Developing a Reinforcement Learning Algorithm to Model Pavlovian Approach Bias on Bidirectional Planning

Reza Kakooee[1], Mohammad Taghi Hamidi Beheshti[1]*, Mehdi Keramati[2]

[1]Department of Control, Faculty of Electrical and Computer Engineering, Tarbiat Modares University, Tehran, Iran
[2]Department of Psychology, School of Social Sciences and Arts, University of London, London, England

Article Info:

Received: **24 July 2021** ***Revised:*** **27 Oct 2021** ***Accepted*****: 15 Nov 2021**

ABSTRACT

Introduction: The decision- making process in the human brain is controlled by two mechanisms: Pavlovian and instrumental learning systems. The Pavlovian system learns the stimulus- outcome association independent of action; a process that manifests itself in the tendency to approach reward- associated stimuli. The instrumental controller, on the other hand, learns the action- outcome association. Instrumental learning is not limited to the current action's outcome and may evaluate a sequence of future actions in the form of forward planning. Nonetheless, forward planning may not be the only planning process used by instrumental learning. Humans may also use backward planning to evaluate actions sequences. However, backward planning has received less attention so far. Previous research has shown that despite the independence of Pavlovian and instrumental learning, they interact with each other such that the Pavlovian approach tendency biases forward planning, causing it to make decisions that may not be optimal actions from the instrumental learning perspective. Nevertheless, the effect of Pavlovian learning on backward planning has not yet been studied. ***Materials and Methods:*** This paper designs a navigation experiment that allows investigating forward, backward, and bidirectional planning. Moreover, we embed Pavlovian approach cues into the maps to investigate how they bias the three forms of planning. ***Results:*** Statistical analysis of the collected data indicates the existence of backward planning and shows that the Pavlovian approach cues bias the planning. This bias is stronger in forward planning compared to backward planning and is even stronger in bidirectional planning. In the context of reinforcement learning, we developed a bidirectional planning algorithm under the Pavlovian approach tendency. ***Conclusion:*** The simulation results are consistent with the experimental results and indicate that the effect of Pavlovian bias can be modeled as pruning of decision trees.

Keywords:

1. Decision Making
2. Strategic Planning
3. Conditioning, Operant
4. Computer Simulation

***Corresponding Author:** Mohammad Taghi Hamidi Beheshti

Email: *mbehesht@modares.ac.ir*

[DOI: 10.52547/shefa.9.4.51]

توسعهٔ الگوریتم یادگیری تقویتی برای مدل کردن اثر ایمای پاولفی روی برنامه‌ریزی دوجهته

رضا کاکویی[۱]، محمد تقی حمیدی بهشتی[۱*]، مهدی کرامتی[۲]

[۱]گروه کنترل، دانشکده برق و کامپیوتر، دانشگاه تربیت مدرس، تهران، ایران
[۲]گروه روانشناسی، دانشکده علوم اجتماعی و هنر، دانشگاه لندن، لندن، انگلستان

