.v37i9.26281.
[50] Colin, T., et al. "Hierarchical reinforcement learning as creative problem solving," in *Robotics and Autonomous Systems*, vol. 86, pp. 196–206, 2016, doi: http://dx.doi.org/10.1016/j.robot.2016.08.021.
[51] C. Tenorio-González and E. F. Morales, "Automatic discovery of concepts and actions," Expert Systems with Applications, vol. 92, pp. 192-205, 2018, doi: http://dx.doi.org/10.1016/j.eswa.2017.09.023.
[52] Memmel, M., et al. "ASID: Active Exploration for System Identification in Robotic Manipulation," in *arXiv preprint arXiv:2404.12308*, 2024.
[53] R. Rayyes, H. Donat, J. Steil. "Efficient online interest-driven exploration for developmental robots," in *IEEE Transactions on Cognitive and Developmental Systems*, vol. 14, no. 4, pp. 1367–1377, 2020, doi: http://dx.doi.org/10.1109/TCDS.2020.3001633.
[54] Rayyes, R., et al. "Interest-driven exploration with observational learning for developmental robots," in *IEEE Transactions on Cognitive and Developmental Systems*, vol. 15, no. 2, pp. 373–384, 2021, doi: http://dx.doi.org/10.1109/TCDS.2021.3057758.
109-134, 1985, doi: http://dx.doi.org/10.1016/0092-6566(85)90023-6.
[24] R. W. White, "Motivation reconsidered: The concept of competence," *Psychological review,* vol. 66, no. 5, p. 297, 1959, doi: http://dx.doi.org/10.1037/14156-005.
[25] P.-Y. Oudeyer and F. Kaplan, "How can we define intrinsic motivation," in *Proc. of the 8th Conf. on Epigenetic Robotics*, 2008, vol. 5, pp. 29-31.
[26] S. Roohi, J. Takatalo, C. Guckelsberger, and P. Hämäläinen, "Review of intrinsic motivation in simulation-based game testing," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 2018: ACM, p. 347, doi: http://dx.doi.org/10.1145/3173574.3173921.
[27] D. Schunk, M. DiBenedetto. "Motivation and social cognitive theory," in *Contemporary educational psychology*, vol. 60, pp. 101832, 2020, doi: http://dx.doi.org/10.1093/oxfordhb/9780195399820.013.0002.
[28] M. Csikszentmihalyi, M. Csikzentmihaly, *Flow: The psychology of optimal experience*. Harper & Row New York, 1990.
[29] Rhinehart, N., et al. "Information is power: Intrinsic control via information capture," in *Advances in Neural Information Processing Systems*, vol. 34, pp. 10745–10758, 2021, doi: https://dx.doi.org/10.48550/arXiv.2112.03899.
[30] Zhang, T., et al. "Made: Exploration via maximizing deviation from explored regions," in *Advances in Neural Information Processing Systems*, vol. 34, pp. 9663–9680, 2021.
[31] M. Mirolli and G. Baldassarre, "Functions and mechanisms of intrinsic motivations," in *Intrinsically Motivated Learning in Natural and Artificial Systems*: Springer, 2013, pp. 49-72, doi: http://dx.doi.org/10.1007/978-3-642-32375-1_3.
[32] R. De Charms, Personal causation: The internal affective determinants of behavior. Routledge, 2013.
[33] M. Csikszentmihalyi, "Toward a psychology of optimal experience," in *Flow and the foundations of positive psychology*: Springer, 2014, pp. 209-226, doi: http://dx.doi.org/10.1007/978-94-017-9088-8_14.
[34] J. Schmidhuber, "Maximizing fun by creating data with easily reducible subjective complexity," in *Intrinsically motivated learning in natural and artificial systems*: Springer, 2013, pp. 95-128, doi: http://dx.doi.org/10.1007/978-3-642-32375-1_5.
[35] Eppe, M., et al. "Intelligent problem-solving as integrated hierarchical reinforcement learning," in *Nature Machine Intelligence*, vol. 4, no. 1, pp. 11–20, 2022, doi: http://dx.doi.org/10.1038/s42256-021-00433-9.
[36] P. Redgrave and K. Gurney, "The short-latency dopamine signal: a role in discovering novel actions?," *Nature reviews neuroscience,* vol. 7, no. 12, p. 967, 2006, doi: http://dx.doi.org/10.1038/nrn2022.
[37] G. Barto, S. Singh, and N. Chentanez, "Intrinsically motivated learning of hierarchical collections of skills," in Proceedings of the 3rd International Conference on Development and Learning, 2004, pp. 112-19.
[38] J. Schmidhuber, "Formal theory of creativity, fun, and intrinsic motivation (1990–2010)," *IEEE Transactions on Autonomous Mental Development,* vol. 2, no. 3, pp. 230-247, 2010.
[39] M. Mirolli and G. Baldassarre, "Intrinsically motivated learning in natural and artificial systems," *Intrinsically*

