دوره هجدهم، شماره بهار ۱۴۰۴

مجله فناوری اطلاعات در طراحی مهندسی

Information Technology in Engineering Design

http://sanad.iau.ir/journal/ited

مروری بر سیستم‌های یادگیری مبتنی برکنجکاوی در هوش مصنوعی

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

(۱) **گروه مهندسی کامپیوتر، واحد تهران مرکزی، دانشگاه آزاد اسلامی، تهران، ایران**

(۲) **گروه مهندسی فیزیک و انرژی، دانشگاه صنعتی امیرکبیر، تهران، ایران***

(۳) **گروه مهندسی مکانیک، واحد شیراز، دانشگاه شیراز، شیراز، ایران**

(۴) **گروه مهندسی کامپیوتر، واحد تهران مرکزی، دانشگاه آزاد اسلامی، تهران، ایران**

(تاریخ دریافت: ۱۴۰۳/۰۶/۲۷ تاریخ پذیرش: ۱۴۰۳/۰۹/۱۳)

چکیده

یکی از جنبه‌های کلیدی که می‌تواند هوش مصنوعی را به سطح بالاتری از توانایی برساند، کنجکاوی است. همانند انسان‌ها، در هوش مصنوعی نیز، کنجکاوی می‌تواند به عنوان یک مکانیسم کلیدی برای بهبود یادگیری فعال و اکتشاف در محیط‌های پیچیده و ناشناخته عمل کند. در این مقاله مروری، تلاش‌ها برای مدل‌سازی و شبیه‌سازی کنجکاوی در ماشین‌ها به منظور ایجاد سیستم‌هایی که بتوانند به طور خودکار و مستقل رفتارهای اکتشافی از خود نشان دهند مورد بررسی قرار گرفته است. در این پژوهش، با بررسی مطالعات روانشناختی در مورد کنجکاوی و مدل‌های محاسباتی موجود در هوش مصنوعی، به دنبال درک عمیق‌تری از مفهوم کنجکاوی و چگونگی شبیه‌سازی آن در ماشین‌ها هستیم. همچنین، به بررسی مزایا و محدودیت‌های رویکردهای موجود پرداخته‌ایم. نتایج این پژوهش نشان می‌دهد که کنجکاوی می‌تواند به عنوان یک عامل مهم در تسریع یادگیری، افزایش توانایی تعمیم‌پذیری مدل‌ها و بهبود عملکرد در وظایف چالش‌برانگیز عمل کند. همچنین با معرفی یک متغیر هم‌سنجش جدید به نام «پوشش فضا»، به دنبال تحقیقات جدیدی برای این نوع مدل‌سازی کنجکاوی در هوش مصنوعی هستیم. در نهایت، در کنار برشمردن برخی کاربردها، با ارائه پیشنهاداتی برای تحقیقات آینده، تلاش کرده‌ایم تا مسیری را برای توسعه سیستم‌های هوش مصنوعی کنجکاوتر و قدرتمندتر هموار کنیم.

کلمات کلیدی: کنجکاوی، هوش مصنوعی، یادگیری ماشین، یادگیری تقویتی، متغیرهای هم‌سنجش، پوشش فضا

*عهده‌دار مکاتبات:

سعید ستایشی

نشانی: گروه مهندسی فیزیک و انرژی، دانشگاه صنعتی امیرکبیر، تهران، ایران

پست الکترونیکی: setayeshi@aut.ac.ir

۱- مقدمه

در عصر حاضر، هوش مصنوعی (AI[۱]) به یکی از مهم‌ترین و پویاترین حوزه‌های پژوهشی تبدیل شده است که هدف آن توسعه سیستم‌هایی است که قادر به انجام وظایف پیچیده‌ای باشند که به‌طور سنتی به هوش انسانی نیاز دارند. این سیستم‌ها با الهام از عملکرد مغز و فرآیندهای شناختی انسان، تلاش می‌کنند تا به سطوح جدیدی از توانایی در یادگیری، استدلال و تصمیم‌گیری برسند. کاربردهای گسترده AI در زمینه‌هایی مانند تشخیص تصویر [1]، پردازش زبان طبیعی [2]، رباتیک [3]، خودروهای خودران [4] و سیستم‌های توصیه‌گر[۲] [5,6] به وضوح نقش حیاتی این فناوری را در دنیای مدرن نشان می‌دهد [7]. قابلیت سیستم‌های AI نیز باید با کارکردهای زیستی و روانشناختی یا شناخت در سطح انسان تقویت شود تا بتواند مزایای هوش انسانی مانند سازگاری سریع، بهره‌وری نمونه بالا و تفسیر قابل اعتماد را به ارث ببرد [8]. یکی از جنبه‌های کلیدی که می‌تواند AI را به سطح بالاتری از توانایی برساند و در سال‌های اخیر توجه بسیاری از پژوهشگران را به خود جلب کرده است، کنجکاوی است. کنجکاوی به عنوان یک عنصر اساسی شناخت، به طور طبیعی انگیزه‌ای درونی را فراهم می‌کند که انسان‌ها را برای کشف اطلاعات جالب و مفید به کاوش در جهان ترغیب می‌کند. تقاضای سیری‌ناپذیر برای اطلاعات می‌تواند در نهایت یادگیری، تصمیم‌گیری و رشد سالم را شکل داده و تقویت کند [8]. در AI نیز، کنجکاوی به عنوان عاملی کلیدی برای یادگیری فعال، اکتشاف و نوآوری در نظر گرفته می‌شود. سیستم‌های AI کنجکاو، به جای اینکه صرفاً به انجام وظایف از پیش تعریف شده بپردازند، به طور فعال به دنبال کشف اطلاعات جدید و گسترش دانش خود هستند. در حوزه AI، تلاش‌ها برای مدل‌سازی و شبیه‌سازی کنجکاوی در ماشین‌ها به منظور ایجاد سیستم‌هایی که بتوانند به‌طور خودکار و مستقل رفتارهای اکتشافی از خود نشان دهند، به تحقیقات گسترده‌ای منجر شده است [9]. این رویکرد نه تنها می‌تواند منجر به بهبود کارایی یادگیری در سیستم‌های AI شود، بلکه می‌تواند باعث افزایش توانایی این سیستم‌ها در حل مسائل پیچیده و تعامل با محیط‌های پویا و متغیر گردد [10]. از دیدگاه یادگیری ماشین، کنجکاوی به عنوان اصول الگوریتمی برای تمرکز یادگیری بر الگوهای جدید و قابل یادگیری در مقابل نویزهای نامنظم پیشنهاد شده است. این اصول در تسریع یادگیری و ساخت رباتیک توسعه‌ای نظارت نشده موفق بوده‌اند [11]. از دیدگاه محاسباتی، کنجکاوی می‌تواند به عنوان یک نیروی محرک برای جستجوی دانش جدید و بهبود مدل‌های موجود عمل کرده و در الگوریتم‌های یادگیری ماشین می‌تواند باعث تسریع فرآیند یادگیری و افزایش توانایی تعمیم‌پذیری مدل‌ها شود [12]. برای مثال، در یادگیری تقویتی

[۱] Artificial Intelligence

[۲] recommender systems

(RL[1])، کنجکاوی به عنوان پاداش‌های درونی برای تشویق عامل‌ها به کاوش در محیط و کشف حالات جدید تعریف می‌شود [13]. این رویکرد می‌تواند منجر به بهبود عملکرد در وظایف چالش‌برانگیز شود و به سیستم‌ها اجازه دهد تا با محیط‌های ناشناخته و پیچیده بهتر سازگار شوند [14]. انواع مختلفی از الگوریتم‌های یادگیری مبتنی بر کنجکاوی پیشنهاد شده است تا در وظایف کلاسیک AI مانند طبقه‌بندی [15]، توصیه [16] و بهینه‌سازی [17] اجرا شود، که در آن هدف نهایی بهبود کارایی یادگیری و توانمندسازی عامل‌های هوشمند برای یادگیری به شیوه‌ای انسانی است. در حوزه RL، کنجکاوی به عنوان پاداش‌های درونی برای کمک به فرآیند یادگیری کمی‌سازی می‌شود. در نتیجه، عامل‌ها به سمت حالت‌های جدید [18] تشویق می‌شوند یا اقداماتی برای کاوش در مناطق بسیار نامطمئن بر اساس دانش موجود خود در مورد محیط انجام می‌دهند [19,20]. با انگیزش کنجکاوی، عامل‌ها می‌توانند محیط را کاوش کرده و مهارت‌های متنوعی را بیاموزند که ممکن است حتی در موقعیت‌های ندیده مفید باشد و ویژگی مطلوب رفتارهای اکتشافی انسانی را نشان دهد. متأسفانه، انسان‌ها اغلب پیچیده‌تر از چیزی هستند که یک مدل یادگیری ماشین می‌تواند توصیف کند. در نتیجه بازگشت به تحقیقات روانشناسی برای درک چگونگی برانگیخته شدن کنجکاوی انسان مفید است [11]. تحقیقات روانشناختی نشان می‌دهد که عوامل مختلفی مانند تازگی[2]، تضاد، عدم قطعیت[3] و پیچیدگی[4] می‌توانند کنجکاوی را تحریک کنند [8]. این عوامل می‌توانند به عنوان متغیرهای هم‌سنجش در مدل‌سازی کنجکاوی محاسباتی مورد استفاده قرار گیرند تا به توسعه الگوریتم‌هایی که قادر به تشخیص و پاسخ‌گویی به این محرک‌ها هستند، کمک کنند [8]. به‌طور خاص، استفاده از نظریه‌های روانشناختی برای طراحی سیستم‌های کنجکاو، می‌تواند به توسعه سیستم‌های AI که قادر به تعامل طبیعی‌تر با انسان‌ها و محیط‌های واقعی هستند، منجر شود [22]. در این مقاله، تلاش می‌کنیم تا با بررسی نقش کنجکاوی در AI و استفاده از مدل‌های محاسباتی مختلف، به درک عمیق‌تری از این مفهوم دست یابیم و راهکارهایی برای بهبود سیستم‌های AI ارائه دهیم. همچنین به بررسی مزایا و محدودیت‌های رویکردهای موجود پرداخته و زمینه‌های پژوهشی جدیدی را پیشنهاد خواهیم کرد که می‌توانند به تقویت قابلیت‌های اکتشافی و یادگیری سیستم‌های AI کمک کنند.

[1] Reinforcement Learning
[2] Novelty
[3] Uncertainty
[4] Complexity

۱-۱- کارهای اصلی

مهم‌ترین دستاوردهای این مطالعه عبارتند از:

- تحقیق پیرامون مفهوم کنجکاوی در روانشناسی و ارتباط آن با انگیزش درونی
- بررسی متغیرهای هم‌سنجش به عنوان محرک‌های اصلی کنجکاوی
- مطالعه کنجکاوی و متغیرهای هم‌سنجش در AI
- معرفی متغیر هم‌سنجش جدیدی به نام پوشش فضا برای کنجکاوی مصنوعی
- بررسی و دسته‌بندی تحقیقات مدل‌های محاسباتی در کنجکاوی مبتنی بر متغیرهای هم‌سنجش
- آنالیز مفهوم، مزایا و چالش‌ها(محدودیت‌ها) و معیارهای پاداش در این تحقیقات

۱-۲- انتخابات مقالات

از میان حجم گسترده‌ای از پژوهش‌های انجام شده در این حوزه، مقالات مورد استفاده در این پژوهش با توجه به معیارهای مشخصی انتخاب شده‌اند که عبارتند از:

- **کلید واژه‌ها**: در مرحله نخست پژوهش، با بهره‌گیری از پایگاه داده‌ی گوگل اسکالر[۱]، مقالاتی که حداقل یکی از مفاهیم کنجکاوی، یادگیری ماشین، و مدل‌های محاسباتی را در کنار متغیرهای کنجکاوی نظیر نوآوری، پیچیدگی و عدم قطعیت و ... بررسی می‌کردند، استخراج شدند. برای این منظور از الگوی زیر استفاده شده است:

[curiosity_concept] + [collative_variables] + [machine_learning_term] and/or [model_term]

. curiosity_concept: curiosity or curiosity-driven or artificial curiosity
. collative_variables: novelty or change or complexity or uncertainty or surprisingness or incongruity or conflict
. machine_learning_term: learning or reinforcement learning or machine learning
. model_term: model or computational model or structure

شایان ذکر است که برخی از پژوهش‌های مورد بررسی، به جای ارائه یک تعریف صریح از کنجکاوی از منظر روانشناسی، به معرفی مدل‌هایی پرداخته‌اند که عملکرد آن‌ها به طور ضمنی مفهوم کنجکاوی را بازتاب می‌دهد. به ویژه در مطالعات مربوط به زیر بخش «پوشش فضایی»، این تطابق به وضوح قابل مشاهده است. در این دسته از پژوهش‌ها، اگرچه محققان لزوماً از واژه «کنجکاوی» استفاده نکرده‌اند، اما مدل‌های پیشنهادی آن‌ها به خوبی توانسته‌اند رفتارهای اکتشافی و جستجوی اطلاعات جدید را شبیه‌سازی کنند که از ویژگی‌های بارز کنجکاوی انسانی است.

[۱] Google Scholar

- **سال و ارجاع:** در گردآوری منابع این پژوهش، تلاش شده است تا تمرکز بر مقالات منتشر شده در سال‌های اخیر باشد. با این حال، به منظور ارائه یک مرور جامع بر موضوع، برخی از مقالات بنیادین و تأثیرگذار که پیشگامان این حوزه محسوب می‌شوند، علی‌رغم قدمت انتشارشان، در فهرست منابع گنجانده شده است. همچنین، مقالاتی که بیشترین تعداد ارجاعات را به خود اختصاص داده‌اند، به عنوان منابع کلیدی مورد توجه قرار گرفته‌اند.
- **محل چاپ:** جهت تضمین اعتبار یافته‌های پژوهش، منابع مورد استفاده در این مطالعه از میان مقالات چاپ شده در ژورنال‌های معتبر بین‌المللی همچون IEEE، ACM، Springer و Oxford University Press انتخاب شده است. همچنین، کنفرانس‌های برجسته‌ای در حوزه رایانه و AI از جمله AAAI، ICLR، ICML، CVPR و NeurIPS به عنوان منابع اصلی مقالات کنفرانسی مدنظر قرار گرفته‌اند. با توجه به نقش فزاینده‌ی آرشیوهای پیش‌چاپ مانند arXiv در انتشار سریع یافته‌های پژوهشی، برخی از مقالات منتشر شده در این پلتفرم نیز در این پژوهش مورد بررسی قرار گرفته‌اند.

۱-۳- ساختار مقاله

این مقاله به شرح زیر سازماندهی شده است. در بخش ۲ مروری کوتاه بر مطالعات روانشناختی کنجکاوی انسان ارائه می‌دهیم و رابطه کنجکاوی با انگیزش درونی بررسی شده و به جایگاه متغیرهای هم‌سنجش در کنجکاوی می پردازیم. در ادامه در بخش ۳، مدل‌های محاسباتی موجود کنجکاوی را در AI از دریچه متغیرهای هم‌سنجش مورد بررسی و ارزیابی قرار داده‌ایم. محدودیت‌ها، کاربردها و جهت‌گیری تحقیقات آینده این حوزه را در بخش ۴ ارائه کرده‌ایم. در نهایت، در بخش ۵، به جمع بندی می‌پردازد.

۲- کنجکاوی در روانشناسی

در چند دهه گذشته، تحقیقات گسترده و تجزیه و تحلیل‌های تجربی بر روی مکانیسم‌های زیستی کنجکاوی در مطالعات رفتاری، نوروبیولوژیکی و شناختی انجام شده است [23,26]. در ابتدا کنجکاوی به عنوان هیجانی نزدیک به ترس طبقه‌بندی می‌شد [27]. در آن زمان، نظریه‌های مبتنی بر «سائق[1]» محبوبیت پیدا کردند، در حالی که کنجکاوی ممکن بود با احساس ناخوشایند محرومیت همراه باشد و از طریق مجموعه‌ای از رفتارهای اکتشافی کاهش یابد [28,29]. منظور از سائق حالتی درونی است که موجود زنده را به فعالیت وا می‌دارد یا رفتار خاصی را برمی‌انگیزد.

[1] drive

پس از آن، اولین چارچوب نظری کامل مبتنی بر سائق توسط لورنز[1] [30] ارائه شد. در نیمه دوم قرن بیستم، هب[2] [31] و پیاژه[3] [32] در تحقیقاتی به طور مستقل نشان دادند که کنجکاوی با نقض انتظار بر اساس دانش موجود ایجاد می‌شود. نظریه شایستگی [33] (گسترش یافته در [34]) بیان داشت که کنجکاوی انگیزه‌ای برای «شایستگی» است که ناشی از میل به تسلط بر محیط است. البته بیشتر روانشناسان معتقدند که کنجکاوی یک انگیزش درونی[4] است که محرک رشد شناختی انسان و حیوان است [31]. در دهه ۱۹۵۰، اکثر تحقیقات در مورد کنجکاوی بر زیربنای روانشناختی آن متمرکز بود [35]. برلین[5] [33] کنجکاوی را در دو طیف دسته بندی کرد: ۱) از کنجکاوی ادراکی (حسی) به کنجکاوی معرفتی[6] (شناختی)، و ۲) از کنجکاوی خاص[7] به کنجکاوی گسترده (گوناگون)[8]. کنجکاوی ادراکی ریشه در حواس ما دارد و در همه موجودات زنده از جمله حیوانات دیده می‌شود (مانند حس لامسه، بینایی، چشایی). کنجکاوی معرفتی سطح بالاتری دارد و مختص انسان است. این کنجکاوی به دنبال کسب دانش و اطلاعات جدید است. کنجکاوی خاص وقتی برانگیخته می‌شود که به دنبال پاسخ سوال خاصی باشیم. کنجکاوی گسترده یک انگیزه کلی برای جستجوی اطلاعات بدون جهت خاص است و عمدتاً برای رفع خستگی به کار می‌رود. اسپیلبرگر[9] و استار[10] [36] کنجکاوی گسترده را با محرک‌های «سطح پایین» و کنجکاوی خاص را با محرک‌های «سطح بالا» مرتبط کردند .با این حال، اشمیت[11] و لاهرودی[12] [37] با این تصور که کنجکاوی می‌تواند گسترده باشد، مخالف هستند آن ها به جای کنجکاوی گسترده، به تمایل کلی به دانش به عنوان «جستجوگری» اشاره کردند. با این وجود، اجماع کلی روانشناسان بر رابطه نزدیک بین کنجکاوی و شناخت، به عنوان محرکی برای کاوش محرک‌های جدید یا علاقه به دانش اشاره دارد [11].

۲-۱- رابطه کنجکاوی با انگیزش درونی

انگیزش درونی و کنجکاوی مفاهیمی نزدیک به هم در حوزه روانشناسی هستند، به ویژه در زمینه یادگیری و رفتار. انگیزش درونی به معنای انجام فعالیتی به خاطر رضایت ذاتی آن است، نه به خاطر نتیجه جداگانه یا پاداش خارجی

[1] Lorenz
[2] Hebb
[3] Piaget
[4] Intrinsic motivation
[5] Berlyne
[6] epistemic
[7] specific
[8] diversive
[9] Spielberger
[10] Starr
[11] Schmitt
[12] Lahroodi

[38]. از سوی دیگر، کنجکاوی تمایل به کسب دانش یا تجربیات جدید است و اغلب تحت تأثیر علاقه‌ی طبیعی یا شیفتگی به موضوع موردنظر شکل می‌گیرد. تحقیقات نشان می‌دهد که کنجکاوی می‌تواند به طور قابل توجهی انگیزش درونی را افزایش دهد و محرک داخلی قوی برای کاوش و یادگیری فراهم کند. طبق نظریه خودتعیینی[1] دسی[2] و رایان[3] [38]، فعالیت‌هایی که نیازهای روانشناختی اساسی برای خودمختاری، شایستگی و ارتباط را برآورده می‌کنند، به احتمال زیاد انگیزش درونی را افزایش می‌دهند. کنجکاوی این نیازها را با ترویج حس خودمختاری از طریق کاوش خودهدایت شده، پرورش شایستگی با تشویق به تسلط بر مهارت‌ها و دانش جدید و تقویت ارتباط برآورده می‌کند [38]. علاوه بر این، مطالعات تجربی نشان داده‌اند که کنجکاوی نه تنها انگیزش درونی را تحریک می‌کند بلکه عملکرد شناختی و تحصیلی را نیز بهبود می‌بخشد [39]. علاوه بر این، تحقیق [40] نشان می‌دهد کودکانی که سطوح بالایی از کنجکاوی را نشان می‌دهند، انگیزش درونی بیشتری دارند که منجر به یادگیری مؤثرتر و دستاوردهای تحصیلی بالاتر در طول زمان می‌شود. تحقیقات علوم اعصاب نیز از ارتباط بین انگیزش درونی و کنجکاوی حمایت می‌کند. مطالعات با استفاده از تصویربرداری تشدید مغناطیسی عملکردی[4] نشان داده‌اند که کنجکاوی، نواحی مغز مرتبط با پاداش و حافظه، مانند میان‌مغز[5] و هیپوکامپ[6] را فعال می‌کند. این نشان می‌دهد که مکانیسم‌های عصبی زمینه‌ساز انگیزش درونی به شدت به مکانیسم‌هایی که کنجکاوی و پردازش پاداش را کنترل می‌کنند، مرتبط هستند [41].

۲-۲- متغییرهای هم‌جنس

متغیرهای هم‌سنجش[7]، مانند پیچیدگی، تازگی، ابهام[8] و شگفتی[9]، نقش مهمی در تحریک کنجکاوی ایفا می‌کنند. طبق نظریه برلین، این متغیرها ویژگی‌های محرک‌هایی هستند که با ایجاد وضعیتی از عدم قطعیت و برانگیختگی در فرد، کنجکاوی را برانگیخته می‌کنند. این برانگیختگی منجر به رفتار کاوشگرانه می‌شود، زیرا فرد می‌خواهد عدم قطعیت را برطرف کرده و اطلاعات بیشتری در مورد محرک به دست آورد [42]. علاوه بر این، رابطه بین متغیرهای هم‌سنجش و کنجکاوی در AI کاربردهای عملی دارد. همانطور که در [11] پیشنهاد شده است، نظریه کنجکاوی روانشناختی ارائه

[1] Self-Determination Theory
[2] Deci
[3] Ryan
[4] Functional magnetic resonance imaging (fMRI)
[5] midbrain
[6] hippocampus
[7] Collative Variables
[8] Ambiguity
[9] Surprise

شده توسط برلین [13] می‌تواند به عنوان مبنای اندازه‌گیری محاسباتی کنجکاوی مصنوعی با تعیین شدت محرک (دریافت شده از محیط) با استفاده از یک یا چند متغیر هم‌سنجش معرفی شود. در بخش بعدی، متغیرهای کلیدی مورد استفاده در مدل‌سازی کمی کنجکاوی مصنوعی تشریح خواهند شد. این متغیرها نقش محوری در ساختار چارچوب یکپارچه‌ای دارند که به منظور اندازه‌گیری دقیق و جامع انواع مختلف کنجکاوی مصنوعی طراحی شده است.

۳- کنجکاوی در هوش مصنوعی

کنجکاوی در AI به عنوان یک جزء حیاتی برای توانمندسازی عامل‌ها در کاوش خودکار محیط‌های خود و کسب دانش جدید ظهور کرده است. اهمیت غیرقابل انکار کنجکاوی برای دستیابی نهایی به دانش مفید، محققان AI را به توسعه الگوریتم‌های یادگیری کنجکاو ترغیب کرده است. این تلاش‌ها منجر به تنوع زیادی از مکانیسم‌ها برای فعال کردن یادگیری مبتنی بر کنجکاوی در عامل‌های مصنوعی شده است [44].