اطلاعات مقاله:
دریافت: ۲ مرداد ۱۴۰۰ اصلاحیه: ۵ آبان ۱۴۰۰ پذیرش: ۲۴ آبان ۱۴۰۰

چـــــکیده

مقدمه: فرآیند تصمیم‌گیری در مغز انسان توسط دو سازوکار یادگیری پاولفی و ابزاری کنترل می‌شود. یادگیری پاولفی با آموختن پیوند محرک- نتیجه به یادگیری منجر می‌شود بدون آن‌که به عمل انتخابی وابسته باشد. همچنین این یادگیری به صورت تمایل به نزدیک شدن به محرک‌های نوید دهندهٔ پاداش ظاهر می‌شود. اما کنترلر ابزاری به دنبال یادگیری پیوند عمل- نتیجه است. البته یادگیری ابزاری تنها به نتیجهٔ عمل کنونی بسنده نکرده، و ممکن است به صورت یک برنامه‌ریزی رو به جلو دنباله‌ای از عمل‌ها را ارزیابی کند. از طرفی، برنامه‌ریزی رو به جلو ممکن است تنها فرآیند برنامه‌ریزی‌ای نباشد که یادگیری ابزاری از آن استفاده می‌کند. ممکن است انسان‌ها از برنامه‌ریزی روبه‌عقب نیز به منظور ارزیابی توالی عمل‌ها بهره برند. با این وجود برنامه‌ریزی روبه‌عقب کمتر تاکنون مورد توجه قرار گرفته است. پژوهش‌های پیشین نشان دادند با وجود مستقل بودن یادگیری پاولفی و ابزاری، آن‌ها با یکدیگر تعامل می‌کنند. در حقیقت یادگیری پاولفی نزدیک شوندگی روی برنامه‌ریزی رو به جلو تأثیر گذاشته و منجر به اتخاذ تصمیماتی می‌شود که ممکن است از نظر کنترلر ابزاری بهینه نباشند. اما تأثیر یادگیری پاولفی روی برنامه‌ریزی روبه‌عقب هنوز مطالعه نشده است. ***مواد و روش‌ها:*** در این مقاله، ما یک آزمایش مسیریابی طراحی کردیم که امکان برنامه‌ریزی‌های رو به جلو، رو به عقب، و دوجهته در آن فراهم است، و ایماهای پاولفی نزدیک شوندگی را نیز در نقشه‌ها تعبیه نمودیم. ***یافته‌ها:*** تحلیل آماری داده‌های جمع‌آوری شده نه تنها از وجود برنامه‌ریزی رو به عقب حکایت می‌کنند، بلکه نشان می‌دهند که ایمای پاولفی نزدیک شوندگی بر روی سه برنامه‌ریزی تاثیر می‌گذارد، هر چند که این تأثیر در برنامه‌ریزی دوجهته بیش‌تر از روبه‌جلو، و در روبه‌جلو بیش‌تر از روبه‌عقب است. همچنین در بستر یادگیری تقویتی، الگوریتم برنامه‌ریزی دوجهته را تحت بایاس پاولفی توسعه دادیم. ***نتیجه‌گیری:*** نتایج شبیه‌سازی با نتایج برآمده از آزمایش سازگار بوده و بیان می‌کنند که تأثیر بایاس پاولفی را می‌توان به نوعی در قالب هرس درختان تصمیم مدل‌سازی نمود.

واژه‌های کلیدی:
۱- تصمیم‌گیری
۲- برنامه‌ریزی راهبردی
۳- یادگیری ابزاری
۴- مدل‌سازی کامپیوتری

***نویسنده مسئول:** محمد تقی حمیدی بهشتی

پست الکترونیک: *mbehesht@modares.ac.ir*

[DOI: 10.52547/shefa.9.4.51]