[55] M. Mutti and M. Restelli, "An Intrinsically-Motivated Approach for Learning Highly Exploring and Fast Mixing Policies," *arXiv preprint arXiv:1907.04662,* 2019, doi: http://dx.doi.org/10.1609/aaai.v34i04.5968.
[56] V. G. Santucci, G. Baldassarre, and M. Mirolli, "GRAIL: a goal-discovering robotic architecture for intrinsically-motivated learning," *IEEE Transactions on Cognitive and Developmental Systems,* vol. 8, no. 3, pp. 214-231, 2016, doi: http://dx.doi.org/10.1109/TCDS.2016.2538961.
[57] J. Achterhold, M. Krimmel, J. Stueckler, "Learning temporally extended skills in continuous domains as symbolic actions for planning," in *Conference on Robot Learning*, 2023, pp. 225–236.
[58] M. Schembri, M. Mirolli, and G. Baldassarre, "Evolving internal reinforcers for an intrinsically motivated reinforcement-learning robot," in *2007 IEEE 6th International Conference on Development and Learning*, 2007: IEEE, pp. 282-287, doi: http://dx.doi.org/10.1109/DEVLRN.2007.4354052.
[59] Cartoni, E., et al. "REAL-X—Robot open-Ended Autonomous Learning Architecture: Building Truly End-to-End Sensorimotor Autonomous Learning Systems," in *IEEE Transactions on Cognitive and Developmental Systems*, 2023, doi: http://dx.doi.org/10.1109/TCDS.2023.3270081.
[60] Baranes and P.-Y. Oudeyer, "Intrinsically motivated goal exploration for active motor learning in robots: A case study," in 2010 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2010: IEEE, pp. 1766-1773, doi: http://dx.doi.org/10.1109/IROS.2010.5651385.
[61] S. Hangl, V. Dunjko, H. J. Briegel, and J. Piater, "Skill learning by autonomous robotic playing using active learning and creativity," *arXiv preprint arXiv:1706.08560,* 2017, doi: https://dx.doi.org/10.48550/arXiv.1706.08560.
[62] H. Qureshi, Y. Nakamura, Y .Yoshikawa, and H. Ishiguro, "Intrinsically motivated reinforcement learning for human–robot interaction in the real-world," Neural Networks, vol. 107, pp. 23-33, 2018, doi: http://dx.doi.org/10.1016/j.neunet.2018.03.014.
[63] Chiappa, A., et al. "Acquiring musculoskeletal skills with curriculum-based reinforcement learning," in *bioRxiv*, pp. 2024–01, 2024, doi: http://dx.doi.org/10.1016/j.neuron.2024.09.002.
[64] D. Tanneberg, J. Peters, and E. Rueckert, "Intrinsic motivation and mental replay enable efficient online adaptation in stochastic recurrent networks," *Neural Networks,* vol. 109, pp. 67-80, 2019, doi: http://dx.doi.org/10.1016/j.neunet.2018.10.005.
[65] U. Nehmzow, Y. Gatsoulis, E. Kerr, J. Condell, N. Siddique, and T. M. McGuinnity, "Novelty detection as an intrinsic motivation for cumulative learning robots," in *Intrinsically Motivated Learning in Natural and Artificial Systems*: Springer, 2013, pp. 185-207, doi: http://dx.doi.org/10.1007/978-3-642-32375-1_8.
[66] S. Marsland, U. Nehmzow, and J. Shapiro, "A real-time novelty detector for a mobile robot," *arXiv preprint cs/0006006,* 2000, doi: https://dx.doi.org/10.48550/arXiv.cs/0006006.
[67] H. V. Neto and U. Nehmzow, "Incremental PCA: An alternative approach for novelty detection," *Towards Autonomous Robotic Systems,* 2005.
[68] D. Tanneberg, J. Peters, and E. Rueckert, "Online learning with stochastic recurrent neural networks using intrinsic motivation signals ",in *Conference on Robot Learning*, 2017, pp. 167-174.