۳-۱- متغییرهای هم‌جنس در هوش مصنوعی

در AI، الگوریتم‌هایی که بر اساس متغیرهای هم‌سنجش، کنجکاوی را مدل می‌کنند، قابلیت یادگیری و تعمیم‌پذیری بهتری را نشان داده‌اند، زیرا این سیستم‌ها برای کاوش و یادگیری مؤثر از محیط خود هدایت می‌شوند. این ادغام متغیرهای هم‌سنجش در سیستم‌های یادگیری، نقش اساسی آن‌ها را در پرورش کنجکاوی و بهبود فرآیندهای یادگیری انسان و ماشین برجسته می‌کند [43].

یکی از روش‌های کمی‌سازی کنجکاوی، اندازه‌گیری آن بر اساس یک یا چند متغیر هم‌سنجش با در نظر گرفتن محرک‌هایی است که از محیط دریافت می‌شود. به طور خاص، متغیرهای هم‌سنجش مجموعه‌ای از اطلاعات از منابع مختلف هستند که بر اساس دانش قبلی یا انتظارات در زمینه‌های مرتبط، شباهت‌ها و تفاوت‌ها را نسبت به محیط تشخیص می‌دهند [43,45]. در اینجا ما نگاهمان به این متغیرها و ترجمان آن‌ها در حوزه AI و مدل‌های محاسباتی مبتنی بر پژوهش‌های [11] و [8] که بر پایه تحقیقات برلین [13] است، می‌باشد. متغیرهای اصلی هم‌سنجش به شرح زیر تعریف می‌شوند:

- **تازگی:** به معنای درک محرک یا اطلاعات جدید (تجربیات) است. میزان تازگی یک محرک با سه عامل به طور معکوس مرتبط است: ۱) تعداد دفعات برخورد با محرک‌های مشابه در گذشته، ۲) زمان آخرین مواجهه با محرک، و ۳) میزان شباهت محرک به محرک‌های قبلی. به عبارت دیگر، هر چه محرک جدیدتر باشد و کمتر با آن روبرو

شده باشیم و از نظر ظاهری یا عملکردی با محرک‌های قبلی تفاوت بیشتری داشته باشد، از تازگی بیشتری برخوردار خواهد بود.

- **تغییر**[1]: هنگام تأثیر محرک روی گیرنده‌ها، حرکت را نشان می‌دهد. تغییر را می‌توان با اختلاف بین دو بازه‌ی زمانی اندازه‌گیری کرد.
- **شگفتی**: زمانی رخ می‌دهد که نتیجه حاصل شده از یک محرک متفاوت با انتظار ما باشد.
- **تناقض**[2]: اگر نتیجه‌ای به هیچ عنوان قرار نیست از یک محرک ایجاد شود ولی دقیقا آن نتیجه حاصل شود، تناقض به وجود آمده است.
- **پیچیدگی**: تقریباً به تنوع یا گوناگونی در یک الگوی محرک اشاره دارد. سه ویژگی اصلی که درجه‌ی پیچیدگی را تعیین می‌کنند عبارتند از: ۱) تعداد عناصر قابل تشخیص در یک محرک، ۲) عدم شباهت بین این عناصر، و ۳) درجه‌ای که چندین عنصر به عنوان یک واحد درک، و به آن‌ها پاسخ داده می‌شود.
- **عدم قطعیت**: زمانی بروز می‌کند که یک موجود زنده در انتخاب پاسخ به یک محرک دچار مشکل شود. درجه‌ی عدم قطعیت را می‌توان با استفاده از اشکال مختلف آنتروپی طبق نظریه اطلاعات، با تشخیص (طبقه‌بندی) محرک دریافتی یا پاسخ به دسته‌های مختلف، به صورت کمی اندازه‌گیری کرد.
- **تعارض**[3]: زمانی رخ می‌دهد که یک محرک دو یا چند پاسخ ناسازگار را در یک موجود زنده برانگیزد. پاسخ‌ها می‌توانند به چندین روش با یکدیگر ناسازگار باشند. اولاً، برخی پاسخ‌ها ذاتاً با یکدیگر متضاد هستند. به عنوان مثال، هیچ موجودی نمی‌تواند همزمان به جلو و عقب حرکت کند. پاسخ‌های دیگر ممکن است در ابتدا بتوانند با هم اجرا شوند، اما از طریق یادگیری ناسازگار شوند. به عنوان مثال، ما به ندرت هنگام دست دادن اخم می‌کنیم. دلیل سوم ناسازگاری به توانایی محدود موجود زنده در انجام چند وظیفه به طور همزمان، نسبت داده می‌شود. به عنوان مثال، توانایی استثنایی در نظر گرفته می‌شود اگر شخصی بتواند همزمان دو کتاب را بخواند.

لازم به ذکر است که تغییر، شگفتی و تناقض می‌توانند به عنوان یک متغیر تکمیلی تازگی در نظر گرفته شوند.

علاوه بر متغیرهای فوق، ما متغیر جدیدی به نام «پوشش فضا[4]» را معرفی کرده‌ایم. چرا که متغیرهای هم‌سنجش شامل تشخیص شباهت‌ها و تفاوت‌ها بین محرک فعلی و تجربیات قبلی هستند. پوشش فضا با این امر همخوانی دارد و نشان می‌دهد که عامل چقدر از محیط (فضای مسئله) را کاوش کرده است. به بیان دقیق‌تر برای تعریف آن می‌توان گفت: پوشش فضایی به میزان کاوش یک عامل در مناطق مختلف درون یک فضای مسئله تعریف می‌شود. این مفهوم تمایل عامل را برای بررسی و تعامل با مجموعه‌ای متنوع از مکان‌ها یا وضعیت‌ها در محیط به تصویر می‌کشد. برای

[1] Change
[2] Incongruity
[3] Conflict
[4] Space Coverage

اینکه یک متغیر بتواند جزء متغیرهای هم سنجش قرار گیرد باید ویژگی‌های زیر را داشته باشد:

- قابلیت اندازه‌گیری کمی: می‌توان پوشش فضایی را به عنوان درصدی از فضای مسئله که عامل بازدید یا با آن تعامل داشته است تعریف کرد.
- اختصاصی بودن: متغیرهای همبستگی کنونی مانند تازگی و پیچیدگی کلی‌تر هستند. ممکن است بخواهیم انواع خاصی از پوشش فضایی را بسته به فضای مسئله تعریف کنیم. به عنوان مثال، کاوش مناطق جدید در مقابل بازدید مجدد از مناطق قبلاً کاوش شده با اعمال مختلف.
- قابلیت ادغام با سایر متغیرهای موجود: ممکن است با سایر متغیرها ترکیب شود یا به عنوان یک عامل وزنی بسته به وظیفه خاص و رفتار کاوش مورد نظر استفاده شود.

۳-۲- مدل‌های محاسباتی در کنجکاوی

با توجه به حجم تحقیقات انجام شده، تقسیم‌بندی ما در هفت گروه: تازگی، تغییر، تعارض، پیچیدگی، عدم قطعیت، شگفتی و تناقض، پوشش فضا خواهد بود. این تقسیم‌بندی در شکل ۱ نشان داده شده است.

شکل ۱: انواع مدل‌های محاسباتی در کنجکاوی

۳-۲-۱- مدل‌های مبتنی بر تازگی

کنجکاوی مبتنی بر تازگی به‌عنوان یک انگیزش درونی کلیدی عمل می‌کند که عامل‌های هوشمند را به کاوش در حالات جدید و ناشناخته سوق می‌دهد. این نوع کنجکاوی بر این اصل استوار است که مواجهه با محرک‌های جدید می‌تواند پاداش‌های درونی ایجاد کند و حتی در محیط‌هایی با پاداش‌های بیرونی محدود، کاوش را تشویق می‌کند. به‌طور کلی، کنجکاوی مبتنی بر تازگی به‌عنوان یک متغیر هم‌سنجش در عامل‌های هوشمند عمل می‌کند و به‌طور معکوس با فرکانس مواجهه با محرک‌های مشابه همبستگی دارد؛ یعنی عامل‌ها تمایل دارند تا از تعامل مکرر با حالات

آشنا اجتناب کرده و به‌دنبال حالات جدید باشند. این امر منجر به یادگیری رفتارهای جدید و کسب مهارت‌های بیشتر می‌شود [51]. با وجود این، رویکرد صرفاً مبتنی بر تازگی ممکن است بهینه نباشد، زیرا در برخی موارد بازبینی حالات قبلی می‌تواند مفید باشد. تازگی را می‌توان بر اساس تراکم مواجهه با محرک‌های مشابه اندازه‌گیری کرد؛ شمارش دفعات بازدید از حالات یا جفت‌های حالت–عمل می‌تواند شاخصی از تازگی و کنجکاوی باشد. کاوش مبتنی بر تازگی، با ایجاد انگیزه برای کشف محرک‌های ناآشنا، درک جامعی از محیط فراهم می‌کند و به بهبود یادگیری و توانایی‌های انطباقی منجر می‌شود. مطالعات اخیر نشان داده‌اند که عامل‌هایی که از مکانیزم‌های هدایت‌شده با تازگی بهره می‌برند، عملکرد بهتری در وظایف مختلف دارند [46,47]. این یافته‌ها بر اهمیت گنجاندن این نوع کنجکاوی در طراحی سیستم‌های مستقل برای تقویت رفتارهای یادگیری مؤثر و سازگار تأکید می‌کند.

یکی از نخستین تحقیقات در زمینه‌ی مدل‌سازی کنجکاوی محاسباتی توسط ماکدو[1] و کاردوزو[2] [49] انجام شد. آنها مدلی برای ربات‌هایی که در محیط‌های ناشناخته کاوش می‌کنند، ارائه دادند که بر مبنای مفاهیمی همچون تازگی، شگفتی و عدم قطعیت عمل می‌کند. این مدل محیط را به صورت شبکه‌ای از اشیاء نمایش می‌دهد، که هر شیء به عنوان یک «محرک» کنجکاوی ربات در نظر گرفته می‌شود. برای سنجش تازگی هر شیء، ابتدا ویژگی‌های اشیاء در یک قالب مشترک قرار می‌گیرند و سپس به کدهای عددی تبدیل می‌شوند. تفاوت بین اشیاء از طریق محاسبه‌ی فاصله‌ی همینگ[3]، که تعداد عدم تطابق‌ها را اندازه‌گیری می‌کند، مشخص می‌شود. شیئی که کمترین شباهت به اشیاء قبلی داشته باشد، بالاترین تازگی و «ارزش کنجکاوی» را دارد و احتمال بیشتری برای کاوش توسط ربات خواهد داشت. این مدل از استراتژی «اضافه کردن تک به تک، بهترین در ابتدا[4]» استفاده می‌کند، به این معنی که ربات همیشه شیء با بالاترین تازگی را برای کاوش انتخاب می‌کند. آیا ماشین‌ها می‌توانند خلاق باشند؟ ساندرس[5] و گرو[6] [50] با طراحی مدلی از کنجکاوی برای عامل‌های طراحی، به بررسی این سؤال پرداخته‌اند. این مدل بر ارزیابی «جذابیت» الگوهای طراحی جدید بر اساس تجربیات گذشته‌ی عامل تمرکز دارد و به عامل کمک می‌کند تا تصمیم بگیرد کدام الگوها ارزش کاوش بیشتری دارند. فرآیند با مقایسه‌ی هر الگوی طراحی جدید با الگوهای موجود در «فضای مفهومی طراحی» آغاز می‌شود. این فضا، که توسط یک نقشه‌ی خودسازمان‌دهنده[7] مدل‌سازی شده، نمایانگر دانش عامل از الگوهای قبلی است. الگوهای

[1] Macedo
[2] Cardoso
[3] Hamming distance
[4] Add one at a time, Best first
[5] Saunders
[6] Gero
[7] Self-Organizing Map

جدید بر اساس میزان شباهت به طرح‌های موجود ارزیابی می‌شوند؛ هرچه شباهت بیشتر باشد، تازگی کمتر و در نتیجه جذابیت آن کاهش می‌یابد. این مدل با استفاده از ایجاد تعادلی بین تازگی و آشنایی مقدار کنجکاوی را محاسبه می‌کند. الگوهایی که نه بیش از حد آشنا و نه کاملاً بیگانه هستند، جذاب‌تر به نظر می‌رسند و مقدار کنجکاوی بالاتری دارند. این مقدار جذابیت عامل را به سمت کاوش ایده‌های جدیدتر و خلاقانه‌تر هدایت می‌کند. این مدل نشان می‌دهد که کنجکاوی، تحت تأثیر تازگی، می‌تواند نقشی کلیدی در تقویت خلاقیت در ماشین‌ها ایفا کند.

[48] یک معماری برای ربات‌های کنجکاو و بازپیکربندی شونده ارائه کرد که بازی و تفکر طراحی خلاقانه را تشویق می‌کند. این ربات‌ها در واکنش به تغییرات در ساختار خود، رفتارهای جدیدی را یاد می‌گیرند و این باعث تشویق به آزمایش، تأمل و تخیل می‌شود. در این تحقیق در مورد چگونگی اینکه این ربات‌ها می‌توانند از طریق یک مدل محاسباتی کنجکاوی و RL، رفتارهای جدیدی را یاد بگیرند، صحبت می‌شود. در بسیاری از تحقیقات، علاقه فزاینده‌ای به استفاده از روش‌های اکتشاف مبتنی بر شمارش برای بهبود عملکرد عامل تقویتی وجود دارد. بلمر[1] و همکاران [54] رویکردی مبتنی بر شمارش را با انگیزش درونی در RL ترکیب کرده‌ند. آن‌ها مفهومی جدید به نام « شمارش کاذب[2]» را معرفی می‌کنند که با استفاده از مدل‌های چگالی، اکتشاف مبتنی بر شمارش را به تنظیمات غیرجدولی تعمیم می‌دهد. این روش به بهبود اکتشاف در محیط‌های پیچیده‌ای مانند بازی‌های ویدیویی کمک می‌کند، جایی که روش‌های سنتی مبتنی بر شمارش با مشکل مواجه می‌شوند. این رویکرد امکان اکتشاف مبتنی بر تازگی را حتی در وظایف کنترل پیوسته، که در آن‌ها تعداد وقوع حالات بسیار کم یا صفر است، فراهم می‌کند. برای کاهش پیچیدگی مدل برای پیاده‌سازی اکتشاف مبتنی بر تازگی (تعداد) و در نتیجه قابلیت تعمیم روش‌های اکتشاف مبتنی بر شمارش به فضاهای حالت با ابعاد بالا و پیوسته، تانگ[3] و همکاران [55]، تابع هش[4] کردن رمزگذاری شده را برای نگاشت فضای حالت پیوسته با ابعاد بالا به یک فضای ویژگی با ابعاد پایین پیشنهاد کردند. البته آموزش یک کد هش مبتنی بر رمزگذار خودکار[5] نیازمند به‌روزرسانی تعداد زیادی پارامتر است که می‌تواند فرآیند یادگیری را کند نماید.

[57] مکانیزم جدیدی برای کنجکاوی در RL پیشنهاد می‌دهد که از حافظه اپیزودیک برای تعیین پاداش‌های نوآوری بر اساس قابلیت دسترسی استفاده می‌کند. به‌جای تکیه صرف بر شگفتی یا غیرقابل پیش‌بینی بودن، این رویکرد بر میزان تلاش لازم برای رسیدن به یک حالت جدید از حالات قبلی تمرکز دارد. هدف این مکانیزم، کاهش مشکل تمرکز بر رفتارهای تصادفی به جای اکتشاف معنادار است. تازگی در این رویکرد به‌صورت ضمنی و بر اساس قابلیت دسترسی

[1] Bellemare
[2] Pseudo-count
[3] Tang
[4] Hash
[5] autoencoder

به حالت‌های جدید سنجیده می‌شود. با استفاده از تعداد مراحل لازم برای رسیدن به حالت‌های جدید از طریق حافظه تجربه‌شده‌ی ذخیره‌شده در یک بافر حافظه، می‌توان درجه تازگی را کمّی‌سازی کرد. این تازگی مبتنی بر قابلیت دسترسی، پاداش‌های درونی را شکل می‌دهد که عامل را به کاوش در حالات کمتر آشنا سوق می‌دهد. کیم[1] و همکاران [58] برای ارزیابی تازگی مرتبط با وظیفه، از چارچوب گلوگاه اطلاعات واریانس [59] استفاده کردند. این روش به‌منظور فشرده‌سازی اطلاعات غیرضروری و حفظ داده‌های مرتبط با وظیفه، از شبکه‌های عصبی و پارامتری‌سازی مجدد بهره می‌برد. هدف این رویکرد، بهبود عملکرد مدل و افزایش مقاومت آن در برابر حملات تهاجمی است. علاوه بر تشویق اکتشاف، کنجکاوی همچنین باعث بهبود کارایی نمونه‌ها[2] می‌شود. با بازپخش مکرر تجربیات جدید، عامل‌ها می‌توانند با سرعت بیشتری سیاست‌های بهینه را یاد بگیرند. الگوریتم‌های RL مستقل از سیاست معمولاً از یک بافر بازپخش برای ذخیره مجموعه‌های گذار[3] استفاده می‌کنند که در طول آموزش مدل به‌طور یکنواخت نمونه‌برداری و در فرآیند یادگیری به کار گرفته می‌شوند.

برای تشویق عامل به تمرکز بر تجربیات کمتر کاوش شده (جدید)، [18] یک چارچوب اولویت‌دهی مبتنی بر کنجکاوی[4] پیشنهاد کرده است که مسیرهایی با حالات هدف نادر را بیش از حد نمونه‌برداری می‌کند. این روش بهبود کارایی یادگیری و عملکرد را هدف قرار داده است. البته باید اشاره کرد که در صورت عدم وجود پاداش بیرونی، این روش ممکن است قادر به یادگیری یک سیاست بهینه نباشد. بسیاری از نظریه‌های کلاسیک RL برای توضیح رفتار انسانی در محیط‌های پویا بدون پاداش‌های خارجی کافی نیستند. [60] استدلال می‌کند که هر دو عامل شگفتی و نوآوری نقش‌های متفاوتی در تصمیم‌گیری انسانی دارند، به‌طوری‌که نوآوری به اکتشاف کمک می‌کند و شگفتی نرخ‌های یادگیری را افزایش می‌دهد. در این مقاله، یک مدل RL هیبریدی معرفی می‌شود تا اثرات شگفتی، نوآوری و پاداش را بر رفتار انسانی و سیگنال‌های EEG[5] تفکیک کند. مدل از یک الگوی تصمیم‌گیری عمیق و دنباله‌دار با پاداش‌های خلوت و تغییرات ناگهانی محیط استفاده می‌کند. سپس آزمایش‌های رفتاری و ضبط‌های EEG تحلیل می‌شوند تا نشان دهند این عوامل چگونه بر اکتشاف، یادگیری و تصمیم‌گیری تأثیر می‌گذارند.

پژوهشگران در [61] به دنبال پاسخ به این سوال هستند که آیا انسان‌ها مانند برخی الگوریتم‌های یادگیری، به دلیل کنجکاوی و جستجوی نوآوری، از مسیر اصلی خود منحرف می‌شوند؟ برای پاسخ به این سوال، آزمایشی طراحی شده

[1] Kim
[2] Sample efficiency
[3] transition tuples
[4] curiosity- driven prioritization
[5] Electroencephalogram

است که در آن شرکت‌کنندگان باید در محیطی پیچیده به دنبال پاداش باشند. محققان با مقایسه رفتار انسان‌ها با پیش-بینی‌های مدل‌های RL، به این نتیجه رسیده‌اند که خوش‌بینی نسبت به پاداش‌های آینده، می‌تواند باعث افزایش حواس‌پرتی و جستجوی بی‌هدف شود. این یافته‌ها نشان می‌دهد که کنجکاوی، اگرچه نیروی محرکه‌ای برای یادگیری و اکتشاف است، اما می‌تواند منجر به اتخاذ تصمیمات غیر بهینه شود. [62] یک روش به نام کاوش هماهنگ چندعاملی [1] را توسعه داده است. این روش به عامل‌های RL چندعاملی غیرمتمرکز اجازه می‌دهد با به اشتراک‌گذاری اطلاعات نوآوری به صورت هماهنگ کاوش کنند. این روش به چالش‌های اندازه‌گیری نوآوری جهانی از مشاهدات محلی و هماهنگی کاوش بین عامل‌ها برای بهبود عملکرد در محیط‌های با پاداش کم پاسخ می‌دهد.

[52] رویکرد جدیدی در RL برای AI عمومی به نام OODA-RL[2] معرفی کرده است که می‌تواند به تغییرات جدید و غیرمنتظره در محیط‌های باز پاسخ دهد. با بهره‌گیری از حلقه مشاهده (جمع‌آوری داده‌ها از محیط)، جهت‌گیری (تحلیل و بروزرسانی دانش بر اساس داده‌های جدید)، تصمیم‌گیری (انتخاب یک عمل)، و عمل (انجام عمل) که از تصمیم‌گیری‌های نظامی الهام گرفته شده، به عامل‌ها اجازه می‌دهد تا بدون قوانین از پیش تعیین شده به موقعیت‌های جدید و حتی چالش‌برانگیز پاسخ دهند و به این ترتیب، عامل‌های RL بتوانند در محیط‌های آشنا و جدید به صورت پویا سازگار شوند. برای کمک به ربات‌ها در راستای کاوش وضعیت‌های جدید به طور کارامدتر، [53] با استفاده از پاداش‌های درونی مبتنی بر شناسایی نوآوری از طریق یک شبکه عصبی عمیق رمزگذار خودکار، کنترل ربات را بهبود می‌بخشد. برای نیل به این هدف، RL همراه با پاداش‌های درونی مبتنی بر فرکانس بازدید وضعیت‌ها، کارایی را افزایش می‌دهد. رمزگذار خودکار سیگنال‌های حسگری را پردازش می‌کند تا ناهنجاریها را شناسایی کرده و به عنوان پاداش‌های درونی استفاده می‌کند. در ادامه در جدول ۱ به مزایا و محدودیت‌ها (چالش‌ها) و معیار پاداش تحقیقات بیان شده در مدل‌های مبتنی بر تازگی پرداخته است.