مقدمه

زندگی روزمرۀ انسان در شرایط متفاوت با ارزیابی تصمیمات مختلف از ساده تا پیچیده درهم آمیخته است، که هر کدام ممکن است منجر به نتایج فوری یا تأخیری، مثبت یا منفی شوند. آزمایش مسیریابی را تصور کنید که می‌خواهیم در آن از نقطۀ شروع به طرف نقطۀ هدف عزیمت کنیم. مسیرهای متنوعی ممکن است در پیش روی داشته باشیم. دستیابی کارآمد به نقطۀ هدف، نیازمند ارزیابی شاخه‌های تصمیم‌گیری مختلفی است که از نقطۀ شروع به سمت هدف به منظور یافتن بهترین مسیر گسترش می‌یابند. فرآیند ارزیابی توالی تصمیماتی که دستیابی به هدف را میسر می‌کنند، به‌عنوان برنامه‌ریزی رو به جلو[1] شناخته می‌شود (۱). اما شاید برنامه‌ریزی رو به جلو تنها راهبرد تصمیم‌گیری انسان‌ها نباشد. در حقیقت این احتمال وجود دارد که انسان‌ها به‌منظور ارزیابی کارآمدتر گزینه‌های موجود، درخت تصمیم را از سمت نقطۀ هدف نیز به سمت نقطۀ شروع گسترش دهند. چنان روند برنامه‌ریزی که به صورت وارونه از نقطۀ هدف به سمت نقطۀ شروع جاری می‌شود در ادبیات هوش مصنوعی به‌عنوان برنامه‌ریزی روبه‌عقب[2] شناخته می‌شود (۲). با وجود این‌که پژوهش‌های هوش مصنوعی از کارآمدی برنامه‌ریزی رو به عقب حکایت می‌کنند، اما گسترش درختان تصمیم تا حداکثر عمق ممکن و در دو جهت رو به جلو و رو به عقب از نظر رایانشی و زمان در دسترس، شاید میسر نباشد (۴-۳). حتی اگر میسر باشد، چنین ارزیابی دقیقی نیازمند رایانش[3] سنگینی است که فرآیند تصمیم‌گیری را بسیار کند می‌نماید. بنابراین به سطحی از کنترل در گسترش درختان تصمیم، یا هرس کردن آن‌ها نیازمند هستیم (۵). درخت تصمیم به توالی مجموعه‌ای از تصمیمات احتمالی اشاره دارد که در مواجهه با پدیده‌ای می‌توانیم اتخاذ کنیم. هر یک از تصمیمات، شاخۀ درختی را می‌سازد که به نوبۀ خود به تصمیمات دیگر و شاخه‌های دیگری می‌تواند منجر شود. در پدیده‌های پیچیده ممکن است شاهد رشد درخت تصمیم به صورت فزاینده‌ای باشیم. در نتیجه به ارزیابی شاخه‌های مختلف درخت تصمیم نیازمندیم. ارزیابی درختان تصمیم به منظور یافتن بهترین توالی تصمیم‌ها توسط یادگیری ابزاری[4] در مغز انسان کنترل می‌شود (۶). سیستم ابزاری با ارزیابی ارتباط مستقیم بین عمل انتخابی و نتیجۀ آن به یادگیری منجر می‌شود. اما عمل انتخابی تنها تحت کنترل سیستم ابزاری نبوده، بلکه توسط یادگیری پاولفی[5] نیز متأثر می‌شود (۷). یادگیری پاولفی، پاسخ‌های غریزی هستند که به‌وسیلۀ ایماهای محیط[6] هدایت می‌شوند.