[69] E. Özbilge, "Experiments in online expectation-based novelty-detection using 3D shape and colour perceptions for mobile robot inspection," *Robotics and Autonomous Systems,* vol. 117, pp. 68-79, 2019, doi: http://dx.doi.org/10.1016/j.robot.2019.04.003
[70] S. Klyubin, D. Polani, and C. L. Nehaniv, "Empowerment: A universal agent-centric measure of control," in 2005 IEEE Congress on Evolutionary Computation, 2005, vol. 1: IEEE, pp. 128-135, doi: http://dx.doi.org/10.1109/CEC.2005.1554676.
[71] P. Capdepuy, D. Polani, and C. L. Nehaniv, "Maximization of potential information flow as a universal utility for collective behaviour," in *2007 IEEE Symposium on Artificial Life*, 2007: Ieee, pp. 207-213, doi: http://dx.doi.org/10.1109/ALIFE.2007.367798.
[72] S. Mohamed and D. J. Rezende, "Variational information maximisation for intrinsically motivated reinforcement learning," in *Advances in neural information processing systems*, 2015, pp. 2125-2133.
[73] K. Gregor, D. J. Rezende, and D. Wierstra, "Variational intrinsic control," *arXiv preprint arXiv:1611.07507,* 2016, doi: https://dx.doi.org/10.48550/arXiv.1611.07507.
[74] D. Emukpere, B. Wu, J. Perez. "SLIM: Skill Learning with Multiple Critics," in *arXiv preprint arXiv:2402.00823*, 2024, doi: https://dx.doi.org/10.48550/arXiv.2402.00823.
[75] F. Leibfried, S. Pascual-Diaz, and J. Grau-Moya, "A Unified Bellman Optimality Principle Combining Reward Maximization and Empowerment," *arXiv preprint arXiv:1907.12392,* 2019, doi: https://dx.doi.org/10.48550/arXiv.1907.12392.
[76] J. Achiam and S. Sastry, "Surprise-based intrinsic motivation for deep reinforcement learning," *arXiv preprint arXiv:1703.01732*, 2017, doi: https://dx.doi.org/10.48550/arXiv.1703.01732
[77] B. C. Stadie, S. Levine, and P. Abbeel, "Incentivizing exploration in reinforcement learning with deep predictive models," *arXiv preprint arXiv:1507.00814,* 2015, doi: https://doi.org/10.48550/arXiv.1507.00814.
[78] M. Bellemare, S. Srinivasan, G. Ostrovski, T. Schaul, D. Saxton, and R. Munos, "Unifying count-based exploration and intrinsic motivation," in *Advances in Neural Information Processing Systems*, 2016, pp. 1471-1479.
[79] D. Pathak, P. Agrawal, A. A. Efros, and T. Darrell, "Curiosity-driven exploration by self-supervised prediction," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 16-17, doi: http://dx.doi.org/10.1109/CVPRW.2017.70.
[80] H.-K. Yang, P.-H. Chiang, K.-W. Ho, M.-F. Hong, and C.-Y. Lee, "Never Forget: Balancing Exploration and Exploitation via Learning Optical Flow," *arXiv preprint arXiv:1901.08486,* 2019, doi: https://doi.org/10.48550/arXiv.1901.08486.
[81] M. Salichs Sánchez-Caballero and M. Á. Malfaz Vázquez, "A new approach to modelling emotions and their use on a decision-making system for artificial agent," 2012.
[82] D. Dörner and C. D. Güss, "PSI: A computational architecture of cognition, motivation, and emotion," *Review of General Psychology,* vol. 17, no. 3, pp. 297-317, 2013, doi: http://dx.doi.org/10.1037/a0032947.
[83] P. Sequeira, "Socio-emotional reward design for intrinsically motivated learning agents," *Unpublished doctoral dissertation). Universidada Técnica de Lisboa,* 2013.