جدول ۱: مزایا، محدودیت‌ها (چالش‌ها) و معیار پاداش مدل‌های مبتنی بر تازگی

مرجع	مزایا	معایب/چالش‌ها/محدویت‌ها	معیار پاداش
[50]	کاهش بار کاری طراح به دلیل تمرکز بر طراحی‌های جالب، یادگیری تطبیقی به علت استفاده از نقشه‌های خودسازمانده	محدودیت به فضاهای طراحی خاص، نیاز به اعتبارسنجی بیشتر	بر اساس سطح نوآوری تشخیص داده شده در طراحی‌ها است. طراحی‌هایی که کمتر معمول و نوآورتر هستند، جالب‌تر محسوب می‌شوند و

[1] Multi-Agent Coordinated Exploration (MACE)

[2] Observation-Orientation-Decision-Action-Reinforcement Learning

			بنابراین پاداش‌های درونی بالاتری دریافت می‌کنند.
[48]	تشویق به تفکر خلاقانه، یادگیری تطبیقی، فراهم کردن پلتفرمی برای یادگیری بازی‌محور مفاهیم الکترونیک و برنامه‌نویسی کامپیوتری.	امکان محدود بودن تعداد کاربر به علت نمایش اولیه به پلتفرم Lego Mindstorms، وابستگی اثربخشی سیستم به میزان مشارکت و تمایل کاربر به آزمایش	معیارهای پاداش بر اساس توانایی ربات در کشف رفتارهای جدید و جالب در پاسخ به تغییرات ساختاری است.
[54]	بهبود اکتشاف در محیط‌های پیچیده که روش‌های سنتی مبتنی بر شمارش بی‌اثر هستند، تعمیم به تنظیمات غیرجدولی، عملکرد مناسب در بازی‌های دشوار	پیچیدگی پیاده‌سازی مدل‌های چگالی برای شمارش‌های کاذب به ویژه در فضاهای با ابعاد بالا، ناتوان در یادگیری از پیکسل‌های خام بدون هزینه‌های سرسام‌آور برای مدل چگالی	حداکثرسازی پاداش تجمعی عامل از طریق بهبود اکتشاف. استفاده از شمارش‌های کاذب به هدایت اکتشاف به سمت حالت‌های با عدم قطعیت بالاتر کمک می‌کند.
[55]	رویکردی ساده برای اکتشاف مبتنی بر شمارش بدون نیاز به ترفندها یا مدل‌های پیچیده، انعطاف‌پذیری بالا چرا که این روش مکمل الگوریتم‌های موجود RL است و می‌تواند به راحتی در تنظیمات مختلف ادغام شود، نگاشت فضای حالت پیوسته با ابعاد بالا به یک فضای ویژگی با ابعاد پایین بوسیله تابع هش کدگذاری شده	وابستگی زیاد اثربخشی روش به طراحی تابع هش[1]، هزینه بر بودن مدیریت و به روزرسانی جداول هش به ویژه در فضاهای حالت بسیار بزرگ، چالش در تعمیم به محیط‌های بسیار پویا یا تغییر سریع به علت عدم سازگاری مناسب تابع هش	بر اساس پاداش‌های اکتشافی ناشی از شمارش حالات است. به دنبال ارائه پاداش‌های اضافی برای حالات کمتر بازدید شده است.
[57]	عملکرد بهتر از روش‌های پیشرفته در معیارهای مختلف از جمله محیط‌های سه‌بعدی پیچیده و وظایف تولیدی تصادفی، عملکرد موثر در محیط‌های بصری غنی مانند VizDoom، DMLab و MuJoCo، حل مشکل مبل‌نشین (تلویزیون نویزی)	چالش مدیریت حافظه اپیزودیک به طور مؤثر و اطمینان از اینکه بسیار بزرگ یا دست و پاگیر نشود، چالش تعیین آستانه مناسب برای اینکه چه چیزی به عنوان یک حالت نوآورانه در نظر گرفته می‌شود، امکان عملکرد ضعیف در وظایف با مشاهده جزئی یا تنظیمات با نمونه‌های کم	پاداش بر اساس نوآوری مشاهده، که با تلاش مورد نیاز برای رسیدن به آن از مشاهدات قبلاً مواجه شده تعیین می‌شود. اگر تلاش بیشتر از آستانه معین باشد، عامل پاداش نوآوری دریافت می‌کند که به پاداش واقعی وظیفه اضافه می‌شود.
[18]	بهبود عملکرد در وظایف دستکاری رباتیک، یادگیری متعادل‌تر و عملکرد کلی بهبود یافته‌تر عامل به خاطر رسیدگی به مشکلات عدم تعادل حافظه و تعصب نمونه، قابلیت	بار محاسباتی اضافی به دلیل تخمین چگالی و اولویت‌دهی به مسیرها، وابستگی شدید اثربخشی روش به تخمین	به طور ضمنی به تعادل تجربیات در بافر حافظه مرتبط هستند. چارچوب پیشنهادی معیارهای پاداش صریحی را تعریف نمی‌کند، بلکه بر روی اولویت‌دهی و نمونه‌برداری بیش از حد از

[1] Hash Function

	ادغام با الگوریتم‌های مختلف RL مستقل از سیاست، بهبود کارایی نمونه از طریق بازپخش تجربه‌ی اولویت‌دار	چگالی، ناتوان در یادگیری کارآمد از پیکسل‌های خام	مسیرهایی که وضعیت‌های هدف نادر دارند، تمرکز دارد.
[62]	به هزینه ارتباطی کمی نیاز دارد که آن را برای تنظیمات غیرمتمرکز عملی می‌سازد، می‌توان در محیط‌های مختلف چندعاملی اعمال کرد که نشان‌دهنده تنوع و قابلیت گسترش آن است.	به یک شبکه ارتباطی کاملاً متصل برای به اشتراک‌گذاری نوآوری نیاز دارد که ممکن است در همه سناریوها عملی نباشد، چالش در سیستم‌های بسیار بزرگ مقیاس به دلیل نیاز به اشتراک‌گذاری نوآوری بین عوامل متعدد	ترکیبی از پاداش مبتنی بر نوآوری: برای کاوش حالت‌ها و مشاهدات جدید داده می‌شود و پاداش‌های مبتنی بر گذشته‌نگر[1]: بر اساس اطلاعات متقابل وزنی که تأثیر اعمال یک عامل بر نوآوری به دست آمده توسط دیگران را اندازه‌گیری می‌کند، فراهم می‌شود.
[61]	ارائه بینش‌هایی در مورد چگونگی تأثیر کنجکاوی و خوش‌بینی بر کاوش و تصمیم-گیری انسانی، افشای الگوهای حواس‌پرتی که نشان می‌دهد چگونه حواس‌پرتی مبتنی بر نوآوری بر کاوش تأثیر می‌گذارد.	محدودیت‌های تنظیمات تجربی، چالش برانگیز بودن اندازه‌گیری و تعریف «خوش‌بینی به پاداش»	پاداش‌های درونی مبتنی بر نوآوری، شگفتی و کسب اطلاعات است. پاداش‌های مبتنی بر نوآوری بر کاوش حالت‌های جدید و شگفت‌انگیز تمرکز دارد، در حالی که پاداش‌های مبتنی بر کسب اطلاعات به جمع‌آوری اطلاعات مفید مربوط می‌شود.
[60]	ارائه پاداش بهبود یافته به خاطر درک دقیق-تری از رفتار انسانی با تفکیک بین شگفتی، نوآوری و پاداش، مدل‌سازی واقعی‌تر رفتار تصمیم‌گیری انسانی در محیط‌های پویا با پاداش‌های پراکنده[2] و تغییرات ناگهانی، امکان ارائه بینش‌های عمیق‌تری از پایه‌های عصبی به علت توانایی تفکیک شگفتی، نوآوری و پاداش در سیگنال‌های EEG	اضافه کردن پیچیدگی به خاطر مدل هیبریدی RL و نیاز به تحلیل دقیق EEG، چالش برانگیز بودن تفسیر و تفکیک اثرات شگفتی، نوآوری و پاداش در هر دو رفتار و سیگنال‌های EEG و نیاز به تکنیک‌های تحلیل پیشرفته	ارزیابی توانایی پیش‌بینی تصمیمات انسانی و تغییر نرخ‌های یادگیری در پاسخ به این عوامل، و همچنین توانایی تفکیک این اثرات در سیگنال‌های EEG
[52]	سازگاری بهبود یافته با تغییرات غیرمنتظره در محیط‌های باز، امکان تنظیمات در زمان واقعی و کاهش وابستگی به قوانین از پیش تعریف شده با استفاده از یادگیری فعال	پیچیدگی در پیاده‌سازی، وابستگی به داده‌های با کیفیت و چالش برانگیز بودن اندازه‌گیری موفقیت در محیط باز به دلیل تنوع و تازگی سناریوها	بر اساس موفقیت عامل در سازگاری با موقعیت-های جدید و دستیابی به اهداف وظیفه
[53]	بهبود کارایی کاوش و کاهش مشکل پاداش-های خلوت	استفاده از رمزگذار خودکار ممکن است پیچیدگی‌هایی را از نظر آموزش و ادغام به همراه داشته باشد. کاهش کارایی پاداش‌های درونی در صورت عدم شناسایی نوآوری به صورت بهینه	شامل پاداش‌های درونی مبتنی بر فرکانس بازدید از وضعیت‌های مختلف در محیط

[1] Hindsight-based
[2] Sparse

۳-۲-۲- مدل‌های مبتی بر تغییر

کنجکاوی می‌تواند با تغییر محرک برانگیخته شود. در RL، این متغیر هم‌سنجش به تغییر قابل توجهی در محیط پس از انجام یک عمل خاص توسط عامل اشاره دارد. این مفهوم که با عنوان «کجکاوی مبتنی بر تغییر» شناخته می‌شود، بر میزان تغییر محیط پس از انجام یک عمل توسط عامل تمرکز دارد [63]. فرض کنید محرکی مانند یک تصویر یا صدا بر حسگرهای عامل تأثیر می‌گذارد. تغییر در محیط را می‌توان با مقایسه‌ی وضعیت (مجموعه‌ای از تمام ادراکات عامل) در دو لحظه‌ی مختلف اندازه‌گیری کرد. به طور شهودی، هنگامی که اعمال عامل منجر به تغییرات قابل توجهی در محیط شود، کنجکاوی آن‌ها افزایش می‌یابد و در نتیجه مهارت‌های تأثیرگذاری را برای کنترل محیط یاد می‌گیرند [8]. نکته‌ی مهم این است که این تغییر، در کنار تازگی، می‌تواند به عنوان یک عامل ثانویه در تعیین میزان «جذابیت» یک موقعیت برای عامل در نظر گرفته شود [63]. در اینجا، تغییر محیط را می‌توان بر اساس ویژگی‌های مختلفی مانند: ۱) تغییر در وضعیت ۲) ارائه اطلاعات جدید (مانند کشف یک ورودی یا خروجی جدید) ۳) ایجاد پیامدهای غیرمنتظره (پاداش‌های بالا یا پایین غیرمنتظره) سنجید.

در زمینه یادگیری حسی–حرکتی ربات، مدلی برای هدایت توجه ربات به تغییرات محیط ارائه شده است [64]. این روش به سیستم این امکان را می‌دهد که بر روی نواحی جالب‌تر در فضای یادگیری تمرکز کند؛ این نواحی معمولاً شامل مواردی هستند که در آن‌ها دقت پیش‌بینی می‌تواند بهبود یابد. ربات به سمت مناطقی هدایت می‌شود که در آن‌ها پیش‌بینی‌ها دقیق نیستند یا تغییرات جدیدی رخ داده است، در حالی که از نواحی بدون قابلیت بهبود دوری می‌کند. هدف این مدل اندازه‌گیری میزان تغییرات در یک محرک نیست، بلکه بر چگونگی واکنش به این تغییرات تمرکز دارد. به همین دلیل، سیستم به بررسی سطح تحریک یا کنجکاوی ناشی از تغییرات نمی‌پردازد، بلکه مکانیزم‌هایی برای هدایت مجدد توجه به تغییرات را ارائه می‌دهد. تمرکز توجه ربات از طریق توزیع گوسی[1] تعیین می‌شود که نمونه‌برداری از داده‌های آموزشی را در فضای موتور لیزر کنترل می‌کند. مرکز این توزیع به‌وسیله میانگین N نمونه آخر به‌دست آمده و توجه ربات را به مناطقی که به‌طور اخیر به‌روزرسانی شده‌اند، هدایت می‌کند. این مدل به‌طور موثری می‌تواند توجه ربات را به سمت تغییرات معطوف کند. [65] یک روش جدید پاداش درونی به نام «اکتشاف مبتنی بر تأثیر پاداش‌دهی[2]» را برای RL در محیط‌های تولید شده به صورت رویه‌ای پیشنهاد می‌کند. این روش، عامل را با پاداش دادن به اعمالی که منجر به تغییرات قابل توجهی در نمایش حالت‌های آموخته شده آن‌ها می‌شود، به اکتشاف تشویق می‌کند. این تحقیق

[1] gaussian distribution
[2] Rewarding Impact-Driven Explration (RIDE)

از [46] برای یادگیری یک نمایش حالت معنی‌دار و یک پاداش درونی مبتنی بر تغییر استفاده می‌کند. سیاست اکتشافی [65] فاقد مؤلفه انتقال است و همچنین نیاز به یادگیری مدل‌ها دارد.

برای رفع این مشکل، [66] ایده مشابهی ارائه کردند که عامل اعمالی را انجام دهد که تغییرات جالبی در محیط ایجاد می‌کنند. ایده‌ی این مقاله پیشنهاد یک روش جدید برای اکتشاف بدون وابستگی به وظیفه در RL است، جایی که عامل بدون هدف خاص یا اهداف خارجی در محیط‌های متعدد اکتشاف می‌کند و سپس سیاست‌های اکتشاف آموخته شده را به محیط‌های جدید و دیده نشده منتقل می‌کند. برای دستیابی به این هدف، روشی به نام انتقال اکتشاف مبتنی بر تغییر معرفی[۱] شده است. این روش شامل دو مرحله است: یادگیری سیاست‌های اکتشاف از یک یا چند محیط بدون وظایف خاص و انتقال این سیاست‌های آموخته شده به محیط‌های جدید. این روش پاداش‌های درونی‌ای که عامل را به اکتشاف بخش‌های دیده نشده محیط تشویق می‌کنند با پاداش‌هایی که تعامل با اشیاء جالب ذاتی را تشویق می‌کنند، ترکیب می‌کند. سپس، سیاست‌های اکتشافی آموخته شده به محیط‌های جدید منتقل می‌شوند. یوآن[۲] و همکاران [67] روش نوینی به نام «پاداش‌دهی به اختلاف بازدید اپیزودیک[۳]» برای بهبود اکتشاف در RL ارائه کردند. با استفاده از یک چارچوب ساده و کارآمد از نظر محاسباتی، پاداش درونی را بر اساس اساس اختلاف در بازدید از حالت‌ها بین اپیزودها محاسبه می‌کند. از یک برآوردگر k–نزدیک‌ترین همسایه با یک کدگذار حالت تصادفی اولیه برای برآورد این تفاوت به طور کارآمد استفاده می‌کند.

[68] روشی نوین برای بهبود اکتشاف در RL با استفاده از معیار شباهت شبیه‌سازی دوتایی[۴] معکوس پویا برای اندازه‌گیری تفاوت حالت‌ها ارائه می‌دهد که باعث می‌شود پاداش‌های متراکم‌تری را بدون تنظیم دستی فراهم کند. عامل، وضعیت‌های دارای خطای TD[۵] بالاتر را کاوش می‌کند، در نتیجه به طور قابل توجهی کارایی آموزش را بهبود می‌بخشد. این روش یک پاداش اکتشافی ایجاد می‌کند که عامل RL را به اکتشاف حالت‌های جدید بدون وابستگی به دانش قبلی انسانی تشویق می‌کند. هدف [72]، بررسی این است که کدام ویژگی‌های روش‌های اکتشافی در RL برای انتقال کارآمد وظایف در مواجهه با تغییرات محیطی در محیط‌های غیر ایستا مفیدتر هستند. این مقاله با دسته‌بندی و ارزیابی یازده الگوریتم اکتشافی RL بر اساس ویژگی‌هایی مانند تنوع صریح و تصادفی بودن، در حوزه‌های گسسته و پیوسته کار می‌کند. این ویژگی‌ها از نظر کارایی آن‌ها در سازگاری با پنج نوع مختلف از نوآوری‌های محیطی آزمایش می‌شوند تا

[۱] Change-Based Exploration Transfer (C-BET)
[2] Yuan
[۳] Rewarding Episodic Visitation Discrepancy (REVD)
[4] bisimulation
[۵] Temporal Difference

مشخص شود که کدام ویژگی‌ها بهترین عملکرد در انتقال را ارائه می‌دهند. جدول ۲ به مزایا و محدودیت‌ها (چالش‌ها) و معیار پاداش تحقیقات بیان شده در مدل‌های مبتنی بر تغییر پرداخته است.

جدول ۲: مزایا، محدودیت‌ها (چالش‌ها) و معیار پاداش مدل‌های مبتنی بر تغییر

مرجع	مزایا	معایب/چالش‌ها/محدویت‌ها	معیار پاداش
[64]	تمرکز بر مناطق تغییر یافته برای سازگاری سریع-تر، جلوگیری از گیر افتادن در مناطقی که یادگیری در آن‌ها غیرممکن است، توانایی سازگاری با تغییرات محیطی بدون نیاز به کالیبراسیون مجدد، بهبود کارایی یادگیری با تمرکز بر مناطق جالب	نیاز به تنظیم دستی پارامترها، پیچیدگی بالقوه در گسترش به فضاهای چند بعدی	سیستم برای اکتشاف مناطقی که پیشرفت یادگیری را به حداکثر می‌رسانند پاداش دریافت می‌کند. این شامل تمرکز بر مناطقی است که تغییر کرده‌اند و نیاز به یادگیری جدید دارند.
[65]	اکتشاف کارآمدتر در محیط‌های تولید شده به صورت رویه‌ای، عملکرد بهتر به خصوص در محیط‌هایی که احتمال بازدید مجدد از حالت‌ها کم است، تشویق به انجام اقداماتی با تأثیرات قابل توجه؛ از اکتشاف مکرر همان حالت‌ها اجتناب کند.	احتمال ناکارآمدی در محیط‌هایی که تغییرات تاثیرگذار نادر یا دشوار به دست می‌آیند، عدم تعمیم‌پذیری برای همه انواع وظایف: مناسب برای وظایفی که در آن عامل نیاز به تأثیرگذاری بر محیط دارد نه وظایف نیازمند به حفظ پایداری	فاصله اقلیدسی بین نمایش حالت‌های متوالی که توسط تعداد بازدیدهای حالت اپیزودیک کاهش می‌یابد تا از بازدید مکرر عامل از همان حالت‌ها جلوگیری شود.
[66]	امکان انتقال مؤثر سیاست‌های اکتشاف به محیط-های جدید و افزایش قابلیت انطباق و اکتشاف مؤثر عامل، اکتشاف جامع‌تر با ترکیب پاداش‌های عامل محور و محیط محور، شبیه‌سازی بهتر از اکتشاف انسانی با در نظر گرفتن دانش و تجربیات قبلی	ناهماهنگی اکتشاف و اهداف وظیفه، مناسب بودن برای فضاهای گسسته و چالش تعمیم به فضاهای پیوسته	شامل پاداش‌های عامل–محور: برای اکتشاف مناطق جدید یا دیده نشده داده و پاداش‌های محیط–محور: برای تعامل با اشیاء یا تغییراتی که ذاتاً جالب تلقی می-شوند.
[67]	ارائه پاداش‌های اکتشافی بدون مشکل از بین رفتن پاداش‌های درونی، بهبود قابل توجهی کارایی نمونه‌گیری الگوریتم‌های RL، سادگی مدل و کارآمد و پایدار از نظر محاسباتی	عدم عملکرد مطلوب در وظایف نیازمند به اکتشاف طولانی‌مدت، تأثیر بر عملکرد اولیه به علت استفاده از یک کدگذار حالت تصادفی	تفاوت بازدید اپیزودیک، که با واگرایی رنی[1] بین اپیزودها اندازه‌گیری می‌شود.
[68]	مقیاس‌پذیرتر بودن در محیط‌هایی با تعداد زیادی حالت منحصربه‌فرد، ارائه پاداش‌های متراکم‌تر و افزایش سرعت آموزش، افزایش اکتشاف	پیچیدگی پیاده‌سازی تابع پتانسیل مبتنی بر معیار شبیه‌سازی دوتایی از نظر محاسباتی، حساسیت به کیفیت و کمیت داده‌های آموزشی موجود	بر اساس نو بودن حالت‌ها، که با معیار شبیه‌سازی دوتایی معکوس اندازه‌گیری می‌شود.

[1] Re´nyi

[72]	تطبیق بهتر با موقعیت‌های جدید، کاربرد گسترده-تر و قابلیت انتقال وظایف بهینه شده	محدودیت در انواع نوآوریهای آزمایش شده	توانایی عامل در سازگاری کارآمد با موقعیت‌های جدی

۳-۲-۳- مدل‌های مبتنی بر تضاد

تضاد زمانی به وجود می‌آید که یک محرک تمایل داشته باشد تا در یک موجود زنده، دو یا چند پاسخ ناسازگار را برانگیخته کند [11]. همان‌طور که در حال یادگیری موضوع جدیدی هستیم، ممکن است با اطلاعاتی مواجه شویم که با درک فعلی ما در تضاد باشد. این امر وضعیتی به نام «تضاد» را ایجاد می‌کند. حالتی که در آن از صحت اطلاعات مطمئن نیستیم.

یکی از اولین تحقیقات در زمینه کنجکاوی از منظر تضاد، توسط وو[1] و همکاران [69] انجام شد. آن‌ها مدل محاسباتی جامعی از کنجکاوی برای یادگیری در محیط‌های مجازی ارائه کردند که در آن تضاد به عنوان محرک کنجکاوی شناخته می‌شود. تضاد زمانی ایجاد می‌شود که درک یادگیرنده (دانش ذخیره شده) با دانش تخصصی موجود در دنیای مجازی در تضاد باشد؛ این امر باعث تحریک کنجکاوی و ترغیب یادگیرندگان به کاوش بیشتر می‌شود. برای اندازه‌گیری تعارض، رابطه بین مفاهیم در دانش کاربر و دانش جهانی مقایسه می‌شود. اگر رابطه میان مفهوم C_i و C_j در دانش کاربر با رابطه مشابه در دانش جهانی متفاوت باشد، درک کاربر در تضاد با درک متخصص خواهد بود. در مرحله اول، سطح تعارض در یک شی مجازی بر اساس تعداد روابط متضاد تعیین می‌شود و در مرحله دوم، سطح کنجکاوی به‌صورت مثبت با این تعارض همبستگی دارد. این همبستگی توسط وزن در رویکرد مبتنی بر نقشه شناختی فازی مشخص می‌شود. در پژوهش [70]، با هدف ارتقاء سیستم‌های توصیه‌گر، رویکردی نوین مبتنی بر ترکیب ترجیحات کاربر و کنجکاوی اجتماعی ارائه می‌شود. برخلاف روش‌های سنتی که صرفاً بر پایه ترجیحات کاربر عمل می‌کنند، مدل پیشنهادی با در نظر گرفتن عوامل روانشناختی همچون شگفتی، عدم قطعیت و تضاد، به دنبال تحریک کنجکاوی کاربر و ارائه توصیه‌های جذاب‌تر است. مدل پیشنهادی با تلفیق فیلترسازی مشارکتی برای استخراج ترجیحات کاربر و منطق فازی برای محاسبه عوامل روانشناختی، رتبه‌بندی شخصی‌سازی شده‌ای از آیتم‌ها ارائه می‌دهد. در این مدل، عوامل رواشناختی با استفاده از یک مکانیزم وزن‌دهی ترکیبی با ترجیحات کاربر ترکیب می‌شوند تا تعادلی بین شخصی‌سازی و کشف آیتم‌های جدید برقرار شود.

[1] Wu

[80] نشان می‌دهد که نظریه‌های مختلف کنجکاوی، با وجود تفاوت‌های ظاهری، به یک هدف مشترک یعنی حداکثر کردن دانش اشاره دارند و به این نتیجه می‌رسند که کنجکاوی به عنوان یک نیروی محرکه، ما را به سمت اطلاعات جدید و مفید سوق می‌دهد. با این حال، این نظریه‌ها به تنهایی نمی‌توانند تمام جنبه‌های کنجکاوی را توضیح دهند. برای درک کامل کنجکاوی، باید به تعامل بین عوامل مختلفی مانند تضاد، عدم قطعیت، تازگی، و انتظار پاداش توجه کنیم. در نتیجه چارچوبی منسجم برای ادغام آن‌ها ارائه می‌دهد. با توصیف فرآیند برانگیختگی تعارض توضیح می‌دهد که در آن تحریک خارجی مجموع‌های از پاسخ‌های رفتاری ناسازگار مانند پاسخ‌های حرکتی، پاسخ‌های ادراکی، پاسخ‌های معرفتی و پیشبینی‌ها را آغاز می‌کند. نظریه اطلاعات، همان‌طور که در اینجا استفاده می‌شود، برای موقعیت‌هایی اعمال می‌شود که پاسخ‌های فعال شده متقابلاً ناسازگار هستند و در نتیجه متناقض هستند. فرض بر این است که حالت‌های تعارض از نظر بیولوژیکی مهم هستند و ارگانیسم را به دنبال اطلاعات می‌اندازد. جدول ۳، به طور خلاصه مزایا، محدودیت‌ها و اهداف پاداش را در پژوهش‌های مختلف در حوزه مدل‌های مبتنی بر تغییر ارائه می‌دهد.