به‌گونه‌ای که انسان‌ها به انتخاب پاسخ‌هایی که نوید دهندۀ پاداش هستند تمایل بیشتری نشان می‌دهند در مقایسه با پاسخ‌های خنثی، و پاسخ‌هایی که نشان از دریافت تنبیهی دارند (۸). با وجود مستقل بودن این دو سیستم، مطالعات فراوانی از تعاملات آن‌ها حکایت دارند (۱۲-۹). به گونه‌ای که تمایلات پاولفی روی انتخاب‌هایی که از نظر ابزاری مناسب ارزیابی می‌شوند تأثیر گذاشته به طوری که تصمیم‌گیرنده به انتخاب عمل‌هایی گرایش پیدا می‌کند که در پیوند با ایماهای پاولفی قرار دارند حتی اگر آن عمل‌ها بهینه‌ترین عمل‌های شناسایی شده توسط سیستم ابزاری نباشند (۱۳). این نوع تأثیرات یادگیری پاولفی روی انتخاب عمل ابزاری به‌عنوان بایاس پاولفی[7] شناخته می‌شود که در دو گونه تمایل به نزدیک‌شدن[8] به سمت ایماهای نوید دهندۀ پاداش، و دوری جستن[9] از انتخاب عمل‌هایی که پیش‌بینی کنندۀ تنبیه هستند، ظاهر می‌شوند (۱۶-۱۴). درحالی‌که نتایج برآمده از بایاس پاولفی ممکن است یادگیری ابزاری را کنترل کنند، پاسخ‌های پاولفی بدون توجه به این‌که چه عملی انتخاب شده است، اتفاق می‌افتند. در نتیجه فرآیند تصمیم‌گیری ممکن است از اتخاذ تصمیم بهینه زاویه بگیرد (۱۷). اما یادگیری ابزاری عمل‌هایی را پیشنهاد می‌دهد که مجموع پاداش مورد انتظار بلند- مدت را بهینه می‌کنند و بدین جهت به پیامد عمل وابسته است (۱۸). پژوهش‌های پیشین نشان داده‌اند که بایاس نزدیک‌شوندگی پاولفی[10] روی برنامه‌ریزی رو به جلو تأثیر می‌گذارد و منجر به انتخاب عمل‌هایی می‌شود که لزوماً از نظر ابزاری بهینه نیستند (۱۹). اما این‌که آیا برنامه‌ریزی رو به عقب نیز توسط این بایاس متأثر می‌شود هنوز به خوبی مورد مطالعه قرار نگرفته است. در نتیجه در این مقاله ما نخست بررسی خواهیم کرد که آیا انسان‌ها از برنامه‌ریزی رو به عقب استفاده می‌کنند یا خیر. دوم، آیا تمایل به نزدیک‌شدن پاولفی، برنامه‌ریزی رو به عقب را بایاس می‌نماید یا خیر. بدین منظور، ما یک آزمایش مسیریابی طراحی نموده و ایمای نزدیک شوندگی پاولفی را در آن جاسازی کردیم تا به این پرسش‌های پژوهشی پاسخ دهیم: آیا انسان‌ها از راهبرد برنامه‌ریزی رو به جلو استفاده می‌کنند؟ و اگر این چنین است آیا این راهبرد توسط تمایل به نزدیک شدن پاولفی بایاس می‌شود؟ دوم، آیا انسان‌ها از راهبرد برنامه‌ریزی رو به عقب بهره می‌برند؟ و اگر چنین باشد آیا این راهبرد نیز توسط تمایل به نزدیک شدن پاولفی بایاس می‌شود؟ سوم، آیا انسان‌ها از دو برنامه‌ریزی رو به جلو و رو به عقب به صورت ترکیبی استفاده می‌کنند، و آیا این راهبرد توسط تمایل به

[1] Forward Planning
[2] Backward Planning
[3] Computation
[4] Instrumental Learning
[5] Pavlovian Learning
[6] Environmental Cues
[7] Pavlovian Bias
[8] Approach
[9] Withdraw
[10] Pavlovian Approach Bias

[DOI: 10.52547/shefa.9.4.51]

نزدیک شدن پاولفی بایاس می‌شود؟ ما این نوع برنامه‌ریزی ترکیبی را برنامه‌ریزی دوجهته[11] می‌نامیم. در نهایت بدنبال پاسخ‌گویی به این پرسش هستیم که تأثیر بایاس پاولفی از نوع تمایل به نزدیک شدن روی کدام یک از این سه نوع برنامه‌ریزی بیشتر است؟