[84] Z. Deng *et al.*, "Deep structured models for group activity recognition," *arXiv preprint arXiv:1506.04191,* 2015, doi: https://dx.doi.org/10.48550/arXiv.1506.04191.
[85] M. McGrath, D. Howard, and R. Baker, "A lagrange-based generalised formulation for the equations of motion of simple walking models," *Journal of biomechanics,* vol. 55, pp. 139-143, 2017, doi: http://dx.doi.org/10.1016/j.jbiomech.2017.02.013.
[86] Baranes and P.-Y. Oudeyer, "Active learning of inverse models with intrinsically motivated goal exploration in robots," Robotics and Autonomous Systems, vol. 61, no. 1, pp. 49-73, 2013, doi: http://dx.doi.org/10.1016/j.robot.2012.05.008.
[87] K. Seepanomwan, V. G. Santucci, and G. Baldassarre, "Intrinsically motivated discovered outcomes boost user's goals achievement in a humanoid robot," in *2017 Joint IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*, 2017: IEEE, pp. 178-183, doi: http://dx.doi.org/10.1109/DEVLRN.2017.8329804.
[88] Péré, S. Forestier, O. Sigaud, and P.-Y. Oudeyer, "Unsupervised learning of goal spaces for intrinsically motivated goal exploration," arXiv preprint arXiv:1803.00781, 2018, doi: https://doi.org/10.48550/arXiv.1803.00781
[89] R. Zhao, X. Sun, and V. Tresp, "Maximum Entropy-Regularized Multi-Goal Reinforcement Learning," *arXiv preprint arXiv:1905.08786,* 2019, doi: https://doi.org/10.48550/arXiv.1905.08786.
[90] Ramamurthy, R., et al, "Novelty-guided reinforcement learning via encoded behaviors," in *2020 International Joint Conference on Neural Networks (IJCNN)*, 2020, pp. 1–8, doi: http://dx.doi.org/10.1109/IJCNN48605.2020.9206982 .
[91] Xu, H., et al. "Novelty is not surprise: Human exploratory and adaptive behavior in sequential decision-making," in *PLOS Computational Biology*, vol. 17, no. 6, pp. e1009070, 2021, doi: http://dx.doi.org/10.1371/journal.pcbi.1009070.
[92] K. Arulkumaran, M. P. Deisenroth, M. Brundage, and A. A. Bharath, "A brief survey of deep reinforcement learning," *arXiv preprint arXiv:1708.05866,* 2017, doi: https://dx.doi.org/10.1109/MSP.2017.2743240.
[93] V. Mnih *et al.*, "Asynchronous methods for deep reinforcement learning," in *International conference on machine learning*, 2016, pp. 1928-1937.
[94] T. Schaul, J. Quan, I. Antonoglou, and D. Silver, "Prioritized experience replay," *arXiv preprint arXiv:1511.05952,* 2015, doi: https://dx.doi.org/10.48550/arXiv.1511.05952.
[95] R. S. Sutton and A. G. Barto, *Introduction to reinforcement learning* (no. 4). MIT press Cambridge, 1998, doi: http://dx.doi.org/10.1109/TNN.1998.712192.
[96] N. Dilokthanakul, C. Kaplanis, N. Pawlowski, and M. Shanahan, "Feature control as intrinsic motivation for hierarchical reinforcement learning," *IEEE transactions on neural networks and learning systems,* 2019, doi: http://dx.doi.org/10.1109/TNNLS.2019.2891792.
[97] J. Schmidhuber, "Artificial curiosity based on discovering novel algorithmic predictability through coevolution," in *Proceedings of the 1999 Congress on Evolutionary Computation-CEC99 (Cat. No. 99TH8406)*, 1999, vol. 3: IEEE, pp. 1612-1618, doi: http://dx.doi.org/10.1109/CEC.1999.785467