جدول ۳: مزایا، محدودیت‌ها (چالش‌ها) و معیار پاداش مدل‌های مبتنی بر تضاد

مرجع	مزایا	معایب/چالش‌ها/محدویت‌ها	معیار پاداش
[69]	توانایی شناسایی اشیاء با قابلیت یادگیری جالب متناسب با سطح دانش فعلی کاربر، کمک به کاربران برای اکتشاف بیشتر در محیط یادگیری مجازی	نیاز به مطالعات میدانی بیشتر برای تأیید اثربخشی مدل، عدم وجود روش یادگیری وزن برای سازگاری خودکار مدل، محدودیت در مقیاس‌پذیری و قابلیت تعمیم به سایر حوزه‌های یادگیری	معیار پاداش مستقیما ذکر نشده است، اما شدت کنجکاوی از طریق فرایند استنتاج عددیِ نقشه شناختی فازی[1] محاسبه می‌شود.
[70]	دقت توصیه بهبود یافته، توصیه طیف وسیع-تری از آیتم‌ها از جمله آیتم‌های انتهایی، پیشنهاد مجموعه متنوع‌تری از آیتم‌ها و جلوگیری از تکراری بودن توصیه‌ها	تخمین نادرست کنجکاوی تحت تأثیر کمیابی داده‌ها، چالش تخمین دقیق کنجکاوی با کاربران جدید با داده کم در دسترس، افزایش پیچیدگی محاسباتی برای ادغام چندین عامل کنجکاوی و محاسبه آن‌ها با استفاده از منطق فازی	ترکیبی از ترجیحات کاربر و کنجکاوی کاربر. آیتم‌ها با ارزیابی هر دو امتیاز پیش‌بینی شده (ترجیحات کاربر) و امتیازات کنجکاوی مشتق شده از عواملی مانند شگفتی، عدم قطعیت و تضاد رتبه‌بندی می‌شوند.
[80]	ارائه‌ی چارچوبی منسجم برای کنجکاوی با بهره‌گیری از چندین نظریه اصلی،	فقدان بررسی عملکرد چارچوب پیشنهادی در محیط‌های مختلف به ورت کاربردی	ترکیبی از نظریه‌های مختلف در راستای کسب دانش بیشتر (کسب دانش بیشتر منجر به پاداش بیشتر)

[1] Fuzzy Cgnitive Map

۳-۲-۴- مدل‌های مبتنی بر پیچیدگی

در مدل‌های یادگیری ماشین مرتبط با کنجکاوی، پیچیدگی معمولاً به دو حوزه زیر مربوط می‌شود: ۱) پیش‌بینی دقیق حالات آینده: این توانایی به چگونگی پیش‌بینی رخدادهای آینده بر اساس تجربیات گذشته سیستم اشاره دارد. ۲) فشرده‌سازی کارآمد اطلاعات: این بخش بر قابلیت سیستم در نمایش اطلاعات پیچیده به شکلی ساده‌تر و قابل مدیریت‌تر تأکید می‌کند. این دو حوزه به عنوان شاخص‌های «جذابیت» یا «پیچیدگی» یک موقعیت تلقی می‌شوند و می‌توانند کنجکاوی سیستم را برای کاوش بیشتر تحریک کنند. در محیط‌های مبتنی بر الگوریتم که وظایف از طریق پارامترها قابل تنظیم هستند، اجرای مستقیم مهارت‌های متنوع به دلیل محدودیت‌های پاداش‌های خارجی غیرعملی است [8]. با این حال، می‌توان با دو رویکرد، کنجکاوی را در چنین محیط‌هایی حفظ کرد: تمرکز بر پیچیدگی بهینه: به‌جای اجرای مستقیم مهارت‌ها، عامل‌ها می‌توانند پیچیدگی اهداف را شناسایی کرده و به‌دنبال وظایفی با سطح بهینه پیچیدگی باشند. حفظ کنجکاوی با چالش: چالش‌ها می‌توانند کنجکاوی را تحریک و حفظ کنند. با جستجوی وظایف با پیچیدگی بهینه، عامل‌ها می‌توانند سطح بالایی از کنجکاوی را حفظ کرده و انگیزه یادگیری خود را افزایش دهند.

اشمیدهوبر[1] [21] سیستم کنترل مدل‌سازی جدیدی معرفی کرد که از «کنجکاوی تطبیقی» برای بهبود پیش‌بینی‌های مدل جهان خود استفاده می‌کند. این سیستم به طور فعال به دنبال شرایطی است که انتظار دارد درباره محیط بیشتر یاد بگیرد. و از طریق یک رویکرد RL الگوریتم یادگیری Q[2]، قابلیت اطمینان پیش‌بینی خود را افزایش می‌دهد. سیستم از این بهبود پیش‌بینی برای تعیین پیچیدگی داده‌های ورودی استفاده می‌کند. این پیچیدگی می‌تواند از اطلاعات آشنا (و به راحتی قابل یادگیری) تا داده‌های نویزدار و غیرقابل پیش‌بینی که یادگیری آن‌ها دشوار است، متغیر باشد. این رویکرد هنگام انتخاب موارد کاوش، هر دو مفهوم «مناطق با قابلیت بهبود» و «مناطق نزدیک به تسلط» را در نظر می‌گیرد. [71] نیز مکانیزمی را برای «کنجکاوی تطبیقی هوشمند» معرفی می‌کند که ربات را به سمت موقعیت‌هایی سوق می‌دهد که در آن‌ها پیشرفت یادگیری‌اش به حداکثر می‌رسد. این مکانیزم باعث می‌شود ربات روی موقعیت‌هایی تمرکز کند که نه چندان قابل پیش‌بینی و نه کاملا غیرقابل پیش‌بینی باشند. نشان داده می‌شود که ربات ابتدا زمان خود را صرف موقعیت-های یادگیری آسان می‌کند و سپس به تدریج توجه خود را به سمت موقعیت‌های دشوارتر معطوف می‌کند و از موقعیت‌هایی که در آن‌ها چیزی برای یادگیری وجود ندارد، اجتناب می‌کند ایده [74] بررسی این است که چگونه رقابت چند عاملی، به طور خاص در بازی قایم‌باشک[3] و اهداف ساده RL می‌توانند منجر به ظهور رفتارهای پیچیده و

[1] Schmidhuber
[2] Q-Learning
[3] hide-and- seek

مرتبط با انسان مانند استفاده از ابزار و سازگاری استراتژیک شوند. مطالعه نشان می‌دهد که از طریق بازی، عامل‌ها می‌توانند استراتژی‌های پیچیده و استفاده از ابزار را توسعه دهند که توسط یک برنامه درسی خودسرپرست (چارچوب یادگیری است که در آن عامل‌ها مهارت‌های خود را از طریق پویایی تعاملات خود در یک محیط توسعه می‌دهند، به ویژه با تأکید بر ظهور استراتژی‌ها در طول زمان) هدایت می‌شود [75] چارچوب نوین RL را ارائه می‌کند که از اهداف ذاتی با انگیزه متخاصم برای آموزش عامل‌ها در محیط‌هایی با پاداش‌های بیرونی پراکنده استفاده می‌کند. این چارچوب شامل یک معلم هدف‌گذار و یک سیاست دانش‌آموز مبتنی بر هدف است. معلم به تدریج اهداف چالش‌برانگیزتری را برای دانش‌آموز تعیین کرده و او را تشویق به یادگیری مهارت‌های عمومی برای عمل در محیط‌های جدید می‌کند. این رویکرد یک برنامه درسی طبیعی از اهداف خودپیشنهادی ایجاد می‌کند که توانایی عامل را برای حل وظایف پیچیده در محیط‌های تولیدی رویه‌ای بهبود می‌بخشد. در نتیجه، با شناسایی پیچیدگی اهداف و تعیین سطح بهینه آن، می‌توان عامل‌ها را به صورت مداوم برای یادگیری با کنجکاوی بالا ترغیب کرد.

ایجاد روشی برای تکامل محیط‌های آموزشی برای عامل‌های RL در [76] مورد بررسی قرار گرفته است. روش پیشنهادی، که «پیچیدگی ترکیبی با ویرایش سطوح[1]» نام دارد، اهداف مبتنی بر پشیمانی را با فرآیندهای تکاملی ترکیب می‌کند تا محیط‌هایی تولید کند که ساده شروع می‌شوند اما به تدریج پیچیده‌تر می‌شوند. این روش از رویکرد مبتنی بر پشیمانی استفاده می‌کند تا اطمینان حاصل کند که عامل همیشه در لبه توانایی‌هایش آزموده می‌شود. سطوح توسط یک عامل معلم طراحی می‌شوند و توسط یک عامل دانش‌آموز حل می‌شوند، و با بهبود عامل دانش‌آموز، پیچیدگی سطوح افزایش می‌یابد. تفاوت مهمی که [77] با کارهای پیشین در این بخش دارد، این است که فاکتور جدیدی به نام بازخورد انسانی را اضافه کرده است. این چارچوب RL سلسله‌مراتبی بازخورد انسانی و مکانیزم محدودیت فاصله پویا را ادغام کرده است. هدف این چارچوب این است که آموزش عامل‌ها را با راهنمایی انتخاب زیرهدف‌ها با بازخورد انسانی و تنظیم پویا دشواری زیرهدف‌ها برای مطابقت با پیشرفت یادگیری عامل، بهبود بخشد.

ایده اصلی [82]، توسعه روشی برای دستکاری رباتیک است که به ربات‌ها اجازه می‌دهد به طور خودمختار مجموعه‌ای از مهارت‌های قابل استفاده مجدد را کشف و یاد بگیرند. این مهارت‌ها به ربات کمک می‌کند تا به روش‌های متنوع و پیچیده با اشیاء تعامل داشته باشد و وظایف مختلف دستکاری را انجام دهد. برای ایجاد مجموعه‌ای از وظایف پیچیده به صورت خودکار استفاده از بازی نامتقارن خودآموز[2] را پیشنهاد می‌دهد. همچنین در اینجا، پیچیدگی به عنوان نوعی کنجکاوی در نظر گرفته می‌شود، جایی که سیستم به طور خودکار وظایف پیچیده‌تری را ایجاد می‌کند. این

[1] Adversarially Compounding Complexity by Editing Levels (ACCEL)
[2] Asymmetric Self-Play

پیچیدگی فرآیند یادگیری را تحریک می‌کند، زیرا ربات به طور مداوم سعی در حل وظایف دشوارتر دارد و در نتیجه مهارت‌ها و قابلیت‌های متنوعی در دستکاری کسب می‌کند. بررسی مزایا، چالش‌ها و معیارهای ارزیابی پاداش در پژوهش‌های مختلف پیرامون مدل‌های مبتنی بر پیچیدگی در جدول ۴ نشان داده شده است.

جدول ۴: مزایا، محدودیت‌ها (چالش‌ها) و معیار پاداش مدل‌های مبتنی بر پیچیدگی

مرجع	مزایا	معایب/چالش‌ها/محدویت‌ها	معیار پاداش
[21]	یادگیری سریعتر در مقایسه با روش‌های جستجوی تصادفی، تمرکز کارآمد بر مناطق محیطی که بیشترین پتانسیل یادگیری را دارند، کاهش زمان هدر رفته بر روی مناطق به خوبی مدل شده یا کم بهبود.	دشواری در مدیریت محیط‌های بسیار غیرقابل پیش‌بینی یا تصادفی، تمرکز بالقوه بر بخش‌های ذاتاً غیرقابل پیش‌بینی محیط، پیچیدگی در پیاده‌سازی و تنظیم مکانیزم کنجکاوی تطبیقی.	سیستم برای اقداماتی که بهبود قابل توجهی در دقت پیش‌بینی‌های مدل جهان خود ایجاد می‌کنند، تقویت می‌شود (بر اساس تغییرات در قابلیت اطمینان پیش‌بینی‌های مدل).
[71]	تنظیم پیچیدگی فعالیت‌های یادگیری بدون نیاز به دخالت انسان، ربات با شروع از موقعیت‌های ساده و حرکت به سمت موقعیت‌های پیچیده تر، به تدریج توانایی‌های خود را توسعه می‌دهد (توسعه تدریجی)	به طور خاص به نحوه پیاده‌سازی این مکانیزم در ربات‌های واقعی نمی‌پردازد.	عدم توضیح کامل معیارهای پاداش، اما اشاره می‌کند که ربات موقعیت‌ها را بر اساس پتانسیل پیشرفت یادگیری آن‌ها با اعداد حقیقی (مثبت یا منفی) برچسب گذاری می‌کند. موقعیت‌هایی با پتانسیل یادگیری بالاتر، پاداش‌های درونی مثبت بیشتری برای ربات به همراه خواهند داشت.
[74]	عملکرد بهتر نسبت به روش‌های انگیزش درونی با افزایش پیچیدگی، کاهش نیاز به مشخص کردن وظایف صریح به وسیله ایجاد برنامه درسی خودسرپرست	پیچیدگی نمونه بالا به علت نیاز به تجربه زیاد، محدود بودن فضای استراتژی در محیط فعلی به طور ذاتی	بر اساس قابلیت دید و نتایج رقابت در بازی قایم باشک می‌باشد. عامل‌ها برای پنهان شدن موفق یا پیدا کردن مخالفان خود پاداش می‌گیرند.
[75]	امکان یادگیری مؤثر در محیط‌هایی با پاداش-های خارجی کم یا هیچ، قابلیت سازگاری با معماری‌ها و تنظیمات مختلف RL، عملکرد بهتر از روش‌های انگیزش درونی پیشرفته در محیط‌های پیچیده و تولیدی رویه‌ای	محدودیت کاربرد در محیط‌های مشاهده پذیرئی؛ چرا که پیاده‌سازی فعلی بر اساس محیط‌های کاملاً قابل مشاهده، امکان محدود شدن اثربخشی توسط نوع هدف و فضای مشاهده، فاقد اشکال انتزاعی‌تر اهداف در دامنه‌های غنی‌تر	شامل تشویق معلم به پیشنهاد اهدافی چالش‌برانگیز ولی غیرممکن نیستند، می-باشد.
[76]	تضمین‌های نظری، منابع محاسباتی کاهش‌یافته، قابلیت تعمیم قوی به محیط‌های متنوع و چالش‌برانگیز	احتمال کند کردن یادگیری به علت عدم هم‌راستایی اهداف خاص وظیفه عامل، توانایی عامل در عبور از حالت‌های ناامن یا غیر عملی به علت تکامل محیط‌های بسیار پیچیده	بر پایه اهداف پشیمانی minimax است. عملکرد عامل بر اساس توانایی‌اش در کمینه‌سازی پشیمانی در سطوح مختلف ارزیابی می‌شود و عامل را تشویق می‌کند تا وظایف پیچیده و چالش‌برانگیز را حل کند.

[77]	تنظیم پویا، تثبیت آموزش به وسیله جدا کردن اکتشاف و بهره‌برداری از تجربیات، بهبود کارایی آموزش با استفاده از مقدار کمی بازخورد انسانی و تطبیق دشواری زیرهدف	وابستگی به بازخورد انسانی، پیچیدگی پیاده‌سازی، مشکلات مقیاس‌پذیری	شامل کمینه‌سازی پشیمانی و حداکثرسازی تکمیل زیرهدف‌ها
[82]	امکان کشف مهارت به صورت خودمختار و بدون دخالت انسان، قابلیت استفاده مجدد مهارت‌های یادگرفته شده	چالش در محیط‌های پیچیده با ابعاد بالا، ایجاد اشکال در وظایفی که نیاز به کنترل دقیق یا تطبیق پویا خارج از مجموعه مهارت‌های تعریف شده دارند.	بر اساس موفقیت در انجام وظایفی که توسط بازی نامتقارن خودآموز ایجاد شده-اند، تعریف می‌شوند.

۳-۲-۵- مدل های مبتنی بر عدم قطعیت

عدم اطمینان به‌عنوان یک متغیر هم‌سنجش، نقش مهمی در تحریک کنجکاوی و پیشبرد یادگیری دارد. در یادگیری ماشین و شبکه‌های عصبی، عدم اطمینان به سطح پیش‌بینی‌ناپذیری یا ابهام موجود در داده‌های ورودی یا پیش‌بینی‌های مدل اشاره دارد. مواجهه با اطلاعات نامشخص، انگیزه‌ای درونی برای کاهش عدم اطمینان ایجاد می‌کند که منجر به اکتشاف و یادگیری می‌شود. کنجکاوی مبتنی بر عدم قطعیت، عامل‌ها را به کاوش در حالت‌ها یا اقدامات ناشناخته ترغیب می‌کند تا با کسب اطلاعات جدید، میزان عدم اطمینان را کاهش دهند. در این زمینه، عدم اطمینان را می‌توان از طریق آنتروپی در نظریه اطلاعات اندازه‌گیری کرد. آنتروپی بیانگر میزان عدم قطعیت در یک محرک است و زمانی به بالاترین حد می‌رسد که محرک با پاسخ دریافتی به‌طور قطعی شناسایی نشود [11,34].

ماکدو[1] و کاردوسو[2] [78] برای بهبود اکتشاف عامل در محیط‌های ناشناخته، در مدل کنجکاوی خود معیار عدم قطعیت را علاوه بر تازگی معرفی کردند. آن‌ها استدلال کردند که تمایل به شناخت یک شیء می‌تواند هم از تازگی و هم از عدم قطعیت ناشی شود؛ به این معنا که اشیاء دارای بخش‌های ناشناخته انگیزه‌ی بیشتری برای کاوش ایجاد می‌کنند. در این مدل، بخش‌های شناخته‌شده‌ی هر شیء با استفاده از فاصله‌ی همینگ برای تازگی اندازه‌گیری می‌شوند، در حالی که عدم قطعیت از طریق محاسبه‌ی آنتروپی بخش‌های ناشناخته، شامل توصیف‌های قیاسی و توابع شیء، تعیین می‌شود. [79] استراتژی اکتشاف VIME[3] را به‌عنوان روشی در RL پیشنهاد کرد که با حداکثرسازی اطلاعات متغیر، کنجکاوی عامل را با کاهش عدم قطعیت نسبت به دینامیک محیط تحریک می‌کند و این کاهش را به‌عنوان پاداش درونی به کار می‌گیرد. این روش عامل‌ها را به کاوش در حالت‌هایی ترغیب می‌کند که بیشترین کاهش عدم قطعیت را دارند

[1] Macedo
[2] Cardoso
[3] Variational Information Maximizing Exploration (VIME)

و برای دامنه‌های پیوسته با فضای حالت و عمل بزرگ مقیاس‌پذیر است. VIME عملکرد بهتری از روش‌های اکتشاف هیوریستیک نشان داده است، اما ناپایداری عملکرد در حضور عناصر تصادفی و محدودیت به مسائل حرکت رباتیک، از جمله محدودیت‌های این روش است. برست[1] و همکاران [81] روش «RL با حداقل‌سازی شگفتی[2]» را برای کاهش عدم قطعیت در محیط‌های پیچیده و با ابعاد بالا معرفی کردند. این روش با استفاده از رمزگذارهای خودکار متغیر[3]، نمایش غیرخطی از وضعیت را یاد می‌گیرد و به‌طور درونی عامل‌ها را تشویق می‌کند تا با حداقل‌سازی آنتروپی، در وضعیت‌های ایمن و پایدار باقی بمانند. این رویکرد در محیط‌های متنوعی مانند بازی‌های ویدیویی و کنترل ربات‌ها به کار رفته است. با این حال، چالش اصلی آن، عدم قابلیت تعمیم به وظایفی است که نیاز به اکتشاف گسترده و کنترل‌های پیچیده دارند.

به منظور ایجاد تعادل موثرتری بین اکتشاف و بهره‌برداری از تجربیات برای توسعه رفتارهای معنادار بدون انحراف از عناصر تصادفی محیط، [83] روش RL بدون نظارت به نام «شگفتی متقابل» را معرفی کرد. در آن، دو سیاست «اکتشاف» و «کنترل» با هم رقابت می‌کنند تا به ترتیب بیشینه‌سازی و کمینه‌سازی آنتروپی مشاهده را انجام دهند. این رقابت به عامل کمک می‌کند تا به کشف حالات جدید پرداخته و سپس بر آن‌ها تسلط یابد تا به تعادل مدنظرش برسد. هدف تحقیق [84]، بیشینه‌سازی همزمان بازدهی مورد انتظار و آنتروپی است که به عامل اجازه می‌دهد تا وظایف را با موفقیت انجام دهد و در عین حال سطح بالایی از تصادفی بودن در اعمالش را حفظ کند که باعث اکتشاف و پایداری می‌شود. با معرفی الگوریتمی به نام SAC[4] که یک روش عامل–منتقد مستقل از سیاست بر اساس چارچوب RL حداکثر آنتروپی است. [85] یک روش تنظیم خودکار برای پارامتر دما معرفی می‌کند که آنتروپی را به یک مقدار هدف تنظیم می‌کند و بدین ترتیب آموزش را پایدار کرده و کارایی نمونه‌ها را بهبود می‌بخشد. به حداکثر رساندن آنتروپی می‌تواند به عامل‌ها در بهبود کاوش با رفتارهای متنوع کمک کند. [87] با هدف تعادل بین اکتشاف و بهره‌برداری از تجربیات، نسخه تغییر یافته SAC را ارائه داده‌اند. این روش، دمای آنتروپی را بر اساس کنجکاوی عامل نسبت به حالات مختلف تنظیم می‌کند. حالات ناآشنا با خطاهای پیش‌بینی بالا دما، آنتروپی بیشتری دارند تا اکتشاف را تشویق کنند. در وظایف عملی مانند مسیریابی و دستکاری رباتیک، اغلب نمونه‌هایی از حالت‌های خروجی موفق در دسترس هستند که هدایت کاوش به سمت هدف نهایی را آسان‌تر می‌کند و می‌توان از اطلاعات یا تجربه‌ی قبلی بهره برد. در همین راستا لی[5] و

[1] Berseth
[2] SMIRL: Surprise MInimizing Reinfrcement Learning in unstable envirnments
[3] Variational AutoEncoders (VAE)
[4] Soft Actor-Critic
[5] Li

همکارن [88]، یک طبقه‌بندی کننده بیزی را برای تمایز بین حالات موفق و ناموفق آموزش می‌دهد. این طبقه‌بندی کننده از توزیع حداکثر درست نمایی نرمال شده[1] برای اندازه‌گیری عدم قطعیت استفاده می‌کند. عدم قطعیت اندازه‌گیری شده سپس به عنوان پاداش درونی برای تشویق اکتشاف به کار می‌رود. [89] چارچوبی پیشنهاد می‌دهد که در دو بخش اصلی عمل می‌کند. اول، از یک طبقه‌بندی کننده بیزی که شامل اندازه‌گیری عدم قطعیت بر اساس احتمال حداکثر نرمالیزه شده شرطی است برای هدایت کاوش به سمت نتایج مورد نظر استفاده می‌کند. دوم، از فاصله واسراشتاین[2]، با یک معیار زمان‌گام برای ارائه پاداش‌های درونی آگاه از فاصله زمانی و شناسایی مرزهای مناطق کاوش شده بهره می‌برد. با ترکیب این عناصر، اهداف برنامه‌ی درسی کالیبره شده‌ای را تولید می‌کند که به طور موثر بین حالت‌های اولیه و نتایج مطلوب واسطه‌گری می‌کند. این روش باعث می‌شود بهره‌وری نمونه بهبود یابد و منجر به عملکرد بهتر در آزمایش‌های کمترشود.