مواد و روش‌ها

طراحی آزمایش: آزمایش طراحی شده شامل نقشه‌هایی مشابه با نقشه‌های مسابقات کامپیوتری سادهٔ اتومبیل‌رانی هستند. تصویر-۱ نمونه‌هایی از این نقشه‌ها را نشان می‌دهد. در همهٔ نقشه‌ها نقطهٔ هدف در شمال، و نقطهٔ شروع در جنوب قرار دارد. نقشه‌ها به گونه‌ای طراحی شده‌اند که ایمای نزدیک‌شوندگی پاولفی در آن‌ها تعبیه شده است. به منظور بررسی وجود بایاس پاولفی روی برنامه‌ریزی روبه‌جلو، ایمای نزدیک شوندگی پاولفی را در نزدیکی نقطهٔ شروع، و برای بررسی این بایاس روی برنامه‌ریزی رو به عقب، ایمای نزدیک‌شوندگی را در نزدیکی نقطهٔ هدف تعبیه کردیم. تصویر ۱-آ و ۱-ب نقشه‌هایی را نشان می‌دهند که در آن‌ها ایمای نزدیک‌شوندگی به ترتیب در نزدیکی نقطهٔ شروع و هدف قرار دارد. هدف بازی این است که فرد سریع‌ترین مسیر برای دستیابی به نقطهٔ هدف را بیابد. همان‌طور که از تصویر ۱-آ پیداست، هنگامی که فرد در نقطهٔ شروع قرارداد، دو عمل می‌تواند انتخاب کند: عمل بالا و عمل پایین. در هر نقشه طول دو مسیر چپ و راست دقیقاً یکسان است. این بدان معناست که از نظر یادگیری ابزاری هدف‌گرا[12]، هیچ تفاوتی بین دو مسیر وجود ندارد. چرا که مدت زمان مورد نیاز برای رسیدن به نقطهٔ هدف کاملاً یکسان است. یادگیری هدف‌گرا از آموختن ارتباط بین عمل و نتیجه، عمل‌های موجود را ارزیابی می‌کند در نتیجه انتخاب عمل بالا و پایین از نظر این یادگیری یکسان است چرا که طول دو مسیر و در نتیجه مدت زمان لازم برای رسیدن به هدف از دو مسیر برابر است. اما از نظر یادگیری پاولفی ممکن است این گونه نباشد. هنگامی که فرد عمل بالا را انتخاب می‌کند در مسیری قرار می‌گیرد که او را به نقطهٔ هدف از لحاظ جغرافیایی نزدیک‌تر می‌کند، در حالی‌که گام نهادن در مسیر پایین او را از نقطه‌ی هدف دور می‌کند. در نتیجه مسیر بالارونده به نوعی نوید دهندهٔ پاداش است چرا که یادگیری پاولفی تمایل ذاتی فرد را در انتخاب عملی که وی را به سمت پاداش سوق می‌دهد تقویت می‌کند. توضیح مشابهی برای برنامه‌ریزی رو به عقب وجود دارد با این تفاوت که ایمای نزدیک‌شوندگی پاولفی در نزدیکی نقطهٔ هدف تعبیه شده است. تصویر ۱-ب چنین نقشه‌ای را نشان می‌دهد. در برنامه‌ریزی رو به عقب، فرد خود را در نقطهٔ هدف تصور می‌کند و مسیرهایی که وی را به سمت این نقطه می‌آورند، ارزیابی می‌کند از این منظر که کدام یک از دو مسیر او را سریع‌تر از نقطهٔ شروع به نقطهٔ هدف می‌رساند. در این‌جا مشابه با حالت برنامه‌ریزی روبه‌جلو، طول دو مسیر چپ و راست کاملاً یکسان است در نتیجه از نقطه نظر کنترلر هدف‌گرا تفاوتی بین دو مسیر وجود ندارد. در این آزمایش، ما همچنین نقشه‌هایی را طراحی کردیم که شامل ایماهای نزدیک‌شوندگی پاولفی تعبیه شده، هم در نزدیکی نقطهٔ شروع و هم در نزدیکی نقطهٔ هدف، هستند. این نقشه‌ها به منظور بررسی تأثیر بایاس نزدیک‌شوندگی پاولفی بر روی برنامه‌ریزی دوجهته طراحی گردیده‌اند. تصویر ۱-پ نمونه‌ای از این نقشه‌ها را نشان می‌دهد.