[98] T. D. Kulkarni, K. Narasimhan, A. Saeedi, and J. Tenenbaum, "Hierarchical deep reinforcement learning: Integrating temporal abstraction and intrinsic motivation," in *Advances in neural information processing systems*, 2016, pp. 3675-3683.
[99] O. Nachum, S. S. Gu, H. Lee, and S. Levine, "Data-efficient hierarchical reinforcement learning," in *Advances in Neural Information Processing Systems*, 2018, pp. 3303-3313.
[100] Levy, R. Platt, and K. Saenko, "Hierarchical actor-critic," arXiv preprint arXiv:1712.00948, 2017.
[101] S. Vezhnevets *et al.*, "Feudal networks for hierarchical reinforcement learning," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, 2017: JMLR. org, pp. 3540-3549.
[102] D. J. Mankowitz, T. A. Mann, and S. Mannor, "Adaptive skills adaptive partitions (ASAP)," in *Advances in Neural Information Processing Systems*, 2016, pp. 1588-1596.
[103] P.-L. Bacon, J. Harb, and D. Precup, "The option-critic architecture," in *Thirty-First AAAI Conference on Artificial Intelligence*, 2017, doi: http://dx.doi.org/10.1609/aaai.v31i1.10916.
[104] V. Mnih *et al.*, "Human-level control through deep reinforcement learning," *Nature,* vol. 518, no. 7540, p. 529, 2015, doi: http://dx.doi.org/10.1038/nature14236.
[105] H. Van Hasselt, A. Guez, and D. Silver, "Deep reinforcement learning with double q-learning," in *Thirtieth AAAI conference on artificial intelligence*, 2016, doi: http://dx.doi.org/10.1609/aaai.v30i1.10295.
[106] T. P. Lillicrap *et al.*, "Continuous control with deep reinforcement learning," *arXiv preprint arXiv:1509.02971,* 2015, doi: https://dx.doi.org/10.48550/arXiv.1509.02971.
[107] N. Heess *et al.*, "Emergence of locomotion behaviours in rich environments," *arXiv preprint arXiv:1707.02286*, 2017, doi: https://dx.doi.org/10.48550/arXiv.1707.02286.
[108] Rajeswaran et al., "Learning complex dexterous manipulation with deep reinforcement learning and demonstrations," arXiv preprint arXiv:1709.10087, 2017, doi: https://dx.doi.org/10.48550/arXiv.1709. 10087.
[109] S. Gu, E. Holly, T. Lillicrap, and S. Levine, "Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates," in *2017 IEEE international conference on robotics and automation (ICRA)*, 2017: IEEE, pp. 3389-3396.
[110] T. Lesort, N. Díaz-Rodríguez, J.-F. Goudou, and D. Filliat, "State representation learning for control: An overview," *Neural Networks,* vol. 108, pp. 379-392, 2018, doi: http://dx.doi.org/10.1016/j.neunet.2018 .07.006.
[111] V. Nair, V. Pong, M. Dalal, S. Bahl, S. Lin, and S. Levine, "Visual reinforcement learning with imagined goals," in Advances in Neural Information Processing Systems, 2018 ,pp. 9191-9200.
[112] V. H. Pong, M. Dalal, S. Lin, A. Nair, S. Bahl, and S. Levine, "Skew-Fit: State-Covering Self-Supervised Reinforcement Learning," *arXiv preprint arXiv:1903.03698,* 2019, doi: https://dx.doi.org/10.48550/arXiv.1903.03698.
[113] Florensa, Y. Duan, and P. Abbeel, "Stochastic neural networks for hierarchical reinforcement learning," *arXiv preprint*