هدف [90] این است که امکان اکتشاف کارآمد در محیط‌های بدون فضای هدف از پیش تعریف شده را فراهم کند و بهره‌وری داده‌ها و عملکرد را در وظایف مختلف دستیابی به هدف بهبودبخشد. برای نیل به این هدف، چارچوبی دو مرحله‌ای معرفی کرده‌اند. ابتدا، فضای هدف معنایی را با کمی‌سازی فضای مشاهدات پیوسته با استفاده از «بردار کوانتیزه شده –رمزگذارهای خودکار متغیر» ایجاد می‌کند و سپس روابط زمانی بین این مشاهدات کمّی‌شده را با استفاده از یک نمودار، بازیابی می‌کند. سپس، اهداف برنامه‌ریزی شده‌ای که عدم قطعیت و فاصله زمانی را در نظر می‌گیرند، پیشنهاد می‌دهد تا عامل را به سمت مناطق ناشناخته و اهداف نهایی هدایت کند. ایده اصلی [86] ارائه روشی جدید به نام ترکیب‌بندی مجدد وزندار[3] است تا استحکام مدل‌های یادگیری ماشین را در برابر تغییرات زیرگروهی بهبود بخشد. این روش به رفع مشکلات بیش‌برازش در شبکه‌های عصبی بیش‌پارامتر[4] کمک می‌کند و عدالت را در میان زیرگروه‌های مختلف از طریق توجه بیشتر به نمونه‌های اقلیت افزایش می‌دهد. در سناریوهایی که عضویت‌های زیرگروه‌ها ناشناخته است، این روش از برآورد مبتنی بر عدم اطمینان برای تخصیص اهمیت استفاده می‌کند، که به مدل اجازه می‌دهد بیشتر بر زیرگروه‌های کم‌نماینده تمرکز کند. جدول ۵، به بررسی مزایا، چالش‌ها و معیارهای ارزیابی پاداش در تحقیقات مختلف پیرامون مدل‌های مبتنی بر تغییر می‌پردازد.

[1] Normalized Maximum Likelihood
[2] Wasserstein
[3] Reweighted Mixup
[4] over-parameterized

جدول ۵: مزایا، محدودیت‌ها (چالش‌ها) و معیار پاداش مدل‌های مبتنی بر عدم قطعیت

مرجع	مزایا	معایب/چالش‌ها/محدویت‌ها	معیار پاداش
[78]	استراتژی‌های اکتشافی طبیعی‌تر و کارآمدتر (تقلید از رفتارهای اکتشافی شبیه به انسان)، استراتژی‌های اکتشافی تطبیقی و انعطاف-پذیرتر (استفاده از انگیزه‌های احساسی)	نیاز به بررسی بیشتر تأثیر توزیع موجودیت‌ها در محیط بر عملکرد عامل-ها، دادن اهمیت یکسان به همه ویژگی-های اشیاء	شامل بازدید از موجودیت‌های جدید، نقشه-برداری از مناطق ناشناخته و دستیابی به اهداف خاص اکتشافی است. همچنین شدت احساسات مثبت (انگیزه‌ها) محاسبه می‌شود تا سودمندی اقدامات مختلف را ارزیابی کند و رفتار عامل‌ها را هدایت کند.
[79]	به طور طبیعی به فضای حالت و عمل پیوسته مقیاس می‌شود، برخلاف بسیاری از روش‌های سنتی، ارائه یک استراتژی اکتشافی کارآمدتر نسبت به روش‌های هیوریستیک، عملکرد قابل توجه در بهبود عملکرد مدل‌های اکتشافی مبتنی بر قواعد کلی در وظایف کنترل پیوسته مختلف، اجتناب از رفتارهای گام تصادفی[1]	به طور کلی غیر قابل حل بودنِ محاسبه توزیع پسین برای مدل دینامیک و نیاز به تکنیک‌های تقریبی مانند استنباط متغیر، پرهزینه بودن از نظر محاسباتی به دلیل نیاز به نگهداری و به روزرسانی یک شبکه عصبی بیزی، نیاز به تنظیم دقیق ابرپارامترها، و محدودیت به مسائل حرکت رباتیک	پاداش‌های خارجی سنتی از محیط و یک پاداش درونی اضافی مبتنی بر کسب اطلاعات درباره مدل دینامیک‌ها است. پاداش درونی به عنوان واگرایی KL[2] بین باورهای پسین و قبلی عامل درباره دینامیک‌های محیط محاسبه می‌شود.
[81]	توانایی عملکرد بدون پاداش‌های خارجی و اتکا به انگیزش درونی، کشف خودکار رفتارهای پیچیده و هماهنگ، تسلط بر رفتارهای نوظهور، یادگیری از پیکسل‌های خام	محاسبات فشرده به دلیل نیاز به به روزرسانی مستمر مدل چگالی و سیاست، وابستگی زیاد رفتار عامل به نمایش حالت انتخاب شده، عدم تعمیم موفقیت آمیز به وظایفی که نیاز به اکتشاف کامل و کنترل‌های پیچیده دارند.	بر اساس حداقل‌سازی شگفتی یا آنتروپی حالات ملاقات شده توسط عامل.
[83]	تعادل بین اکتشاف و کنترل با ترکیب سیاست‌های اکتشافی و کنترلی، جلوگیری از انحرافات تصادفی توسط عناصر غیرقابل پیش‌بینی محیط، پشتیبانی نظری (اثبات‌های رسمی) از عملکرد موثر در محیط‌های تصادفی به منظور بیشینه‌سازی پوشش	پیچیدگی پیاده‌سازی به خاطر تنظیم متقابل با دو سیاست رقابتی، امکان وابسته بودن به ویژگی‌های خاص محیط برای عملکرد بهینه، ناتوانی در سازگاری سریع با تعاملات محدود محیطی	معیارهای پاداش بر اساس آنتروپی مشاهده: سیاست اکتشاف برای افزایش آنتروپی (یافتن حالات شگفت‌انگیز) پاداش دریافت می‌کند، در حالی که سیاست کنترل برای کاهش آنتروپی (بازگشت به حالات قابل پیش‌بینی) پاداش دریافت می‌کند.

[1] Random walk
[2] Kullback–Leibler divergence

	حالات		
[84]	نیاز به تعاملات کمتری با محیط برای یادگیری سیاست‌های مؤثر، عملکرد پایدار در وظایف مختلف و seedهای تصادفی، عملکرد نهایی و کارایی نمونه بهتر در مقایسه با روش‌های مستقل از سیاست و مبتنی بر سیاست	نیاز به منابع محاسباتی قابل توجه به ویژه برای وظایف پیچیده، عملکرد ضعیف با وجود پیکسل‌های خام، یادگیری با کارایی نمونه پایین	بر پایه بیشینه‌سازی بازدهی مورد انتظار و آنتروپی
[85]	بهبود عملکرد در وظایف مرجعی مانند بازی‌های آتاری[1]	پیچیدگی پیاده‌سازی به علت نیاز به درک عمیق از هر دو معادله بلمن[2] و انتشار عدم قطعیت، بار محاسباتی بالا به علت محاسبه و انتشار عدم قطعیت در کاربردهای بزرگ مقیاس	علاوه بر پاداش‌های فوری از محیط، یک پاداش برای اکتشاف حالت‌ها و اقدامات با عدم قطعیت بالا
[87]	افزایش قابل توجه کارایی نمونه با تنظیم پویای اکتشاف بر اساس کنجکاوی عامل، برقراری تعادل بهتری بین اکتشاف و بهره برداری از تجربیات با تنظیم دمای آنتروپی بر اساس آشنایی با حالت، سازگاری با محیط‌های مختلف با تنظیم دمای آنتروپی بر اساس بازخورد بلادرنگ از مدل کنجکاوی عامل	چالش ادغام دمای آنتروپی آگاه به کنجکاوی و توسعه‌ی یک مدل کنجکاوی مناسب، وابستگی زیاد اثربخشی رویکرد به دقت مدل کنجکاوی	بیشینه‌سازی پاداش مورد انتظار در حالی که دمای آنتروپی به طور پویا بر اساس کنجکاوی عامل نسبت به حالات مختلف تنظیم می‌شود.
[88]	برطرف کردن نیاز به تابع پاداش از پیش طراحی شده، بهبود کارایی از یادگیری فراسطح، تسلط بر وظایف چالش‌برانگیز ناوبری و دستکاری رباتیک	تنها برای وظایفی با نمونه‌های خروجی موفق قابل اعمال است و گسترش آن به فرآیند تصمیم‌گیری مارکوف[3] با تنها مشاهدات بصری یا تنظیمات مسئله با مشاهده‌ی جزئی همچنان یک چالش باقی مانده است.	استفاده از عدم قطعیت خروجی طبقه‌بندی کننده به عنوان پاداش درونی
[89]	عملکرد خوب در محیط‌هایی با ساختارهای هندسی پیچیده، بهبود بهره‌وری نمونه که منجر به عملکرد بهتر در آزمایش‌های کمتر می‌شود.	پیچیدگی محاسباتی بالا به دلیل فرایند استنتاج متا-یادگیری، چالش مقیاس-پذیری به محیط‌های بسیار بزرگ و پیچیده	ارائه پاداش بر اساس آگاهی از فاصله زمانی با استفاده از فاصله واسرشتاین[4] و هدایت عامل به سمت نتایج مطلوب با پیشنهاد اهداف برنامه درسی که هر دو عدم قطعیت و فاصله زمانی را

[1] Atari
[2] Bellman
[3] Markov Decision Process
[4] Wasserstein

			در نظر می‌گیرند.
[90]	عدم نیاز به دانش قبلی در مورد محیط، مشاهدات یا فضای هدف، توانایی مدیریت مشاهدات با بعد بالا و تبدیل آن‌ها به فضای هدف قابل استفاده	وابستگی شدید اثربخشی به برآورد دقیق عدم قطعیت، ظرفیت محدود نمایندگی و تحت تاثیر قرار گرفتن توانایی مدل در مدیریت وظایف بسیار پیچیده	دریافت پاداش برای پیشرفت به سمت اهداف برنامه‌ریزی شده میانی که به عدم قطعیت و فاصله زمانی آگاه هستند.
[86]	تعمیم دقیقتری نسبت به روش‌های وزندهی اهمیت سنتی، جلوگیری از بیش‌برازش در شبکه‌های عصبی بیش‌پارامتر	پیچیدگی محاسباتی برای داده‌های بزرگ یا زیرگروه‌های پیچیده، عدم امکان شناسایی دقیق زیرگروهها	بر اساس اطمینان از عملکرد استوار و عادلانه مدل در میان همه زیرگروه‌ها

۳-۲-۶- مدل‌های مبتنی برشگفتی و تناقض

در ادبیات کنجکاوی محاسباتی، شگفتی به دو صورت تفسیر می‌شود. نخست، به‌عنوان نتیجه‌ای غیرمنتظره که از اختلاف بین پیش‌بینی عامل و واقعیت مشاهده‌شده ناشی می‌شود [91]؛ این تفاوت از طریق خطای پیش‌بینی یا تفاوت بین بردارهای پیش‌بینی و مشاهده اندازه‌گیری می‌شود [54,92]. تفسیر دوم شگفتی را به‌عنوان درجه عدم انتظار چیزی توصیف می‌کند و از افزایش اطلاعات قبل و بعد از مشاهده برای مدل‌سازی آن استفاده می‌شود [93]. با این حال، عامل‌های مبتنی بر شگفتی ممکن است به مشاهدات نویزی یا غیرقابل‌پیش‌بینی جذب شوند که به «حواس‌پرتی[1]» یا «شگفتی کاذب[2]» منجر می‌شود (مانند تماشای صفحه تلویزیونی با نویز سفید). در برخی محیط‌ها، رویدادهای غیرمنتظره و برجسته[3] مانند روشن و خاموش شدن چراغ‌ها یا شروع و پایان موسیقی، کنجکاوی مبتنی بر تناقض را تحریک می‌کنند. می‌توان با ارائه پاداش مبتنی بر تناقض، عامل‌ها را به یادگیری از این رویدادها ترغیب کرد. در RL، شگفتی و تناقض را می‌توان در قالب انحراف از باور داخلی عامل ادغام کرد و از خطاهای پیش‌بینی برای تشویق به کاوش در حالت‌های کمتر کاوش‌شده و به‌روزرسانی دانش عامل بهره برد [51].

در ادامه، به بررسی پژوهش‌های صورت گرفته در این حوزه پرداخته می‌شود منتها باید اشاره کرد که در بعضی تحقیقات فقط از شگفتی، برخی فقط تناقض و در سایر تحقیقات از هر دو متغیر استفاده شده است. اشمیدهوبر[4]، به‌عنوان یکی از پیشگامان، کنجکاوی مصنوعی را در سیستم‌های کنترل ساخت مدل معرفی کرد [94]. هدف این سیستم،

[1] distraction
[2] fake surprise
[3] salient
[4] Schmidhuber

یادگیری نگاشت ورودی–خروجی در محیطی نویزی است، که در آن کنجکاوی به‌منظور تقویت تمایل درونی سیستم به بهبود دانش خود از جهان تزریق می‌شود. این هدف با افزودن واحد تقویتی به کنترل‌کننده حاصل می‌شود که اعمال منجر به خطاهای پیش‌بینی بالا را پاداش می‌دهد. این خطای پیش‌بینی، که عدم تطابق میان باور و واقعیت را اندازه‌گیری می‌کند، مطابق با اولین تفسیر از شگفتی است. مکانیزم پاداش‌دهی به کنجکاوی بالا، سیستم را تشویق می‌کند تا موقعیت-های ناشناخته و دارای شگفتی را کاوش کرده و از یکنواختی اجتناب کند، در نتیجه سیستم به سمت یادگیری از موقعیت‌های مغایر و چالش‌برانگیز هدایت می‌شود. [96] چگونگی کمک یادگیری با انگیزه درونی به عامل‌های مصنوعی برای توسعه و گسترش سلسله مراتب مهارت‌های قابل استفاده مجدد را بررسی می‌کند. این رویکرد با هدف توانمندسازی خودمختاری عامل‌ها طراحی شده است، به‌گونه‌ای که مهارت‌هایی را نه‌تنها برای حل مسائل خاص بلکه برای خود یادگیری، مشابه انسان‌ها و حیوانات، کسب کنند. این روش با ارائه چارچوبی محاسباتی در RL، یادگیری با انگیزه درونی را ممکن می‌سازد و از الگوریتم‌های RL سلسله‌مراتبی برای ساخت و گسترش مجموعه‌ای از مهارت‌های قابل استفاده مجدد بهره می‌گیرد. در این فرایند، پاداش‌های درونی ناشی از کنجکاوی و علاقه به رویدادهای جدید و غیرمنتظره، یادگیری را هدایت می‌کنند.

برای توسعه مفهوم کنجکاوی مبتنی بر تناقض، [97] الگوریتمی در RL با چارچوب گزینه‌ها ارائه کرده است که به بررسی چگونگی استفاده از انگیزه درونی برای ترغیب عامل‌ها به کاوش گسترده و یادگیری عمومی می‌پردازد. در این الگوریتم، احتمال عدم انتظار یک رویداد برجسته به‌عنوان پاداش درونی به کار می‌رود. عامل با برخورد مکرر با یک رویداد غیرمنتظره، یاد می‌گیرد آن را پیش‌بینی کند و به مرور با بهبود سیاست گزینه، پاداش درونی کاهش می‌یابد؛ سپس عامل نسبت به آن رویداد «خسته[1]» شده و به کاوش جنبه‌های جدید محیط می‌پردازد. این الگوریتم، با استفاده از پاداش‌های ذاتی مبتنی بر تناقض، یادگیری مؤثر و عملکرد مطلوبی را در عامل‌ها ایجاد می‌کند. اما فرض بنیادی این روش، قابل مشاهده و قطعی بودن رویداد برجسته است که ممکن است در محیط‌های پیچیده یا وظایفی که شناسایی رویدادهای برجسته در آن‌ها دشوار است، محدودیت ایجاد کند. ایده اصلی [98]، روش Plan2Explore است که با بهره‌گیری از یادگیری خودسرپرست و مدل‌های جهانی، چالش‌های اکتشاف و انطباق در RL را مدیریت می‌کند. این روش به عامل امکان می‌دهد محیط را از طریق برنامه‌ریزی برای نوآوری‌های آینده به‌طور کارآمد کاوش کند و با تعامل حداقلی به سرعت با وظایف جدید سازگار شود. این روش با شبیه‌سازی وضعیت‌های آینده و ایجاد مسیرهایی برای حداکثرسازی پاداش‌های درونی، به حداکثر اطلاعات دست می‌یابد. نشان داده شده است که حداکثرسازی واریانس

[1] Bored

میانگین‌های مجموعه، تقریباً به حداکثر کردن به دست آوردن اطلاعات منجر می‌شود. برای دستیابی به عملکرد بهینه و سازگاری سریع، مدل جهانی باید با اکتشاف کافی آموزش ببیند تا کارایی نمونه‌ها را نیز افزایش دهد.

برای بهبود کارایی نمونه در وظایف کنترل پیوسته، [99] چارچوبی نوین به نام «مدل دینامیک پیشرو کنجکاوی تقابلی[1]» از یادگیری متضاد با یک مدل دینامیک پیش‌رو برای یادگیری خود–نظارت‌شده‌ی بازنمایی استفاده می‌کند. این مقاله با ادغام چندین مؤلفه کار می‌کند: رمزگذار تصویر، رمزگذار تکانه[2]، مدل دینامیک پیشرو FDM[3] و ماژول کنجکاوی. FDM پیش‌بینی وضعیت‌های آینده از وضعیت‌های فعلی را انجام می‌دهد و یادگیری تقابلی به آموزش رمزگذار تصویر برای استخراج ویژگی‌های فضایی و زمانی کمک می‌کند. در طول آموزش، پاداش‌های درونی بر اساس خطاهای پیش‌بینی FDM برای تشویق به کاوش ارائه می‌شود. این چارچوب به‌طور کامل با الگوریتم RL آموزش می‌بیند و بهبودهایی در کارآیی نمونه‌برداری و تعمیم نشان می‌دهد. [102] روش جدیدی به نام RL چندعامله با کاوش کنجکاوی محور اپیزودیک معرفی کرده است. هدف این تحقیق، برطرف کردن چالش‌های کاوش کارآمد و آموزش سیاست‌ها در RL چندعامله عمیق است. برای این منظور، از خطاهای پیش‌بینی مقادیر Q-value فردی به عنوان پاداش‌های درونی برای کاوش هماهنگ استفاده می‌کند و از حافظه اپیزودیک برای بهره‌برداری از تجربیات اطلاعاتی استفاده می‌کند. پاداش درونی به دست آمده از دینامیک تابع Q-value یک عامل، کاوش هماهنگ‌شده به سمت حالات جدید یا امیدوارکننده را تشویق می‌کند.

روشی برای اکتشاف خود–نظارتی در RL که از یک پاداش درونی مبتنی بر کنجکاوی استفاده می‌کند توسط [103] ارائه شده است. این پژوهش یک مدل دینامیک نهفته را معرفی می‌کند که حالت مشاهده نشده محیط را ثبت کرده و مشاهدات را در یک فضای ویژگی فشرده بازسازی می‌کند. شگفتی بیزی که نشان دهنده کسب اطلاعات نسبت به متغیر حالت نهفته است، به عنوان پاداش درونی استفاده می‌شود و عامل را تشویق می‌کند تا محیط خود را به طور کامل‌تری کاوش کند، بدون اینکه به پاداش‌های طراحی شده خارجی وابسته باشد. مدل از طریق وظایف شبیه‌سازی رباتیک با اقدامات پیوسته و بازی‌های ویدئویی با اقدامات گسسته ارزیابی شده و کارایی آن در اکتشاف و مقاومت در برابر اقدامات تصادفی نشان داده می‌شود. هدف [104] این است که مشابه کنجکاوی انسان، عامل‌های مصنوعی را به طور طبیعی و مؤثرتری به کاوش در محیط‌های خود بکشاند. این روش با توسعه یک مدل دینامیک پنهان که فهم عامل از دینامیک سیستم را به تصویر می‌کشد، عمل می‌کند. بهبود کاوش در RL با استفاده از پاداش‌های درونی مبتنی بر شگفتی

[1] Curiosity Contrastive Forward Dynamics Model
[2] Momentum
[3] forward dynamics model

بیزی در فضای پنهان انجام شده است. این پاداش به عنوان تفاوت بین باورهای پسین و پیشین در این فضای پنهان محاسبه می‌شود. [105] مدل جدیدی برای پاداش‌های درونی در RL معرفی کرده است که روش‌های اکتشاف مبتنی بر شگفتی موجود را بهبود می‌بخشد. به جای پاداش دادن به عامل‌ها بر اساس میزان شگفتی، به نویی شگفتی پاداش می‌دهد و بدین ترتیب کارایی اکتشاف را افزایش داده و حواس‌پرتی‌ها از مشاهدات پر سر و صدا یا غیرقابل پیش‌بینی را کاهش می‌دهد. این مسئله توسط یک سیستم حافظه شگفتی برای ذخیره و بازسازی شگفتی‌ها استفاده می‌کند. این سیستم با استفاده از یک رمزگذار خودکار، نوآوری شگفتی‌ها را بر اساس خطاهای بازسازی ارزیابی می‌کند. پاداش درونی متناسب با نوآوری شگفتی است و عامل را تشویق می‌کند تا جنبه‌های کمتر قابل پیش‌بینی یا جدید محیط را کشف کند. در ادامه، برای آنالیز بهتر، مزایا، چالش‌ها و معیارهای ارزیابی پاداش در تحقیقات ذکر شده در بالا در جدول ۶ مورد بررسی قرار گرفته است.

جدول ۶: مزایا، محدودیت‌ها (چالش‌ها) و معیار پاداش مدل‌های مبتنی بر شگفتی و تناقض

مرجع	مزایا	معایب/چالش‌ها/محدویت‌ها	معیار پاداش
[94]	تطبیق با محیط‌های مختلف با تغییر رفتار خود بر اساس سطح اعتماد، جلوگیری از بیش‌برازش به داده‌های اولیه و شاید جانبدارانه	پیچیده بودن و بار محاسباتی بالای ادغام مکانیزم‌های اعتماد تطبیقی و کنجکاوی، وابستگی اثربخشی سیستم به اندازه‌گیری دقیق اعتماد	مرحله اکتشاف: اختصاص پاداش برای کشف حالت‌های جدید و اطلاعاتی که عدم قطعیت را کاهش می‌دهند. مرحله بهره‌برداری از تجربیات: بر اساس دستیابی به اهداف از پیش تعیین شده یا به حداکثر رساندن پاداش‌های شناخته شده
[96]	توسعه مهارت‌های خودمختار، انعطاف-پذیری	پیچیدگی ساختار سلسله‌مراتبی و مکانیزم-های انگیزش درونی، نیاز به به منابع محاسباتی قابل توجه، چالش گسترش به محیط‌های بسیار بزرگ و پیچیده	دریافت پاداش برای انجام فعالیت‌هایی که ذاتاً جالب یا جدید هستند. این شامل کشف مهارت‌های جدید، کاوش محیط‌های ناشناخته و مواجهه با رویدادهای غیرمنتظره است.
[97]	الهام از علوم اعصاب با بهره‌مندی از بینش-های مربوط به نقش دوپامین[1] در تازگی و کاوش، میانگین‌گیری از رویدادهای برجسته غیرمنتظره برای بهبود کارایی یادگیری	عدم پوشش دهی تمامی اشکال کاوش و دستکاری جالب با تمرکز اولیه بر رویدادهای غیرمنتظره به عنوان پاداش‌های درونی، نیاز به مشاهده‌پذیر و قطعی بودن رویداد برجسته	علاوه بر پاداش‌های بیرونی معمول، پاداش‌های درونی توسط منتقد عامل تولید می‌شوند. در این پیاده‌سازی، پاداش درونی عامل به روشی مشابه پاسخ تازگی نورون‌های دوپامین تولید می‌شود. پاداش درونی برای هر رویداد برجسته متناسب با خطای پیش‌بینی رویداد برجسته طبق مدل گزینه یاد گرفته شده برای آن رویداد است.