به‌منظور بررسی دقیق رفتار افراد ما نیازمند طراحی نقشه‌هایی نیز هستیم که در آن‌ها هیچ‌گونه ایمای پاولفی وجود ندارد. بدین منظور دو دسته از این نقشه‌های خنثی را طراحی نمودیم: نقشه‌هایی که دو مسیر چپ و راست در آن‌ها کاملاً متقارن هستند؛ و نقشه‌هایی که در آن‌ها یکی از دو مسیر به طور آشکاری از دیگری کوچک‌تر است. تصویر ۱-ت و ۱-ث به ترتیب نمونه‌ای از نقشه‌های خنثی متقارن و نامتقاران را نشان می‌دهند. انتظار داریم که در دستهٔ اول احتمال انتخاب هر یک از دو مسیر چپ و راست توسط افراد تقریبا ۰/۵ باشد، و در دستهٔ دوم احتمال انتخاب مسیر کوتاه‌تر نزدیک به یک باشد.

جمع‌آوری داده: تصویر ۱-ج ساختار بازی‌ای (آزمایشی) که در بستر آن به جمع‌آوری داده پرداختیم را نشان می‌دهد. ۳۰ شرکت کننده به صورت آنلاین در بازی شرکت کردند و به کمک کلیدهای جهت نمای روی صفحه کلید انتخاب‌های خود را ثبت نمودند. چگونگی انجام بازی و جزییات مربوط به آن در ابتدا در اختیارشان قرار داده شد. همچنین بازی شامل یک فاز یادگیری بوده که افراد با نحوهٔ انجام بازی و ثبت انتخاب‌ها بیشتر آشنا گردند. در فاز تست هر بازی شامل ۱۶ نقشهٔ رو به جلو، ۱۶ نقشهٔ رو به عقب، و ۱۶ نقشهٔ دوجهته، و تصویر آینه‌ای آن‌ها به منظور پارسنگ[13] کردن، تشکیل شده است (۱۶*۳*۲=۹۶). همچنین ۱۲ نقشه از نوع خنثای متقارن و ۱۲ نقشهٔ خنثای نامتقارن نیز وجود دارد. بنابراین فاز تست در مجموع شامل ۱۲۰ نقشهٔ متفاوت می‌باشد. این آزمایش تحت نظارت کمیتهٔ اخلاق پژوهشی دانشگاه تربیت مدرس انجام شده است.