of the genetic and evolutionary computation conference, 2018, pp. 21–28, doi: http://dx.doi.org/10.1145/3205455.3205650.

[130] Wang, A., et al, "A unifying framework for social motivation in human-robot interaction," in *The AAAI*

[131] *2020 Workshop on Plan, Activity, and Intent Recognition (PAIR 2020)*, 2020.

[132] K. Iinuma, K. Kogiso. "Emotion-involved human decision-making model," in *Mathematical and Computer Modelling of Dynamical Systems*, vol. 27, no. 1, pp. 543–561, 2021, doi: http://dx.doi.org/10.1080/ 13873954.2021.1986846.

[133] Kirtay, M., et al. "Emotion as an emergent phenomenon of the neurocomputational energy regulation mechanism of a cognitive agent in a decision-making task," in *Adaptive Behavior*, vol. 29, no. 1, pp. 55–71, 2021, doi: http://dx.doi.org/10.1177/ 1059712319880649.

[134] J. Taverner, E. Vivancos, V. Botti. "A Multidimensional Culturally Adapted Representation of Emotions for Affective Computational Simulation and Recognition," in *IEEE Transactions on Affective Computing*, vol. 14, no. 01, pp. 761–772, 2023, doi: http://dx.doi.org/10.1109/TAFFC.2020.3030586.

[135] Ren, Z., et al. "Exploration via hindsight goal generation," in *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[136] Bing, Z., et al. "Complex robotic manipulation via graph-based hindsight goal generation," in *IEEE transactions on neural networks and learning systems*, vol. 33, no. 12, pp. 7863–7876, 2021, doi: http://dx.doi.org/10.1109/TNNLS.2021.3088947.

[137] J. Kim, Y. Seo, J. Shin. "Landmark-guided subgoal generation in hierarchical reinforcement learning," in *Advances in neural information processing systems*, vol. 34, pp. 28336–28349, 2021.

[138] Bagaria, A., et al. "Scaling goal-based exploration via pruning proto-goals," in *arXiv preprint arXiv:2302.04693*, 2023, doi: https://dx.doi.org/ 10.48550/arXiv.2302.04693.

[139] Park, S., et al, "Offline Goal-Conditioned RL with Latent States as Actions," in ICML Workshop on New Frontiers in Learning, Control, and Dynamical Systems, 2023.

[140] L. Wu, K. Chen. "Goal Exploration via Adaptive Skill Distribution for Goal-Conditioned Reinforcement Learning," in arXiv preprint arXiv:2404.12999, 2024, doi: https://dx.doi.org/10.48550/arXiv.2404.12999.

[141] M. Hameed, M. Khan, A. Schwung. "Curiosity Based Reinforcement Learning on Robot Manufacturing Cell," in *arXiv preprint arXiv:2011.08743*, 2020, doi: https://dx.doi.org/10.48550/arXiv.2011.08743.

[142] N. Bougie, R. Ichise. "Fast and slow curiosity for high-level exploration in reinforcement learning," in *Applied Intelligence*, vol. 51, pp. 1086–1107, 2021, doi: http://dx.doi.org/10.1007/s10489-020-01849-3.

[143] Mazzaglia, P., et al, "Curiosity-driven exploration via latent bayesian surprise," in *Proceedings of the AAAI conference on artificial intelligence*, 2022, pp. 7752–7760, doi: http://dx.doi.org/10.1609/aaai.v36i7.20743.

[144] Jarrett, D., et al. "Curiosity in hindsight: intrinsic exploration in stochastic environments," ,2023.

arXiv:1704.03012, 2017, doi: https://dx.doi.org/10.48550/arXiv.1704.03012.

[114] Eysenbach, A. Gupta, J. Ibarz, and S. Levine, "Diversity is all you need: Learning skills without a reward function," *arXiv preprint arXiv:1802.06070,* 2018, doi: https://dx.doi.org/10.48550/arXiv. 1802. 06070.

[115] Warde-Farley, T. Van de Wiele, T. Kulkarni, C. Ionescu, S. Hansen, and V. Mnih, "Unsupervised control through non-parametric discriminative rewards," *arXiv preprint arXiv:1811.11359,* 2018, doi: https://dx.doi.org/10.48550/arXiv.1811.11359.

[116] J. D. Co-Reyes, Y. Liu, A. Gupta, B. Eysenbach, P. Abbeel, and S. Levine, "Self-consistent trajectory autoencoder: Hierarchical reinforcement learning with trajectory embeddings," *arXiv preprint arXiv:1806.02813,* 2018, doi: https://doi.org/10.48550/ arXiv.1806.02813.