[1] Dopamine

[98]	انطباق سریع به وظایف جدید با تعامل حداقلی، کاهش نیاز به جمع آوری داده‌های گرانقیمت، نیاز به اکتشاف کافی برای مدل جهان، یادگیری با بازدهی نمونه کم	چالش برانگیز بودن آموزش یک مدل جهانی دقیق از ورودی‌های با ابعاد بالا، نیاز به تحقیقات بیشتری برای اطمینان از عمومیت خوب آن به طیف وسیعی از وظایف و محیط‌ها	شامل حداکثر کردن پاداش‌های درونی بر اساس نوآوری آینده مورد انتظار: تمرکز بر وضعیت-هایی که پیش‌بینی می‌شود در آینده نوآورانه باشند و نه وضعیت‌هایی که در گذشته نوآورانه بوده‌اند.
[99]	بهبود تعمیم با استخراج اطلاعات فضایی و زمانی، قابلیت اکتشاف در محیط‌های تصادفی	نیاز به منابع محاسباتی زیاد برای آموزش با داده‌های با ابعاد بالا و مدل‌های پیچیده، وابستگی اثربخشی روش به دقت مدل دینامیک پیشرفته	شامل پاداش‌های درونی و بیرونی: پاداش‌های درونی بر اساس خطاهای پیش‌بینی مدل دینامیک پیشرفته و پاداش‌های خارجی بر اساس بازخورد محیط
[102]	بهبود کارایی یادگیری با استفاده از حافظه اپیزودیک، مقیاس‌پذیر و مناسب برای محیط‌های چندعامله بزرگ با استفاده از تکنیک‌های فاکتوریزاسیون ارزش	فاقد تکنیک‌های کاوش تطبیقی برای اطمینان از پایداری در محیط‌های مختلف، مشکلات حافظه اپیزودیک در تنظیمات تصادفی	شامل پاداش‌های بیرونی از محیط و پاداش‌های درونی ناشی از خطاهای پیش‌بینی مقادیر Q فردی
[105]	جلوگیری از حواس‌پرتی‌ها از مشاهدات پر سر و صدا یا غیرقابل پیش‌بینی با تمرکز بر نوآوری تعجب‌ها، عملکرد بهبود یافته در محیط‌های با پاداش‌های کم و وظایف چالش‌برانگیز مانند بازی‌های آتاری	بار محاسباتی به خاطر حافظه اضافی و محاسبات مورد نیاز برای سیستم رمزگذار خودکار و حافظه تعجب، چالش در مقیاس‌بندی آن به محیط‌های پیچیده‌تر	شامل نوآوری تعجب است که توسط خطای بازسازی از رمزگذار خودکار در سیستم حافظه تعجب ارزیابی می‌شود.
[103]	مقاومت در برابر تصادفی بودن، ارزان بودن محاسباتی نسبت به سایر روش‌های پیشرفته	پیچیدگی مدل‌سازی حالت نهفته و نیاز به تنظیم دقیق، محدودیت به برخی محیط‌ها	بر اساس تعجب بیزی است که میزان کسب اطلاعات نسبت به متغیر حالت نهفته را اندازه-گیری می‌کند.
[104]	فراهم کردن کاوش عمیق‌تر به خصوص در وظایف با ابعاد پایین و بالا	چالش در محیط‌هایی با سطوح بالای تصادفی بودن، تفاوت جدی عملکرد در حوزه‌ها و انواع وظایف مختلف	بر پایه تعجب بیزی پنهان[1] است که تفاوت بین باورهای پسین و پیشین عامل در فضای پنهان را نشان می‌دهد.

۳-۲-۷- مدل‌های مبتنی بر پوشش فضا

همان‌طور که در بخش ۳-۱ بیان شد، ما متغیر جدیدی به نام پوشش فضا برای متغیرهای هم‌سنجش معرفی کرده‌ایم. پوشش فضایی نشان می‌دهد که چه بخشی از یک فضای مسئله توسط یک عامل کاوش شده است. این متغیر جدید با سایر متغیرها ارتباط نزدیکی دارد:

[1] Latent Bayesian Surprise

- تازگی: پوشش فضا با تازگی مرتبط است، زیرا اکتشاف مناطق جدید به طور ذاتی شامل مواجهه با محرک‌های جدید است.
- تغییر: حرکت و اکتشاف مکرر منجر به درجه بالایی از تغییر در محرک‌های درک شده توسط عامل می‌شود.
- شگفتی: کشف ویژگی‌ها یا عناصر غیرمنتظره در مناطق جدید می‌تواند به شگفتی تجربه کمک کند.
- پیچیدگی: پیچیدگی محیط می‌تواند بر رفتار پوشش فضایی عامل تأثیر بگذارد، زیرا محیط‌های پیچیده‌تر ممکن است فرصت‌های اکتشافی غنی‌تری را ارائه دهند.
- عدم قطعیت: پوشش فضایی بالاتر می‌تواند منجر به افزایش عدم قطعیت به دلیل قرار گرفتن در معرض مناطق ناشناخته یا کمتر قابل پیش‌بینی شود.
- تعارض: تصمیم به اکتشاف مناطق جدید ممکن است شامل تعارض بین تمایل به اکتشاف و انگیزه‌های رقابتی دیگر مانند نیاز به ایمنی یا کارایی باشد.

این متغیر همانند دیگر متغیرهای هم‌سنجش، پوشش فضایی به عنوان جنبه‌ای از کنجکاوی را کمی‌سازی می‌کند. به طور خاص، این متغیر گستردگی اکتشاف را اندازه‌گیری می‌کند و نشان می‌دهد که عامل چه مقدار از فضای مسئله را بازدید یا با آن تعامل داشته است.

- اندازه‌گیری: درجه پوشش فضایی را می‌توان با موارد زیر اندازه‌گیری کرد:
  - درصد فضای کل کاوش شده: نسبت فضای کاوش شده به کل فضای موجود در محیط.
  - تنوع مناطق بازدید شده: تنوع در انواع مناطق یا بخش‌های درون فضای مسئله که عامل کاوش کرده است.
  - فرکانس بازدید از مناطق جدید: تعداد دفعاتی که عامل به مناطق جدید و قبلاً بازدید نشده می‌رود.

روش‌های این دسته به دنبال حداکثرسازی پوشش فضایی (بازدید از حالت‌های جدید یا جفت‌های (حالت،عمل) هستند که از نوعی انگیزش درونی برای کنترل رفتار اکتشافی عامل استفاده می‌کنند. پوشش فضا اساساً به معنای بازدید از حالت‌های بیشتر کاوش نشده در مدت زمان کوتاه‌تر و در نتیجه یادگیری بیشتر در مورد محیط است [106].

ماچادو[1] و همکاران [107] از مفهوم توابع ارزش پروتو (PVF)[2] برای کشف گزینه‌هایی استفاده می‌کنند که عامل را به سمت اکتشاف کارآمد فضای حالت هدایت کند. برای تقریب این توابع ارزش از خواص توپولوژیکی فضای حالت استفاده می‌شود. با معرفی مفهوم اهداف ویژه[3] که توابع پاداش درونی مشتق شده از PVFها هستند عامل را به کاوش کارآمد در فضای حالت با پیروی از جهت‌های اصلی نمایش یادگیری شده هدایت می‌کند. به طور خاص، با

[1] Machado
[2] proto-value functions
[3] Eigenpurposes

استفاده از ماتریس انتقال MDP، یک مدل انتشار[1] تولید می‌شود که شکل قطری آن متعاقباً منجر به تولید PVFها می‌شود. این مدل، جریان اطلاعات انتشار در محیط را فراهم می‌کند تا به PVFها اجازه دهد اطلاعات مفیدی در مورد هندسه محیط، از جمله نقاط تنگنا[2]، ارائه دهد. کار بعدی ماچادو و همکاران [108] نسخه بهبود یافته‌ای از پژوهش قبلیشان [107] بود که آن را به محیط‌های تصادفی با حالت‌های غیر جدولی گسترش می‌دهد. [110] روش پوشش گزینه‌ها[3]را معرفی کردند که در آن گزینه‌ها با هدف کمینه کردن زمان پوشش تولید می‌شوند. روش پیشنهادی آن‌ها بدون استفاده از اطلاعات به دست آمده از پاداش‌های بیرونی، عامل را تشویق می‌کند تا مناطق کمتر کاوش شده فضای حالت را با تولید گزینه‌ها برای آن قسمت‌ها کاوش کند. ارزیابی تجربی آن‌ها در دامنه‌های پاداش پراکنده گسسته، زمان یادگیری را در مقایسه با برخی از کارهای پیشین کاهش می‌دهد.

روش گزینه‌های پوشش عمیق معرفی شده در [112]، گزینه‌های پوشش را به فضاهای حالت بزرگ یا پیوسته گسترش می‌دهد، در حالی که زمان پوشش مورد انتظار عامل در فضای حالت را به حداقل می‌رساند. نویسندگان با موفقیت رفتار گزینه‌های پوشش عمیق را در وظایف چالش‌برانگیز پاداش پراکنده نشان داده‌اند. امین[4] و همکاران [113] رویکردی متمایز برای اکتشاف با انگیزش درونی به منظور تاکید بر پوشش فضا در نظر گرفته‌اند. آن‌ها با الهام از تئوری زنجیرهایی با چرخش آزاد در فیزیک پلیمر، روشی را برای اکتشاف توسعه داده‌اند. ین تئوری به توضیح رفتار مدل‌های ساده شده پلیمری کمک می‌کند، جایی که زنجیره‌ها از بخش‌های مرتبط به‌لحاظ جهت‌گیری تشکیل شده‌اند. این رویکرد برای وظایفی با فضای حالت و عمل پیوسته و همچنین پاداش‌های پراکنده طراحی شده است. استراتژی اکتشافی آن‌ها، اعمالی را با جهت‌گیری‌های همبسته در فضای عمل انتخاب می‌کند، که منجر به مسیرهای ماندگاری در فضای حالت می‌شود. آن‌ها نشان می‌دهند که روش پیشنهادی‌شان در مقایسه با نمونه‌های مشابه، عملکرد باثبات‌تری دارد و حساسیت کمتری نسبت به افزایش پراکندگی پاداش‌های بیرونی نشان می‌دهد. [114] برای مقابله با مشکلات مربوط به اعمال با فرکانس بالا، ایده «پایداری عمل[5]» را ارائه می‌کند تا عامل مناطق وسیع‌تری از فضای حالت را کاوش کرده و برآورد اثرات اعمال را بهبود بخشد. در نتیجه، درک خود را از نتایج حاصل از اعمالشان افزایش دهند. آن‌ها یک عملگر بلمن با پایداری جدید توسعه می‌دهند که امکان استفاده کارآمد از تجربیات با هر دو پایداری کم و زیاد را برای به‌روزرسانی مقادیر اعمال فراهم می‌کند. این مقاله شامل ارزیابی تجربی در سناریوهای مختلف، از جمله زمینه‌های جدولی و محیط‌های پیچیده مانند بازی‌های آتاری است.

[1] Diffusion model
[2] Bottlenecks
[3] Covering options
[4] Amin
[5] action persistence

ایده اصلی [109] معرفی روشی جدید در RL چندعاملی برای بهبود کاوش در محیط‌هایی با پاداش خلوت است. به جای تمرکز بر یک سیاست مشترک، این روش از یک هدف کاوش مبتنی بر آگاهی از اعضا استفاده می‌کند که در آن هر عضو به حداکثر رساندن کاوش در مناطقی که سایر اعضا کاوش نکرده‌اند، می‌پردازد. همچنین از یک محدودیت سیاست افزایش‌یافته برای کاوش استفاده می‌کند که باعث می‌شود سیاست‌ها از یکدیگر متمایز شوند. این تکنیک بر روی سه محیط پیچیده آزمایش شده است و عملکرد بهتری در شرایط پاداش خلوت از خود نشان داده است. آمورتیلا[1] و همکاران [111] بر معرفی «اهداف کاوش» به عنوان چارچوبی برای بهبود کاوش در RL تمرکز دارند. برای کارآمدسازی محاسباتی کاوش در محیط‌های با ابعاد بالا هدفی به نام L1-Coverage تعریف می‌کنند و نشان می‌دهند که چگونه این هدف می‌تواند از طریق کاهش وظایف کاوش به بهینه‌سازی خط مشی استاندارد به طور کارآمدی کار کند. این هدف با روش‌های رایج RL، مانند گرادیان خط مشی یا یادگیری Q، یکپارچه می‌شود تا به صورت سیستماتیک و کارآمد فضای حالت را پوشش دهد. جدول ۷ مزایا، چالش‌ها و معیارهای ارزیابی پاداش در تحقیقات مذکور را مورد بررسی قرار داده است.

جدول ۷: مزایا، محدودیت‌ها (چالش‌ها) و معیار پاداش مدل‌های مبتنی بر پوشش فضا

مرجع	مزایا	معایب/چالش‌ها/محدویت‌ها	معیار پاداش
[113]	سازگاری با فضاهای پیوسته، کاهش حساسیت به پراکندگی پاداش، بهبود کاوش	پیچیدگی در ابعاد بالا، وابستگی به تنظیم پارامترها	تمرکز بر استراتژی‌های کاوش که پوشش فضای حالت را به حداکثر می‌رسانند. عملکرد بر اساس میزان مؤثر بودن کاوش در مناطق جدید و جلوگیری از گرفتار شدن ارزیابی می‌شود.
[111]	کارآمد برای فضاهای با ابعاد بالا، یکپارچگی با روشهای رایج RL مانند یادگیری Q	وابستگی به ساختار MDP و محدود شدن کاربردپذیری کلی، پیچیدگی پیاده‌سازی	کاوش در فضای حالت به منظور جمع‌آوری اطلاعات به جای تمرکز صرف بر پاداش‌های بیرونی
[114]	همگرایی سریع‌تر با به‌روزرسانی همزمان تخمین‌های ارزش برای تمامی پایداری‌های عمل، تخمین ارزش بهتر (عملگر بلمن)	سوگیری در تخمین بیش از حد، پیچیدگی پیاده‌سازی عملگر بلمن در محیط‌های پیچیده	بر پایه پایداری عمل و توانایی استفاده از انتقالات در مقیاس‌های زمانی مختلف برای به‌روزرسانی تخمین‌های ارزش است.
[110]	کاوش بهبود یافته در محیط‌هایی با پاداش خلوت، پایه نظری قوی، تضمین یافتن گزینه‌های بهینه	پیچیدگی در فضاهای حالت بزرگ، وابستگی به گراف فضای حالت.	پاداش درونی بر اساس کاهش زمان پوشش مورد انتظار است.

[1] Amortila

[112]	کاهش موثر زمان پوشش مورد انتظار، امکان استفاده در هر دو تنظیمات پیش آموزشی و آنلاین سنتی	پرهزینه بودن تخمین توابع ویژه لاپلاسین[1] از نظر محاسباتی، نیاز به یک گراف فضای حالت با تعریف خوب که ممکن است در برخی حوزه‌ها به راحتی در دسترس نباشد.	معیارهای پاداش شامل کشف و بازدید از نواحی کمتر کاوش شده فضای حالت است که به طور موثر گام‌های لازم برای پوشش کل فضای حالت را کاهش می‌دهد. پاداش درونی بر اساس به حداقل رساندن زمان پوشش مورد انتظار است.
[107]	مستقل از وظیفه، امکان تعمیم به هر دو روش جدولی و تقریب تابع	چالش در محیط‌هایی با داده پراکنده چرا که نیاز به یک مرحله یادگیری نمایش اولیه دارد، قابلیت اجرا تنها برای دامنه‌های گسسته	معیارهای پاداش شامل پاداش‌های درونی (اهداف ویژه) مشتق شده از PVFها یادگیری شده است که عامل را به کاوش در فضای حالت با پیروی از جهت‌های اصلی تشویق می‌کند.
[108]	مدیریت انتقال‌های تصادفی، عدم اتکا به ویژگی‌های از پیش تعریف شده، قابل استفاده در هر دو محیط جدولی و محیط‌های پیچیده مانند بازی‌های آتاری ۲۶۰۰	نیاز به داده زیاد برای یادگیری نمایش‌ها از پیکسل‌های خام، عدم پرداختن مقاله به چگونگی انتقال گزینه‌های یادگیری شده در محیط‌های مختلف	شامل پاداش‌های درونی مشتق شده از نمایش جانشین[2] است.
[109]	بهبود کاوش در محیط‌های با پاداش کم، تنوع بیشتر در رفتار سیاست‌ها که باعث کاهش تکرار و بهبود نتایج یادگیری می‌شود.	تعاملات ترتیبی بین هر سیاست و محیط ممکن است منجر به ناکارآمدی در استفاده از منابع محاسباتی شود، نگهداری تعداد زیادی از سیاست‌ها برای محیط‌های با ابعاد بالا ممکن است از نظر محاسباتی هزینه‌بر باشد.	تمرکز بر حداکثر پوشش کاوش توسط هر عضو

۴- بنچمارک‌ها و پایگاه‌های داده، محدویت‌ها، کاربردها و جهت‌گیری آینده

این بخش به بررسی بنچمارک‌ها و پایگاه‌های داده، محدودیت‌ها، کاربردها و جهت‌گیری‌های آینده سیستم‌های یادگیری مبتنی بر کنجکاوی را از منظر کاربردی مورد بررسی قرار می‌دهد.

۴-۱- بنچمارک‌ها و پایگاه‌های داده

برای ارزیابی و مقایسه عملکرد مدل‌های محاسباتی کنجکاوی، استفاده از بنچمارک‌های معتبر بسیار مهم است. این بنچمارک‌ها، مجموعه داده‌ها و معیارهای ارزیابی مشخصی را ارائه می‌دهند که محققان می‌توانند از آن‌ها برای مقایسه مدل‌های خود با سایر مدل‌ها استفاده کنند. برخی از منابع و پایگاه‌های داده‌ای که در این زمینه مفید هستند، در جدول ۸ معرفی شده‌اند.

[1] Laplacian
[2] successor representation

جدول ۸: برخی از مجموعه داده‌ها و بنچمارک‌های معتبر و پرکاربرد

نام	توضیحات	آدرس
Gymnasium	نسخه‌ای بهبودیافته از OpenAI Gym است که محیط‌های استاندارد، متنوع و سازگار برای RL ارائه می‌دهد.	https://github.com/Farama-Foundation/Gymnasium
MuJoCo[1]	شبیه‌ساز فیزیکی است که به‌طور گسترده در رباتیک و تست الگوریتم‌های کنترل استفاده می‌شود و از مدل‌های دینامیکی پیچیده پشتیبانی می‌کند.	https://www.mujoco.org
DeepMind Lab	پلتفرم تحقیقاتی سه‌بعدی است که برای آزمایش مدل‌ها و شبیه‌سازی رفتارهای عامل‌ها در محیط‌های پیچیده طراحی شده است.	https://github.com/deepmind/lab
RoboSuite	مجموعه‌ای از محیط‌های شبیه‌سازی رباتیک است که تمرکز ویژه‌ای بر وظایف پیچیده دستکاری و کنترل در ربات‌ها دارد.	https://robosuite.ai
Meta-World	مجموعه‌ای از محیط‌های شبیه‌سازی رباتیک و وظایف متنوع است که برای یادگیری چندوظیفه ای و متا استفاده می‌شود و برای تست توانایی الگوریتم‌ها در تعمیم به وظایف جدید طراحی شده است.	https://meta-world.github.io
ALE[2]	مجموعه‌ای از بازی‌های کلاسیک است که به‌عنوان یک بنچمارک استاندارد با ساختار ساده و محیط‌های چالش‌برانگیز از محبوبیت بالایی برخوردار است.	https://github.com/Farama-Foundation/Arcade-Learning-Environment
RLlib	یک کتابخانه مقیاس‌پذیر است که روی پلتفرم Ray ساخته شده و امکان اجرای الگوریتم‌های یادگیری توزیع‌شده و مقیاس‌پذیر را فراهم می‌کند.	https://github.com/ray-project/ray/tree/master/rllib
ViZDoom	یک پلتفرم یادگیری است که بر اساس بازی Doom ساخته شده و به عامل‌ها اجازه می‌دهد تا در یک محیط سه‌بعدی چالش‌برانگیز آموزش ببینند. این بنچمارک به دلیل قابلیت آزمایش تعاملات پیچیده و واقع‌گرایانه عامل با محیط و دشمنان، برای تست یادگیری و کنترل در محیط‌های پویا بسیار استفاده می‌شود.	https://github.com/Farama-Foundation/ViZDoom
Jaco Robotic Arm	این مجموعه، یک شبیه‌ساز برای بازوی رباتیک Jaco است که به‌طور خاص برای آزمایش الگوریتم‌های یادگیری در کنترل دقیق و انجام وظایف پیچیده‌ی دستکاری طراحی شده است.	https://github.com/panagiotamoraiti/Jaco_Robotic_Arm
MiniGrid	مجموعه‌ای از محیط‌های ساده و انعطاف‌پذیر که برای تست الگوریتم‌های مختلف، به‌ویژه در زمینه‌های کاوش و تعمیم در محیط‌های ساده و دو بعدی، طراحی شده است.	https://minigrid.farama.org
CARLA[3]	یک شبیه‌ساز منبع‌باز برای تحقیق در حوزه‌ی خودران‌ها است که در محیط‌های واقع‌گرایانه شهری استفاده می‌شود.	https://carla.org

در ادامه، بررسی‌های محدودیت‌ها، کاربردها و جهت‌گیری‌های آینده سیستم‌های یادگیری مبتنی بر کنجکاوی در یک نگاه در شکل ۲ نشان داده شده است.

[1] Multi-Joint dynamics with Contact

[2] ATARI 2600 Games (Arcade Learning Environment

[3] Car Learning to Act

۴-۲- محدودیت‌ها

مفهوم کنجکاوی محاسباتی، به ویژه در زمینه یادگیری مبتنی بر کنجکاوی، توجه قابل توجهی در حوزه AI جلب کرده است. در حالی که یادگیری مبتنی بر کنجکاوی پتانسیل بهبود اکتشاف و کارایی یادگیری در سیستم‌های AI را نشان داده است، اما با مجموعه‌ای از محدودیت‌هایی همراه است که برای کاربرد گسترده‌تر و اثربخشی باید برطرف شوند. در ادامه به برخی از این محدودیت‌ها اشاره می‌شود:

- **کمی‌سازی:** یکی از محدودیت‌های اصلی مدل‌های یادگیری مبتنی بر کنجکاوی، وابستگی آن‌ها به اندازه‌گیری و کمی‌سازی دقیق تازگی است. تازگی، به عنوان یک مؤلفه کلیدی در یادگیری مبتنی بر کنجکاوی، اغلب با استفاده از معیارهایی مانند خطای پیش‌بینی یا افزایش اطلاعات کمی‌سازی می‌شود. با این حال، این معیارها می‌توانند نادقیق و وابسته به زمینه باشند که منجر به استراتژی‌های اکتشاف نامطلوب می‌شود. برای مثال، الگوریتم‌ها ممکن است اکتشاف را در مناطقی از فضای حالت که جدید اما بی‌ربط با وظیفه هستند، بیش از حد تأکید کنند که منجر به یادگیری ناکارآمد می‌شود.
- **هزینه‌های محاسباتی:** محدودیت دیگر، هزینه محاسباتی مرتبط با پیاده‌سازی الگوریتم‌های یادگیری مبتنی بر کنجکاوی است. این مدل‌ها اغلب نیازمند محاسبات پیچیده برای تخمین تازگی و پاداش‌های مبتنی بر کنجکاوی هستند که می‌تواند از نظر محاسباتی گران باشد، به ویژه در فضاهای حالت با ابعاد بالا یا پیوسته. این سربار محاسباتی می‌تواند کاربرد بلادرنگ یادگیری مبتنی بر کنجکاوی را در محیط‌های پویا که تصمیم‌گیری سریع ضروری است، مانع شود.
- **سوگیری[1] اکتشاف:** علاوه بر این، مدل‌های یادگیری مبتنی بر کنجکاوی می‌توانند منجر به چیزی شوند که به عنوان «سوگیری اکتشاف» شناخته می‌شود، جایی که تمرکز عامل بر اکتشاف مبتنی بر تازگی منجر به غفلت از یادگیری مرتبط با وظیفه می‌شود. این سوگیری می‌تواند از عامل در بهره‌برداری موثر از دانش قبلی، که برای حل موثر وظایف ضروری است، جلوگیری کند. برای مثال، در RL، عامل ممکن است به طور مکرر حالت‌های جدید را کاوش کند بدون اینکه به اندازه کافی از حالت‌های شناخته‌شده‌ای که پاداش‌های بالاتری دارند بهره‌برداری کنند، و در نتیجه در بهینه‌سازی عملکرد شکست می‌خورند.