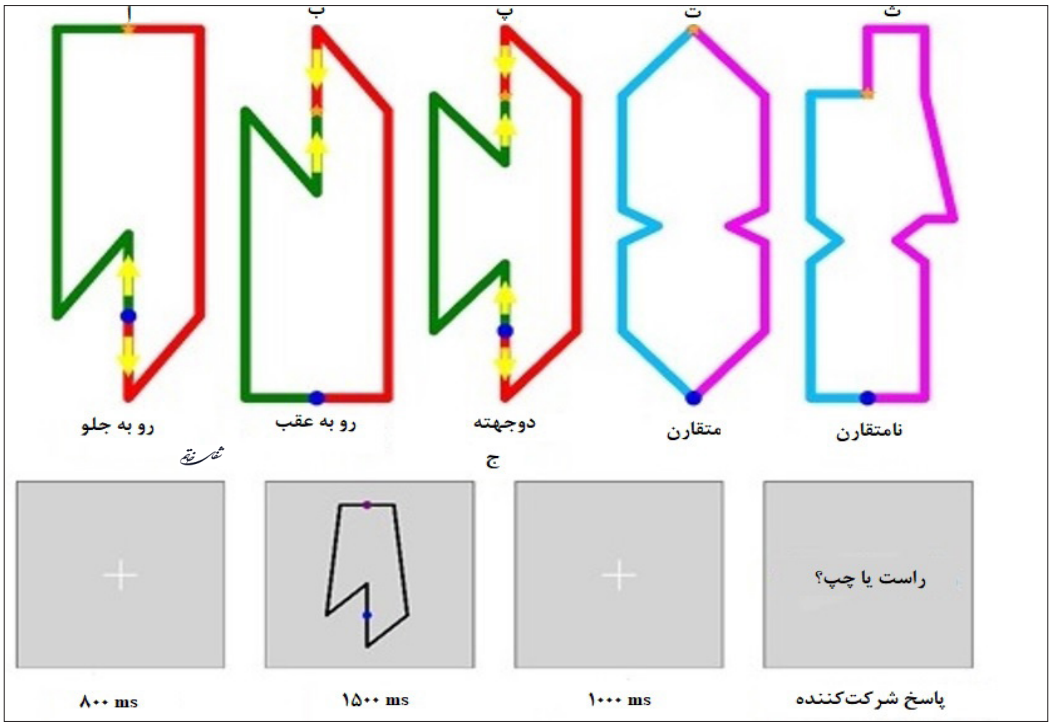
یافته‌ها

[11] Bidirectional Planning

[12] Goal- Directed

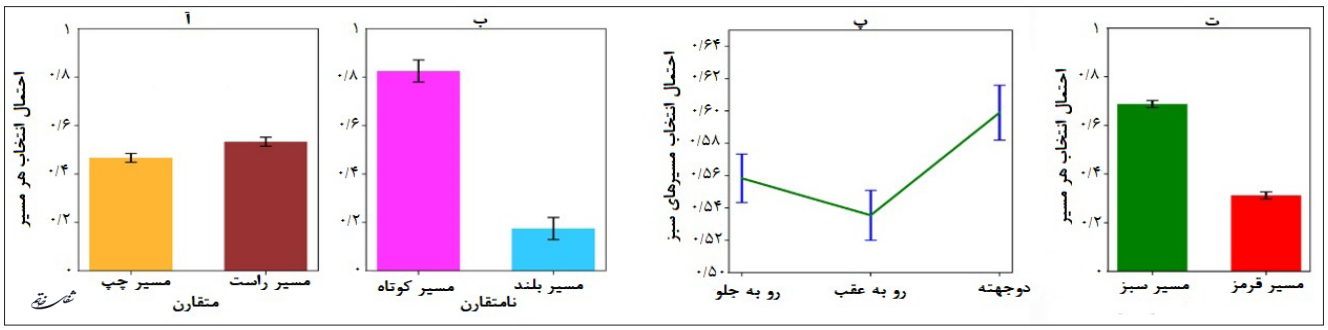
[13] Counterbalancing

[DOI: 10.52547/shefa.9.4.51]



تصویر ۱- سطر اول نمونه‌ای از نقشه‌هایی را نشان می‌دهد که در پنج کلاس مختلف رو به جلو، رو به عقب، دوجهته، متقارن و نامتقارن مورد استفاده قرار گرفته است. سه کلاس اول شامل ایمای پاولفی بوده، و دو کلاس آخر فاقد ایمای پاولفی هستند. به جز در نقشه‌های نامتقارن که یکی از دو مسیر چپ و راست به طور آشکاری از دیگری کوتاه‌تر است، در سایر کلاس‌ها هر دو مسیر دقیقاً طول یکسانی دارند. نقطهٔ شروع (دایرهٔ آبی) همواره در جنوب و نقطهٔ هدف (ستارهٔ طلایی) همواره در شمال قرار دارد. در این‌جا صرفاً به‌منظور بهتر دیده شدن دو مسیر راست و چپ دارای رنگ‌های متفاوتی هستند اما در آزمایش هر دو به یک رنگ هستند. همچنین پیکان‌های زرد رنگ نیز صرفاً به منظور نشان دادن ایمای پاولفی به نقشه‌ها اضافه شده‌اند اما در آزمایش هیچ‌گونه پیکانی وجود ندارد. سطر دوم ساختار آزمایش را نشان می‌دهد که شامل چهار تصویر می‌باشد که به ترتیب و با مدت زمان مشخصی نشان داده می‌شوند. در نهایت شرکت‌کنندگان باید مسیری را انتخاب کنند که از نظر آن‌ها مسیر کوتاه‌تری است.