[117] Ozbilge, E. Ozbilge. "Fusion of Novelty Detectors Using Deep and Local Invariant Visual Features for Inspection Task," in *IEEE Access*, vol. 10, pp. 121032–121047, 2022, doi: http://dx.doi.org/10.1109/ ACCESS .2022.3222810.

[118] Modirshanechi, A., et al. "The curse of optimism: a persistent distraction by novelty," in *bioRxiv*, pp. 2022–07, 2022.

[119] Le, H., et al, "Beyond Surprise: Improving Exploration Through Surprise Novelty.," in *AAMAS*, 2024, pp. 1084–1092.

[120] H. Jiang, Z. Ding, Z. Lu. "Settling Decentralized Multi-Agent Coordinated Exploration by Novelty Sharing," in *arXiv preprint arXiv:2402.02097*, 2024, doi: https://dx.doi.org/10.48550/arXiv.2402.02097.

[121] R. Zhao, P. Abbeel, S. Tiomkin. "Efficient online estimation of empowerment for reinforcement learning," in *arXiv preprint arXiv:2007.07356*, 2020, doi: https://dx.doi.org/10.48550/arXiv.2412.07762.

[122] Choi, J., et al. "Variational empowerment as representation learning for goal-based reinforcement learning," in *arXiv preprint arXiv:2106.01404*, 2021, doi: https://dx.doi.org/10.48550/arXiv.2106.01404.

[123] Brändle, F., et al. "Intrinsically motivated exploration as empowerment,",2022.

[124] Dai, S., et al. "An empowerment-based solution to robotic manipulation tasks with sparse rewards," in *Autonomous Robots*, vol. 47, no. 5, pp. 617–633, 2023, doi: http://dx.doi.org/10.1007/s10514-023-10087-8.

[125] Brändle, F., et al. "Empowerment contributes to exploration behaviour in a creative video game," in *Nature Human Behaviour*, vol. 7, no. 9, pp. 1481–1489, 2023, doi: http://dx.doi.org/10.1038/s41562-023-01661-2.

[126] Becker-Ehmck, P., et al, "Exploration via empowerment gain: Combining novelty, surprise and learning progress," in *ICML 2021 Workshop on Unsupervised Reinforcement Learning*, 2021.

[127] Heiden, T., et al, "Reliably Re-Acting to Partner's Actions with the Social Intrinsic Motivation of Transfer Empowerment," in *ALIFE 2022: The 2022 Conference on Artificial Life*, 2022.

[128] Andrychowicz, M., et al. "Hindsight experience replay," in *Advances in neural information processing systems*, vol. 30, 2017.

[129] Guzzi, J., et al, "A model of artificial emotions for behavior-modulation and implicit coordination in multi-robot systems," in *Proceedings*

[145] C. Zhou, T. Machado, C. Harteveld, "Cautious curiosity: a novel approach to a human-like gameplay agent," in *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, 2023, pp. 370–379, doi: http://dx.doi.org/10.1609/aiide.v19i1.27533.

[146] C. Sun, H. Qian, C. Miao, "CUDC: A Curiosity-Driven Unsupervised Data Collection Method with Adaptive Temporal Distances for Offline Reinforcement Learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024, pp. 15145–15153, doi: http://dx.doi.org/10.1609/aaai.v38i13.29437.

[147] Dewan, S., et al. "Curiosity & Entropy Driven Unsupervised RL in Multiple Environments," in *arXiv preprint arXiv:2401.04198*, 2024, doi: https://dx.doi.org/10.48550/arXiv.2401.04198.

[148] P. Oudeyer, F. Kaplan, V. Hafner. "Intrinsic motivation systems for autonomous mental development," in *IEEE transactions on evolutionary computation*, vol. 11, no. 2, pp. 265–286, 2007, doi: http://dx.doi.org/ 10.1109/TEVC.2006.890271.

[149] S. Hart, R. Grupen. "Learning generalizable control programs," in *IEEE Transactions on Autonomous Mental Development*, vol. 3, no. 3, pp. 216–231, 2011, doi: http://dx.doi.org/10.1109/TAMD.2010.2103311.

[150] N. Duminy, D. Duhaut, undefined. others, "Strategic and interactive learning of a hierarchical set of tasks by the Poppy humanoid robot," in *2016 Joint IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*, 2016, pp. 204–209, doi: http://dx.doi.org/10.1109/DEVLRN.2016.7846820.