[1] Bias

شکل ۲: محدودیت‌ها، کاربردها و جهت‌گیری‌های آینده سیستم‌های یادگیری مبتنی بر کنجکاوی در یک نگاه

- **تعمیم‌پذیری:** انتقال‌پذیری و تعمیم‌پذیری استراتژی‌های یادگیری مبتنی بر کنجکاوی نیز چالش‌های قابل توجهی را ایجاد می‌کند. مدل‌های مبتنی بر کنجکاوی که در یک حوزه موثر هستند، لزوماً در حوزه دیگری به دلیل تفاوت در ساختار وظایف و ماهیت فضاهای حالت–عمل عملکرد خوبی ندارند. این محدودیت نیاز به مدل‌های سازگار دارد که بتوانند استراتژی‌های اکتشاف خود را بر اساس نیازهای خاص وظیفه و محیط به صورت پویا تنظیم کنند.
- **ادغام در سیستم‌های هوش مصنوعی:** ادغام یادگیری مبتنی بر کنجکاوی در سیستم‌های AI موجود نیازمند توجه دقیق به نحوه تراز اهداف مبتنی بر کنجکاوی با ساختارهای پاداش خارجی است. در مواردی که این اهداف با هم در تضاد هستند، یادگیری مبتنی بر کنجکاوی می‌تواند به طور ناخواسته منجر به رفتارهای زیر بهینه یا تضاد در اولویت‌بندی اکتشاف نسبت به بهره‌برداری شود. این مسئله به ویژه در محیط‌هایی که پاداش‌های خارجی کم یا ضعیف تعریف شده‌اند، آشکار است و تعادل بین اکتشاف و عملکرد وظیفه را پیچیده‌تر می‌کند.

۴-۳- کاربردها

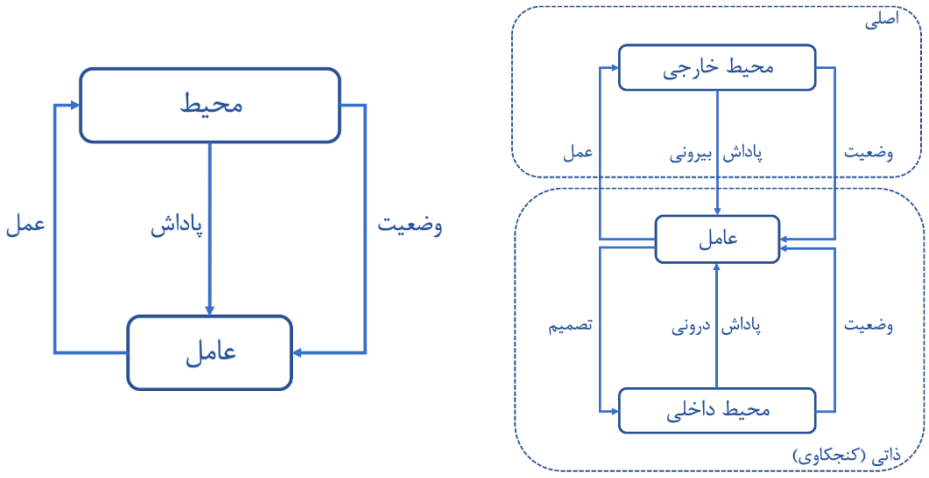
در سال‌های اخیر، همکاری علوم مهندسی و AI منجر به توسعه مدل‌های پیچیده و قابل تطبیق شده است. این پیوند به‌ویژه در زمینه‌هایی نظیر طراحی کنترل‌های تطبیقی، مدیریت داده‌ها و بهینه‌سازی ساختار مدل‌ها مشاهده شود، که در آن اصول مهندسی به‌طور مستقیم بر بهبود ساختار و دقت مدل‌های AI تأثیرگذار بوده‌اند. به عنوان نمونه، روش‌های مهندسی در مدیریت پایدار داده‌ها و طراحی شبکه‌های عصبی کاربرد گسترده‌ای دارند و کمک می‌کنند تا این مدل‌ها در محیط‌های مختلف عملکرد پایدارتری داشته باشند. می‌توان به کاربرد یادگیری ماشینی در حوزه طراحی ساختاری ساختمان و ارزیابی عملکرد اشاره کرد که برای ارتقاء قابلیت‌های طراحی سازه‌های مهندسی استفاده می‌شوند [138].

این مدل‌ها با پردازش داده‌های حجیم و پیچیده، قادر به کشف الگوهای نهفته و شناسایی ناهنجاری‌ها در ساختارها هستند که این موضوع امکان پیش‌بینی بهتر و بهینه‌سازی طراحی را فراهم می‌کند. در حوزه‌هایی مانند مهندسی شیمی و مهندسی فرآیند، مدل‌های AI برای شبیه‌سازی و بهینه‌سازی فرآیندهای شیمیایی و صنعتی به کار گرفته می‌شوند [137]. با استفاده از مدل‌های یادگیری ماشین، سیستم‌های هوشمند می‌توانند پارامترهای فرآیند را بهینه کنند و خروجی-های مطلوب را با کمترین هزینه و مصرف انرژی فراهم کنند. به این ترتیب، ترکیب اصول مهندسی و AI نه تنها منجر به بهبود فرآیندهای موجود می‌شود، بلکه به کاهش مصرف منابع، بهینه‌سازی تولید و در نتیجه به صرفه‌جویی اقتصادی نیز کمک می‌کند. در طراحی و پیاده‌سازی مدل‌های AI برای کاربردهای مهندسی، برخی از الگوریتم‌ها و مدل‌ها بیشتر از سایرین کاربرد دارند. به‌عنوان مثال، الگوریتم‌های RL برای بهینه‌سازی و کنترل سیستم‌های پیچیده مورد استفاده قرار می‌گیرند. این الگوریتم‌ها به عامل‌ها یا سیستم‌ها کمک می‌کنند تا به‌طور مستقل و با یادگیری از تعاملات خود با محیط، عملکرد خود را در شرایط مختلف بهینه کنند.این اتفاق مخصوصا در مدل‌های محاسباتی کنجکاوی رخ می‌دهد. مدل‌های کنجکاوی محاسباتی، به ویژه در شکل یادگیری مبتنی بر کنجکاوی، تأثیر قابل توجهی بر توسعه سیستم‌های AI گذاشته-اند و امکان اکتشاف موثرتر و یادگیری سازگارتر را فراهم کرده‌اند. این مدل‌ها در حوزه‌های مختلفی اعمال شده‌اند و توانایی AI در یادگیری از محیط‌هایی با پاداش‌های خلوت یا بدون پاداش صریح را افزایش داده‌اند. در ادامه به برخی از این کاربردها اشاره شده است.

- **یادگیری تقویتی:** RL یک تکنیک یادگیری ماشینی است که در آن یک عامل با تعامل با محیط یاد می‌گیرد که تصمیم‌گیری کند. همان‌طور که در شکل ۳ (سمت چپ) نشان داده شده است، در هر مرحله زمانی، عامل حالت فعلی محیط را مشاهده می‌کند، یک عمل را انتخاب می‌کند و بر اساس نتیجه پاداش دریافت می‌کند. هدف عامل یادگیری یک سیاستی است که پاداش تجمعی را در طول زمان به حداکثر برساند. در یک فرآیند MDP، حالت آینده عامل فقط به حالت فعلی و عمل انتخاب شده بستگی دارد و این باعث می‌شود که فرآیند تصمیم‌گیری ساده‌تر شود. شکل ۳ (سمت راست) نشان می‌دهد که در RL، عامل می‌تواند دو نوع پاداش دریافت کند: بیرونی و درونی. پاداش‌های بیرونی توسط محیط خارجی ارائه می‌شوند، در حالی که پاداش‌های درونی توسط کنجکاوی یا علاقه عامل به کاوش محیط تولید می‌شوند. هدف عامل، یادگیری سیاستی است که مجموع این پاداش‌ها را در طول زمان به حداکثر برساند. پیچیدگی مسائل RL می‌تواند بر اساس ماهیت فضاهای حالت و عمل متفاوت باشد، و مشاهدات بصری با ابعاد بالا مشکل را چالش‌برانگیزتر می‌کنند. تحقیقات اخیر نشان داده است که حداکثرسازی پاداش یک اصل اساسی برای توسعه عامل‌های هوشمند با قابلیت‌های گسترده مانند ادراک، درک زبان و تقلید است [51]. در RL، از مدل‌های مبتنی بر کنجکاوی برای رسیدگی به چالش سیگنال‌های پاداش خلوت استفاده شده است. با بهره‌گیری از انگیزش درونی، این مدل‌ها عامل‌ها را تشویق می‌کنند تا حالت‌های

جدید یا نامشخص را کاوش کنند و در نتیجه توانایی آن‌ها در کشف استراتژی‌های بهینه را بهبود می‌بخشند [۵۵،۶۷،۶۸،۷۷،۹۰،۹۵،۱۰۰،۱۰۹].

- **رباتیک**: در رباتیک، از یادگیری مبتنی بر کنجکاوی برای تقویت اکتشاف خودکار و اکتساب مهارت استفاده شده است. ربات‌های مجهز به مکانیسم‌های یادگیری مبتنی بر کنجکاوی می‌توانند با تمرکز بر محرک‌های جدید و یادگیری از تعاملات با محیط، محیط اطراف خود را به طور موثرتر کاوش کنند [۱۲۰–۱۱۵، ۱۰۱، ۵۶]. این رویکرد به ویژه در وظایف دستکاری رباتیک مفید بوده است، جایی که اکتشاف مبتنی بر کنجکاوی به ربات‌ها امکان می‌دهد تا بدون دخالت انسان، مهارت‌های حرکتی جدید را کشف و اصلاح کنند [12]. چنین قابلیت‌هایی برای توسعه ربات‌های خودکاری که می‌توانند با محیط‌های متنوع و پویا سازگار شوند، حیاتی هستند.
- **پردازش زبان طبیعی**: یادگیری مبتنی بر کنجکاوی همچنین در پردازش زبان طبیعی کاربردهایی پیدا کرده است، جایی که به آموزش مدل‌ها برای درک بهتر و تولید زبان کمک می‌کند. برای مثال، می‌توان از مدل‌های مبتنی بر کنجکاوی برای کاوش داده‌های زبانی گسترده و تمرکز بر ساختارهای زبانی کم‌نماینده یا پیچیده استفاده کرد و در نتیجه توانایی مدل در تولید متن منسجم‌تر و مناسب‌تر با زمینه را افزایش داد [121]. این کاربرد در بهبود ترجمه ماشینی، خلاصه‌سازی متن و سیستم‌های گفتگوی AI بسیار مفید است [122,123].



شکل ۳: تعاملات عامل–محیط در RL [51]

- **بازی**: در حوزه AI بازی، از یادگیری مبتنی بر کنجکاوی برای توسعه عامل‌هایی استفاده شده است که قادر به تسلط بر بازی‌های پیچیده بدون راهنمایی صریح انسانی هستند. با استفاده از کنجکاوی به‌عنوان یک پاداش ذاتی، عامل‌ها برای کاوش استراتژی‌ها و حالت‌های بازی متنوع انگیزه می‌یابند که منجر به کشف راه‌حل‌های نوآورانه می‌شود. به طور قابل توجه، استفاده از یادگیری مبتنی بر کنجکاوی به سیستم‌های AI اجازه داده است تا به عملکرد فوق‌انسانی در بازی‌هایی مانند Go و بازی‌های ویدیویی مختلف دست یابند، جایی که محیط گسترده است و اکتشاف حیاتی است [۴۶،۴۷،۷۳،۹۲،۱۰۵،۱۱۴،۱۲۴].

- **مراقبت‌های بهداشتی:** کاربرد یادگیری مبتنی بر کنجکاوی به حوزه مراقبت‌های بهداشتی نیز گسترش یافته است، جایی که مدل‌های AI از کنجکاوی برای شناسایی الگوها و ناهنجاری‌ها در داده‌های پزشکی استفاده می‌کنند. این رویکرد می‌تواند تشخیص زودهنگام و برنامه‌ریزی درمان را با تشویق به کاوش موارد نادر یا غیرمعمول که ممکن است توسط مدل‌های سنتی نادیده گرفته شوند، تسهیل کند. با تمرکز بر داده‌های جدید بیماران، سیستم‌های AI مبتنی بر کنجکاوی می‌توانند بینش‌های ارزشمندی در مورد پیشرفت بیماری و اثربخشی درمان ارائه دهند و در نتیجه نتایج بیماران را بهبود بخشند [129,128].
- **خودروهای خودران:** در حوزه خودروهای خودران، یادگیری مبتنی بر کنجکاوی به ناوبری و تصمیم‌گیری کمک می‌کند و اکتشاف سناریوهای رانندگی متنوع را تقویت می‌کند. مدل‌های مبتنی بر کنجکاوی به خودروها امکان می‌دهند تا از رویدادهای غیرمنتظره بیاموزند و با شرایط ترافیکی متغیر سازگار شوند و توانایی آن‌ها را برای عملکرد ایمن و کارآمد در محیط‌های دنیای واقعی افزایش دهند [130,136]. این کاربرد برای توسعه سیستم‌های رانندگی خودران قوی که قادر به مدیریت پیچیدگی‌های جاده‌های مدرن هستند، حیاتی است.

به طور کلی، کاربرد مدل‌های کنجکاوی محاسباتی در AI پتانسیل قابل توجهی در بهبود کارایی یادگیری و سازگاری در حوزه‌های مختلف نشان داده است. با تشویق اکتشاف و کشف، یادگیری مبتنی بر کنجکاوی توسعه سیستم‌های هوشمند قادر به یادگیری از نظارت کم و عملکرد در محیط‌های پیچیده و پویا را تقویت می‌کند.

۴-۴- جهت‌گیری‌های آینده

حوزه یادگیری مبتنی بر کنجکاوی در AI به سرعت در حال تکامل است و فرصت‌های متعددی برای پیشبرد مرزهای دانش ارائه می‌دهد. با افزایش تقاضا برای سیستم‌های خودمختار و هوشمندتر، یادگیری مبتنی بر کنجکاوی مسیرهای امیدوارکننده‌ای را برای توسعه AI که قادر به یادگیری و سازگاری در محیط‌های پیچیده است، ارائه می‌دهد. این بخش چندین جهت‌گیری آینده برای یادگیری مبتنی بر کنجکاوی در AI را بررسی می‌کند و حوزه‌های بالقوه تحقیق و کاربرد را برای بهبود بیشتر اثربخشی و قابلیت کاربرد آن برجسته می‌کند.

- **بهبود تعمیم‌پذیری و انتقال یادگیری:** یکی از چالش‌های مهم در یادگیری مبتنی بر کنجکاوی، بهبود قابلیت‌های تعمیم‌پذیری مدل‌های مبتنی بر کنجکاوی است. تحقیقات آینده باید بر توسعه الگوریتم‌هایی تمرکز کنند که بتوانند به طور موثر دانش را بین وظایف و حوزه‌های مختلف انتقال دهند و در نتیجه نیاز به آموزش خاص برای هر وظیفه را کاهش دهند. این کار را می‌توان با بررسی رویکردهای متا-یادگیری انجام داد که به مدل‌ها امکان می‌دهد رفتارهای مبتنی بر کنجکاوی را در محیط‌های مختلف تعمیم دهند و منجر به سیستم‌های AI قوی‌تر و متنوع‌تر شوند. بهبود تعمیم‌پذیری برای استقرار یادگیری مبتنی بر کنجکاوی در کاربردهای دنیای واقعی که

محیط‌ها در آن پویا و غیرقابل پیش‌بینی هستند، حیاتی خواهد بود.

- **ادغام کنجکاوی انسان‌مانند:** ادغام جنبه‌های کنجکاوی انسان‌مانند در سیستم‌های AI جهت‌گیری امیدوارکننده دیگری برای یادگیری مبتنی بر کنجکاوی است. کنجکاوی انسان نه تنها توسط تازگی بلکه توسط اهداف ذاتی و برنامه‌ریزی بلندمدت نیز هدایت می‌شود. تحقیقات آینده می‌توانند بررسی کنند که چگونه می‌توان این عناصر را در مدل‌های یادگیری مبتنی بر کنجکاوی ادغام کرد تا به سیستم‌های AI اجازه دهد اهداف ذاتی را دنبال کنند و در رفتار اکتشافی بلندمدت شرکت کنند. این می‌تواند شامل توسعه چارچوب‌های سلسله مراتبی مبتنی بر کنجکاوی باشد که به عامل‌های AI امکان می‌دهد تعادل بین اکتشاف و بهره‌برداری را به طور موثرتر برقرار کنند، مشابه نحوه برخورد انسان‌ها با وظایف پیچیده حل مسئله.
- **ترکیب کنجکاوی با انگیزه‌های دیگر:** در حالی که کنجکاوی محرک قدرتمندی برای اکتشاف است، ترکیب آن با عوامل انگیزشی دیگر مانند تعامل اجتماعی یا یادگیری مشارکتی می‌تواند منجر به سیستم‌های AI موثرتر شود. تحقیقات آینده می‌توانند بررسی کنند که چگونه مدل‌های مبتنی بر کنجکاوی را می‌توان با الگوریتم‌های یادگیری اجتماعی ادغام کرد تا رفتار همکاری بین چندین عامل را تسهیل کند. این رویکرد می‌تواند در محیط‌هایی که همکاری و ارتباط برای تکمیل وظیفه ضروری است، مانند سیستم‌های چندعاملی و رباتیک، به ویژه مفید باشد.
- **رسیدگی به نگرانی‌های اخلاقی و ایمنی:** با خودمختارتر شدن سیستم‌های یادگیری مبتنی بر کنجکاوی و توانایی آن‌ها در کاوش محیط‌های ناشناخته، رسیدگی به نگرانی‌های اخلاقی و ایمنی اهمیت فزایندهای پیدا می‌کند. تحقیقات آینده باید بر توسعه تکنیک‌های اکتشاف ایمن تمرکز کنند که تضمین می‌کنند سیستم‌های AI در محدوده‌های اخلاقی از پیش تعریف شده عمل می‌کنند و از رفتارهای مضر اجتناب می‌کنند. این می‌تواند شامل طراحی چارچوب‌های نظارتی و دستورالعمل‌هایی باشد که استقرار سیستم‌های یادگیری مبتنی بر کنجکاوی را در کاربردهای حیاتی مانند مراقبت‌های بهداشتی و خودروهای خودران تنظیم می‌کند. اطمینان از پایبندی سیستم‌های مبتنی بر کنجکاوی به اصول اخلاقی برای جلب اعتماد و پذیرش عمومی ضروری است.
- **بهره‌برداری از پیشرفت‌های علوم اعصاب:** پیشرفت‌های علوم اعصاب بینش‌های ارزشمندی در مورد مکانیسم‌های زیربنایی کنجکاوی و اکتشاف انسانی ارائه می‌دهند. تحقیقات آینده می‌توانند از این بینش‌ها برای اطلاعرسانی توسعه مدلهای یادگیری مبتنی بر کنجکاوی پیچیدهتر استفاده کنند. با درک نحوه پردازش تازگی و کنجکاوی توسط مغز، محققان می‌توانند سیستم‌های AI را طراحی کنند که این فرآیندها را تقلید کرده و منجر به استراتژی‌های اکتشاف طبیعی‌تر و کارآمدتر شوند. همکاری بین محققان AI و دانشمندان اعصاب می‌تواند ایجاد مدل‌های الهام گرفته از زیست‌شناسی را تسهیل کند که قابلیت‌های یادگیری سیستم‌های AI مبتنی بر کنجکاوی را افزایش می‌دهد.
- **گسترش کاربردها به حوزه‌های جدید:** در حالی که یادگیری مبتنی بر کنجکاوی در زمینه‌هایی مانند رباتیک و RL نویدبخش بوده است، کاربرد آن در حوزه‌های جدید همچنان یک حوزه هیجان‌انگیز برای اکتشافات آینده

است. کاربردهای بالقوه شامل آموزش شخصی‌سازی شده است که در آن مدل‌های مبتنی بر کنجکاوی می‌توانند تجربیات یادگیری را بر اساس ترجیحات فردی تنظیم کنند، و صنایع خلاق، جایی که سیستم‌های AI می‌توانند ایده‌ها و راه‌حل‌های جدید تولید کنند. گسترش دامنه یادگیری مبتنی بر کنجکاوی به این حوزه‌ها و حوزه‌های دیگر نیازمند رویکردهای نوآورانه‌ای است که از نقاط قوت منحصر به فرد اکتشاف مبتنی بر کنجکاوی بهره‌برداری کنند.

- **توسعه مدل‌های کنجکاوی چندعاملی:** بررسی می‌کند چگونه چندین عامل کنجکاو می‌توانند با یکدیگر تعامل داشته و یادگیری و اکتشاف را به صورت مشترک پیش ببرند. این رویکرد، در پی بهره‌برداری از تعاملات اجتماعی و همکاری بین عامل‌ها است تا بهره‌وری یادگیری و توانایی اکتشاف را بهبود بخشد. برخی چالش‌های پیش رو در این حوزه عبارتند از: هماهنگی بین عامل‌ها، تعاملات اجتماعی و تقسیم وظایف.

۵- نتیجه‌گیری

ما در این مقاله به بررسی جامع و دقیق نقش کنجکاوی در AI و اهمیت آن به عنوان محرک اصلی برای یادگیری فعال و اکتشاف پرداخته‌ایم. در ادامه مفهوم کنجکاوی را از دیدگاه روانشناسی و رابطه آن با انگیزش درونی را مورد بررسی قرار دادیم. همچنین متغیرهای هم‌سنجش همچون تازگی، تضاد، عدم قطعیت و پیچیدگی را به‌عنوان محرک‌های اصلی کنجکاوی در سیستم‌های AI معرفی و بررسی کرده‌ایم. از طرفی، معرفی متغیر هم‌سنجش جدیدی به نام «پوشش فضا» به منظور بهبود مدل‌سازی کنجکاوی محاسباتی، نشان داد که می‌توان به شکلی کارآمدتر به گسترش قابلیت‌های اکتشافی سیستم‌های AI پرداخت. با تحلیل و دسته‌بندی مدل‌های محاسباتی موجود که بر اساس این متغیرها طراحی شده‌اند، به درک بهتری از مزایا و چالش‌های این رویکردها دست یافته‌ایم. این تحلیل‌ها همچنین ما را به بررسی معیارهای پاداش و تأثیر آن‌ها بر بهبود فرآیند یادگیری و اکتشاف در سیستم‌های AI رهنمون کرد. با استفاده از این پژوهش‌ها، می‌توان راهکارهای مؤثرتری برای توسعه سیستم‌های هوشمند ارائه داد که قادر به تعامل طبیعی‌تر با انسان‌ها و محیط‌های واقعی باشند. این سیستم‌ها می‌توانند به شکل‌گیری رفتارهای خلاقانه و نوآورانه کمک کنند و در حل مسائل پیچیده و ناشناخته مؤثر واقع شوند. پیشنهادات این مقاله برای پژوهش‌های آینده، شامل ارتقای مدل‌های محاسباتی کنجکاوی و گسترش کاربردهای آن در زمینه‌های مختلف علمی و صنعتی می‌باشد. با توجه به روند رو به رشد فناوری‌های AI و نیاز به سیستم‌هایی که به‌طور خودکار و مستقل به اکتشاف و یادگیری بپردازند، تقویت مکانیزم-های کنجکاوی محاسباتی می‌تواند به شکل قابل‌توجهی به پیشرفت در این حوزه کمک کند. این پژوهش راه را برای توسعه ابزارها و الگوریتم‌های پیشرفته‌تر در جهت افزایش بهره‌وری و کارایی سیستم‌های AI هموار می‌سازد.