[151] Gerken, M. Spranger, "Continuous Value Iteration (CVI) Reinforcement Learning and Imaginary Experience Replay (IER) for learning multi-goal, continuous action and state space controllers," in 2019 International Conference on Robotics and Automation (ICRA), 2019, pp. 7173–7179, doi: http://dx.doi.org/10.1109/ICRA.2019.8794347.

[152] R. Rayyes, H. Donat, J. Steil, "Hierarchical interest-driven goal babbling for efficient bootstrapping of sensorimotor skills," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, 2020, pp. 1336–1342, doi: http://dx.doi.org/10.1109/ ICRA40945.2020.9196763.

[153] Huang, S., et al. "Learning gentle object manipulation with curiosity-driven deep reinforcement learning. arXiv 2019," in *arXiv preprint arXiv:1903.08542*, doi: https://dx.doi.org/10.48550/arXiv.1903.08542

[154] Schulman, J., et al. "Proximal policy optimization algorithms," in arXiv preprint arXiv:1707.06347, 2017, doi: https://dx.doi.org/10.48550/arXiv.1707.06347.

[155] J. Lee, K. Toutanova. "Pre-training of deep bidirectional transformers for language understanding," in arXiv preprint arXiv:1810.04805, vol. 3, no. 8, 2018, doi: https://dx.doi.org/10.48550/arXiv.1810.04805.

[156] Chen, T., et al, "A simple framework for contrastive learning of visual representations," in International conference on machine learning, 2020, pp. 1597–1607.

[157] Frans, K., et al. "Unsupervised Zero-Shot Reinforcement Learning via Functional Reward Encodings," in arXiv preprint arXiv:2402.17135, 2024, doi: https://dx.doi.org/10.48550/arXiv.2402.17135.

A. Radford. "Improving language understanding by generative pre-training,", 2018.

[158] Tarvainen, H. Valpola. "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Advances in neural information processing systems*, vol. 30, 2017.

[159] Burda, Y., et al. "Exploration by random network distillation," in *arXiv preprint arXiv:1810.12894*, 2018, doi: https://dx.doi.org/10.48550/arXiv.1810.12894.

[160] Haarnoja, T., et al, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *International conference on machine learning*, 2018, pp. 1861–1870.

[161] Hafner, D., et al. "Dream to control: Learning behaviors by latent imagination," in arXiv preprint arXiv:1912.01603, 2019, doi: https://dx.doi.org/ 10.48550/arXiv.1912.01603.

[162] Schrittwieser, J., et al. "Mastering atari, go, chess and shogi by planning with a learned model," in Nature, vol. 588, no. 7839, pp. 604–609, 2020, doi: http://dx.doi.org/ 10.1038/s41586-020-03051-4.

[163] OpenAI, O., et al. "Asymmetric self-play for automatic goal discovery in robotic manipulation," in *arXiv preprint arXiv:2101.04882*, 2021, doi: https://dx.doi.org/10.48550/arXiv.2101.04882.

[164] Cao, J., et al. "Hierarchical multi-agent reinforcement learning for cooperative tasks with sparse rewards in continuous domain," in *Neural Computing and Applications*, vol. 36, no. 1, pp. 273–287, 2024, doi: http://dx.doi.org/10.1007/s00521-023-08882-6.

[165] J. Lehman, K. Stanley. "Abandoning objectives: Evolution through the search for novelty alone," in *Evolutionary computation*, vol. 19, no. 2, pp. 189–223, 2011, doi: http://dx.doi.org/10.1162/EVCO_a_00025.

[166] Nguyen, D., et al, "Social Motivation for Modelling Other Agents under Partial Observability in Decentralised Training.," in *IJCAI*, 2023, pp. 4082–4090, doi: http://dx.doi.org/10.24963/ijcai.2023/454.

[167] Duminy, N., et al. "Intrinsically motivated open-ended multi-task learning using transfer learning to discover task hierarchy," in *Applied Sciences*, vol. 11, no. 3, pp. 975, 2021, doi: http://dx.doi.org/10.3390/app11030975.