منابع

[1] R. Archana, P. Jeevaraj. "Deep learning models for digital image processing: a review," in Artificial Intelligence Review, vol. 57, no. 1, pp. 11, 2024.

[2] Mehrish, A., et al. "A review of deep learning techniques for speech processing," in Information Fusion, vol. 99, pp. 101869, 2023.

[3] M. Soori, B. Arezoo, R. Dastres. "Artificial intelligence, machine learning and deep learning in advanced robotics, a review," in Cognitive Robotics, vol. 3, pp. 54–70, 2023.

[4] Badue, C., et al. "Self-driving cars: A survey," in Expert systems with applications, vol. 165, pp. 113816, 2021.

[5] Yin, H., et al. "On-device recommender systems: A comprehensive survey," in arXiv preprint arXiv:2401.11441, 2024.

[6] Stray, J., et al. "Building human values into recommender systems: An interdisciplinary synthesis," in ACM Transactions on Recommender Systems, vol. 2, no. 3, pp. 1–57, 2024.

[7] T. Kashdan, M. Steger. "Curiosity and pathways to well-being and meaning in life: Traits, states, and everyday behaviors," in Motivation and Emotion, vol. 31, pp. 159–173, 2007.

[8] Sun, C., Qian, H., & Miao, C. (2022). From psychological curiosity to artificial curiosity: Curiosity-driven learning in artificial intelligence tasks. *arXiv preprint arXiv:2201.08300*.

[9] George Loewenstein. 1994. The psychology of curiosity: A review and reinterpretation. Psychological bulletin 116, 1 (1994), 75.

[10] J. Schmidhuber. "Developmental robotics, optimal artificial curiosity, creativity, music, and the fine arts," in Connection Science, vol. 18, no. 2, pp. 173–187, 2006.

[11] Wu, Q., & Miao, C. (2013). Curiosity: From psychology to computation. *ACM Computing Surveys (CSUR)*, *46*(2), 1-26.

[12] P. Oudeyer, F. Kaplan. "What is intrinsic motivation? A typology of computational approaches," in Frontiers in neurorobotics, vol. 1, pp. 108, 2007.

[13] D.E. Berlyne. 1960. *Conflict, arousal, and curiosity*. McGraw-Hill New York.[1]

[14] R. Saunders, J. Gero, "The digital clockwork muse: A computational model of aesthetic evolution," in Proceedings of the AISB, 2001, pp. 12–21.

[15] S. Reichhuber, S. Tomforde. "Active Reinforcement Learning–A Roadmap Towards Curious Classifier Systems for Self-Adaptation," in arXiv preprint arXiv:2201.03947, 2022.

[16] Z. Fu, X. Niu. "Modeling Users' Curiosity in Recommender Systems," in ACM Transactions on Knowledge Discovery from Data, vol. 18, no. 1, pp. 1–23, 2023.

[17] Schaul, T., et al, "Curiosity-driven optimization," in 2011 IEEE Congress of Evolutionary Computation (CEC), 2011, pp. 1343–1349.

[18] R. Zhao, V. Tresp. "Curiosity-driven experience prioritization via density estimation," in arXiv preprint arXiv:1902.08039, 2019.

[19] T. Blau, L. Ott, F. Ramos. "Bayesian curiosity for efficient exploration in reinforcement learning," in arXiv preprint arXiv:1911.08701, 2019.

[20] D. Rezende, S. Mohamed, "Variational inference with normalizing flows," in International conference on machine learning, 2015, pp. 1530–1538.

[21] J. Schmidhuber, "Curious model-building control systems," in Proc. international joint conference on neural networks, 1991, pp. 1458–1463.

[22] P. Oudeyer, F. Kaplan, "How can we define intrinsic motivation?," in the 8th international conference on epigenetic robotics: Modeling cognitive development in robotic systems, 2008.

[23] Dobrynin, D., et al. "Physical and biological mechanisms of direct plasma interaction with living tissue," in New Journal of Physics, vol. 11, no. 11, pp. 115020, 2009.

[24] Jepma, M., et al. "Neural mechanisms underlying the induction and relief of perceptual curiosity," in Frontiers in behavioral neuroscience, vol. 6, pp. 5, 2012.

[25] Todd B Kashdan and Paul J Silvia. 2009. Curiosity and interest: The benefits of thriving on novelty and challenge. Oxford handbook of positive psychology 2 (2009), 367–374.

[26] Modirshanechi, A., et al. "Curiosity-driven exploration: foundations in neuroscience and computational modeling," in Trends in Neurosciences, 2023.

[27] G Stanley Hall and Theodate L Smith. 1903. Curiosity and interest. The Pedagogical Seminary 10, 3 (1903), 315–358.

[28] Daniel E Berlyne. 1950. Novelty and curiosity as determinants of exploratory behaviour. British Journal of Psychology 41, 1 (1950), 68.

[29] Abraham Harold Maslow. 1943. A theory of human motivation. Psychological review 50, 4 (1943), 370.

[30] Konrad Z Lorenz. 1981. Exploratory behavior or curiosity. In the Foundations of Ethology. Springer, 325–335.

[31] Donald O Hebb. 1946. On the nature of fear. Psychological review 53, 5 (1946), 259.

[32] Jean Piaget. 2003. The psychology of intelligence. Routledge.

[33] Robert W White. 1959. Motivation reconsidered: The concept of competence. Psychological review 66, 5 (1959), 297.

[34] Edward L Deci and Richard M Ryan. 2010. Intrinsic motivation. The corsini encyclopedia of psychology (2010), 1–2.

[35] William N Dember and Robert W Earl. 1957. Analysis of exploratory, manipulatory, and curiosity behaviors. Psychological review 64, 2 (1957), 91.

[36] C.D. Spielberger and L.M. Starr. 1994. *Curiosity and exploratory behavior.* NJ: Lawrence Erlbaum Associates, 221–243.

[37] F.F. Schmitt and R. Lahroodi. 2008. The epistemic value of curiosity. *Educational Theory* 58, 2 (2008), 125–148.

[38] Deci, E. L., & Ryan, R. M. (2000). Self-Determination Theory: Theoretical issues and practical applications. Rochester: University of Rochester Press.

[39] Jirout, J. J., & Klahr, D. (2012). Children's scientific curiosity: In search of an operational definition of an elusive concept. Developmental Review, 32(2), 125-160.

[40] Gottfried, A. E. (1990). Academic intrinsic motivation in young elementary school children. Journal of Educational Psychology, 82(3), 525-538.

[41] Gruber, M. J., Gelman, B. D., & Ranganath, C. (2014). States of curiosity modulate hippocampus-dependent learning via the dopaminergic circuit. *Neuron*, *84*(2), 486-496.

[42] Cantor, G. N., Cantor, J. H., & Ditrichs, R. (1963). Observing behavior in preschool children as a function of stimulus complexity. *Child Development*, 683-689.

[43] Daniel E Berlyne. 1978. Curiosity and learning. Motivation and emotion 2, 2 (1978), 97–175.

[44] A. Ten, P. Oudeyer, C. Moulin-Frier. "Curiosity-driven exploration," in The Drive for Knowledge: The Science of Human Information Seeking, pp. 53, 2022.

[45] Paul J Silvia. 2005. Cognitive appraisals and interest in visual art: Exploring an appraisal theory of aesthetic emotions. Empirical studies of the arts 23, 2 (2005), 119–133.

[46] Pathak, D., Agrawal, P., Efros, A. A., & Darrell, T. (2017). Curiosity-driven exploration by self-supervised prediction. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2017.

[47] Burda, Y., Edwards, H., Storkey, A., & Klimov, O. (2018). Exploration by random network distillation. arXiv preprint arXiv:1810.12894.

[48] O. Nahirnyi. "Reinforcement Learning Agents in Procedurally-generated Environments with Sparse Rewards,", 2022.

[49] Macedo, L., & Cardoso, A. (1999, January). Towards artificial forms of surprise and curiosity. In Proceedings of the European Conference on Cognitive Science, S. Bagnara, Ed (pp. 139-144).

[50] Saunders, R., & Gero, J. S. (2001). A curious design agent. In CAADRIA (Vol. 1, pp. 345-350).

[51] C. Sun. "Curiosity-driven learning in artificial intelligence and its applications," 2023.

[52] Yadav, P., et al. "OODA-RL: A reinforcement learning framework for artificial general intelligence to solve open world novelty," in Authorea Preprints, 2023.

[53] M. Kubovč\ik, I. Dirgová Luptáková, J. Posp\'\ichal. "Signal Novelty Detection as an Intrinsic Reward for Robotics," in Sensors, vol. 23, no. 8, pp. 3985, 2023.

[54] Bellemare, M., et al. "Unifying count-based exploration and intrinsic motivation," in Advances in neural information processing systems, vol. 29, 2016.

[55] Tang, H., et al. "# exploration: A study of count-based exploration for deep reinforcement learning," in Advances in neural information processing systems, vol. 30, 2017.

[56] Han, C., et al. "Learning robotic manipulation skills with multiple semantic goals by conservative curiosity-motivated exploration," in Frontiers in Neurorobotics, vol. 17, pp. 1089270, 2023.

[57] Savinov, N., et al. "Episodic curiosity through reachability," in arXiv preprint arXiv:1810.02274, 2018.

[58] Kim, Y., et al, "Curiosity-bottleneck: Exploration by distilling task-specific novelty," in International conference on machine learning, 2019, pp. 3379–3388.

[59] Alemi, A., et al. "Deep variational information bottleneck," in arXiv preprint arXiv:1612.00410, 2016.

[60] 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.

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

[62] H. Jiang, Z. Ding, Z. Lu, "Settling Decentralized Multi-Agent Coordinated Exploration by Novelty Sharing," in Proceedings of the AAAI Conference on Artificial Intelligence, 2024, pp. 17444–17452.

[63] Sun, C., Qian, H., & Miao, C. (2022). From psychological curiosity to artificial curiosity: Curiosity-driven learning in artificial intelligence tasks. arXiv preprint arXiv:2201.08300.

[64] Karaoguz, C., et al, "Curiosity driven exploration of sensory-motor mappings," in Capo Caccia Cognitive Neuromorphic Engineering Workshop, 2011.

[65] R. Raileanu, T. Rocktäschel. "Ride: Rewarding impact-driven exploration for procedurally-generated environments," in arXiv preprint arXiv:2002.12292, 2020.

[66] Parisi, S., et al. "Interesting object, curious agent: Learning task-agnostic exploration," in Advances in Neural Information Processing Systems, vol. 34, pp. 20516–20530, 2021.

[67] Yuan, M., et al. "Rewarding episodic visitation discrepancy for exploration in reinforcement learning," in arXiv preprint arXiv:2209.08842, 2022.

[68] Wang, Y., et al. "Efficient potential-based exploration in reinforcement learning using inverse dynamic bisimulation metric," in Advances in Neural Information Processing Systems, vol. 36, 2024.

[69] Q. Wu, C. Miao, Z. Shen, "A curious learning companion in virtual learning environment," in 2012 IEEE International Conference on Fuzzy Systems, 2012, pp. 1–8.

[70] Q. Wu, S. Liu, C. Miao, "Recommend interesting items: How can social curiosity help?," in Web Intelligence, 2019, pp. 297–311.

[71] P. Oudeyer. "Intelligent adaptive curiosity: a source of self-development," ,2004.

[72] Balloch, J., et al, "Characteristics of Effective Exploration for Transfer in Reinforcement Learning," in Finding the Frame: An RLC Workshop for Examining Conceptual Frameworks, 2024.

[73] Jiang, R., et al, "OB-HPPO: An Option and Intrinsic Curiosity Based Hierarchical Reinforcement Learning Approach for Real-Time Strategy Games," in International Conference on Intelligent Computing, 2024, pp. 443–454.

[74] Baker, B., et al. "Emergent tool use from multi-agent autocurricula," in arXiv preprint arXiv:1909.07528, 2019.

[75] Campero, A., et al. "Learning with amigo: Adversarially motivated intrinsic goals," in arXiv preprint arXiv:2006.12122, 2020.

[76] Parker-Holder, J., et al, "Evolving curricula with regret-based environment design," in

International Conference on Machine Learning, 2022, pp. 17473–17498.

[77] Zhou, X., et al. "MENTOR: Guiding Hierarchical Reinforcement Learning with Human Feedback and Dynamic Distance Constraint," in arXiv preprint arXiv:2402.14244, 2024.

[78] L. Macedo, A. Cardoso, "The role of surprise, curiosity and hunger on exploration of unknown environments populated with entities," in 2005 portuguese conference on artificial intelligence, 2005, pp. 47–53.

[79] Houthooft, R., et al. "Vime: Variational information maximizing exploration," in Advances in neural information processing systems, vol. 29, 2016.

[80] Ten, A., et al. "The curious U: Integrating theories linking knowledge and information-seeking behavior," ,2024.

[81] Berseth, G., et al. "Smirl: Surprise minimizing reinforcement learning in unstable environments," in arXiv preprint arXiv:1912.05510, 2019.

[82] Jansonnie, P., et al. "Unsupervised Skill Discovery for Robotic Manipulation through Automatic Task Generation," in arXiv preprint arXiv:2410.04855, 2024.

[83] Fickinger, A., et al. "Explore and control with adversarial surprise," in arXiv preprint arXiv:2107.07394, 2021.

[84] Haarnoja, T., et al. "Soft actor-critic algorithms and applications," in arXiv preprint arXiv:1812.05905, 2018.

[85] O'Donoghue, B., et al, "The uncertainty bellman equation and exploration," in International conference on machine learning, 2018, pp. 3836–3845.

[86] Han, Z., et al. "Reweighted mixup for subpopulation shift," in arXiv preprint arXiv:2304.04148, 2023.

[87] Lin, J., et al. "Cat-sac: Soft actor-critic with curiosity-aware entropy temperature," 2020.

[88] Li, K., et al, "Mural: Meta-learning uncertainty-aware rewards for outcome-driven reinforcement learning," in International conference on machine learning, 2021, pp. 6346–6356.

[89] D. Cho, S. Lee, H. Kim. "Outcome-directed reinforcement learning by uncertainty & temporal distance-aware curriculum goal generation," in arXiv preprint arXiv:2301.11741, 2023.

[90] Lee, S., et al. "CQM: curriculum reinforcement learning with a quantized world model," in Advances in Neural Information Processing Systems, vol. 36, 2024.

[91] A. Barto, M. Mirolli, G. Baldassarre. "Novelty or surprise?," in Frontiers in psychology, vol. 4, pp. 907, 2013.

[92] J. Schmidhuber. "Formal theory of creativity, fun, and intrinsic motivation (1990–2010)," in IEEE transactions on autonomous mental development, vol. 2, no. 3, pp. 230–247, 2010.

[93] Storck, J., et al, "Reinforcement driven information acquisition in non-deterministic environments," in Proceedings of the international conference on artificial neural networks, Paris, 1995, pp. 159–164.

[94] J. Schmidhuber, "Adaptive confidence and adaptive curiosity," Citeseer, Tech. Rep., 1991.

[95] S. Bohara, M. Hanif, M. Shafique, "CuriousRL: Curiosity-Driven Reinforcement Learning for Adaptive Locomotion in Quadruped Robots," in 2024 International Joint Conference on Neural Networks (IJCNN), 2024, pp. 1–8.

[96] Barto, A., et al, "Intrinsically motivated learning of hierarchical collections of skills," in Proceedings of the 3rd International Conference on Development and Learning, 2004, pp. 19.

[97] N. Chentanez, A. Barto, S. Singh. "Intrinsically motivated reinforcement learning," in Advances in neural information processing systems, vol. 17, 2004.

[98] Sekar, R., et al, "Planning to explore via self-supervised world models," in International conference on machine learning, 2020, pp. 8583–8592.

[99] Nguyen, T., et al, "Sample-efficient reinforcement learning representation learning with curiosity contrastive forward dynamics model," in 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2021, pp. 3471–3477.

[100] Poli, F., et al. "Curiosity and the dynamics of optimal exploration," in Trends in Cognitive Sciences, vol. 28, no. 5, pp. 441–453, 2024.

[101] Schwarke, C., et al, "Curiosity-driven learning of joint locomotion and manipulation tasks," in Proceedings of The 7th Conference on Robot Learning, 2023, pp. 2594–2610.

[102] Zheng, L., et al. "Episodic multi-agent reinforcement learning with curiosity-driven exploration," in Advances in Neural Information Processing Systems, vol. 34, pp. 3757–3769, 2021.

[103] Mazzaglia, P., et al, "Self-supervised exploration via latent Bayesian surprise," in ICLR2021, the 9th International Conference on Learning Representations, 2021.

[104] Mazzaglia, P., et al, "Curiosity-driven exploration via latent bayesian surprise," in Proceedings of the AAAI conference on artificial intelligence, 2022, pp. 7752–7760.

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

[106] Amin, S., et al. "A survey of exploration methods in reinforcement learning," in arXiv

preprint arXiv:2109.00157, 2021.

[107] M. Machado, M. Bellemare, M. Bowling, "A laplacian framework for option discovery in reinforcement learning," in International Conference on Machine Learning, 2017, pp. 2295–2304.

[108] Machado, M., et al. "Eigenoption discovery through the deep successor representation," in arXiv preprint arXiv:1710.11089, 2017.

[109] P. Xu, J. Zhang, K. Huang, "Population-based diverse exploration for sparse-reward multi-agent tasks," in Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, 2024, pp. 283–291.

[110] Jinnai, Y., et al, "Discovering options for exploration by minimizing cover time," in International Conference on Machine Learning, 2019, pp. 3130–3139.

[111] P. Amortila, D. Foster, A. Krishnamurthy. "Scalable Online Exploration via Coverability," in arXiv preprint arXiv:2403.06571, 2024.

[112] Jinnai, Y., et al, "Exploration in reinforcement learning with deep covering options," in International Conference on Learning Representations, 2020.

[113] Amin, S., et al. "Locally persistent exploration in continuous control tasks with sparse rewards," in arXiv preprint arXiv:2012.13658, 2020.

[114] Sabbioni, L., et al, "Simultaneously updating all persistence values in reinforcement learning," in Proceedings of the AAAI Conference on Artificial Intelligence, 2023, pp. 9668–9676.

[115] Hartikainen, K., et al. "Dynamical distance learning for semi-supervised and unsupervised skill discovery," in arXiv preprint arXiv:1907.08225, 2019.

[116] F. Stulp, O. Sigaud. "Robot skill learning: From reinforcement learning to evolution strategies," in Paladyn, Journal of Behavioral Robotics, vol. 4, no. 1, pp. 49–61, 2013.

[117] S. Nguyen, P. Oudeyer. "Socially guided intrinsic motivation for robot learning of motor skills," in Autonomous Robots, vol. 36, pp. 273–294, 2014.

[118] G. Gordon. "Infant-inspired intrinsically motivated curious robots," in Current Opinion in Behavioral Sciences, vol. 35, pp. 28–34, 2020.

[119] Zeng, H., et al. "AHEGC: Adaptive Hindsight Experience Replay With Goal-Amended Curiosity Module for Robot Control," in IEEE Transactions on Neural Networks and Learning Systems, 2023.

[120] Wang, T., et al. "Curiosity model policy optimization for robotic manipulator tracking control with input saturation in uncertain environment," in Frontiers in Neurorobotics, vol. 18, pp. 1376215, 2024.

[121] Luo, Y., et al, "Curiosity-driven reinforcement learning for diverse visual paragraph generation," in Proceedings of the 27th ACM International Conference on Multimedia, 2019, pp. 2341–2350.

[122] Colas, C., et al. "Language as a cognitive tool to imagine goals in curiosity driven exploration," in Advances in Neural Information Processing Systems, vol. 33, pp. 3761–3774, 2020.

[123] Hong, Z., et al. "Curiosity-driven red-teaming for large language models," in arXiv preprint arXiv:2402.19464, 2024.

[124] Roohi, S., et al, "Review of intrinsic motivation in simulation-based game testing," in Proceedings of the 2018 chi conference on human factors in computing systems, 2018, pp. 1–13.

[125] Esteva, A., et al. "Dermatologist-level classification of skin cancer with deep neural networks," in nature, vol. 542, no. 7639, pp. 115–118, 2017.

[126] Niu, X., et al, "Surprise me if you can: Serendipity in health information," in Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, 2018, pp. 1–12.

[127] Z. Fu, X. Niu, M. Maher. "Deep learning models for serendipity recommendations: a survey and new perspectives," in ACM Computing Surveys, vol. 56, no. 1, pp. 1–26, 2023.

[128] Song, S., et al. "Serious information in hedonic social applications: affordances, self-determination and health information adoption in TikTok," in Journal of Documentation, vol. 78, no. 4, pp. 890–911, 2022.

[129] T. Hasan, R. Bunescu, "Topic-Level Bayesian Surprise and Serendipity for Recommender Systems," in Proceedings of the 17th ACM Conference on Recommender Systems, 2023, pp. 933–939.

[130] Codevilla, F., et al, "Exploring the limitations of behavior cloning for autonomous driving," in Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 9329–9338.

[131] M. Albilani, A. Bouzeghoub, "Dynamic Adjustment of Reward Function for Proximal Policy Optimization with Imitation Learning: Application to Automated Parking Systems," in 2022 IEEE Intelligent Vehicles Symposium (IV), 2022, pp. 1400–1408.

[132] F. Carton, "Exploration of reinforcement learning algorithms for autonomous vehicle visual perception and control," Ph.D. dissertation, Institut Polytechnique de Paris, 2021.

[133] M. Hutsebaut-Buysse, "Learning to navigate through abstraction and adaptation," Ph.D. dissertation, University of Antwerp, 2023.

[134] Huang, C., et al. "Deductive reinforcement learning for visual autonomous urban driving navigation," in IEEE Transactions on Neural Networks and Learning Systems, vol. 32, no. 12, pp. 5379–5391, 2021.

[135] Wu, Y., et al. "Deep reinforcement learning on autonomous driving policy with auxiliary critic network," in IEEE transactions on neural networks and learning systems, vol. 34, no. 7, pp. 3680–3690, 2021.

[136] Yan, Y., et al, "An improved proximal policy optimization algorithm for autonomous driving decision-making," in Fourth International Conference on Sensors and Information Technology (ICSI 2024), 2024, pp. 837–845.

[137] L. Bustillo, T. Laino, T. Rodrigues. "The rise of automated curiosity-driven discoveries in chemistry," in Chemical Science, vol. 14, no. 38, pp. 10378–10384, 2023.

[138] H. Sun, H. Burton, H. Huang. "Machine learning applications for building structural design and performance assessment: State-of-the-art review," in Journal of Building Engineering, vol. 33, pp. 101816, 2021.