در این بخش به کمک تحلیل آماری T-test به بررسی انتخاب‌های افراد شرکت کننده در آزمایش می‌پردازیم. تصویر ۲-آ و ۲-ب میانگین احتمال انتخاب‌های هر یک از دو مسیر موجود در نقشه‌های متقارن و نامتقارن را نشان می‌دهد. همان‌طور که انتظار می‌رفت در نقشه‌های نامتقارن (تصویر ۲-ب)، احتمال انتخاب مسیر کوتاه‌تر به طور آشکاری (P<۰/۰۰۱) بیشتر از مسیر بلندتر می‌باشد که نشان می‌دهد که افراد تمایل زیادی به انتخاب مسیر کوتاه‌تر دارند. از آن‌جا که این نقشه‌ها فاقد ایمای پاولفی هستند، می‌توان نتیجه گرفت که تصمیمات افراد توسط یادگیری ابزاری هدایت می‌شود. یادگیری ابزاری از نوع هدف‌گرا موجب می‌شود که افراد مسیری را انتخاب کنند که در زمان کوتاه‌تری به مقصد رسیده و منجر به برنده شدن در بازی می‌شود. تصویر ۲-آ میانگین احتمال انتخاب‌های مسیر چپ و راست در نقشه‌های خنثای متقارن را نشان می‌دهد. همان‌طور که از تصویر پیداست احتمال انتخاب هر یک از دو مسیر نزدیک به ۰/۵ است (P<۰/۱۷) که مشابه با نقشه‌های نامتقارن بر هدایت تصمیمات توسط یادگیری ابزاری دلالت دارد.



تصویر ۲- نتایج داده‌های جمع‌آوری شده از آزمایش. تصویر آ نشان می‌دهد که در زمانی‌که ایمای پاولفی وجود ندارد (و دو مسیر متقارن هستند)، افراد اولویت خاصی در انتخاب مسیرها ندارند بلکه به صورت شانسی یکی از دو مسیر را انتخاب می‌کنند. اما تصویر ب نشان می‌دهد چنانچه یکی از مسیرها به طور مشخصی از دیگری کوتاه‌تر باشد، افراد تمایل دارند در اکثر موارد مسیر کوتاه را انتخاب کنند. تصویر پ بیان می‌کند که افراد نه تنها از برنامه‌ریزی روبه‌جلو، بلکه از برنامه‌ریزی رو به عقب و در نتیجه دوجهته نیز استفاده می‌کنند چرا که هر سه کلاس توسط ایمای نزدیک‌شوندگی پاولفی بایاس شده‌اند. به این دلیل که احتمال انتخاب مسیر سبز (یعنی مسیری که شامل ایمای نزدیک‌شوندگی پاولفی است)، در هر سه کلاس بیش‌تر از ۰/۵ می‌باشد. همچنین قدرت بایاس کنندگی ایمای نزدیک‌شوندگی پاولفی در کلاس روبه‌جلو قوی‌تر از رو به عقب، و در کلاس دوجهته قوی‌تر از روبه‌جلو می‌باشد.

[DOI: 10.52547/shefa.9.4.51]

حال به بررسی انتخاب افراد در نقشه‌های شامل ایمای پاولفی می‌پردازیم. تصویر ۲-پ میانگین احتمال انتخاب مسیر سبز (مسیری که شامل ایمای پاولفی است) توسط ۳۰ شرکت کننده را در ۳ کلاس مختلف (رو به جلو، رو به عقب، و دوجهته) نشان می‌دهد. یافتهٔ نخست این است که احتمال انتخاب مسیر سبز در هر سه کلاس بیشتر از ۰/۵ است (P<۰/۰۰۱). این امر نشان می‌دهد افراد تحت تأثیر ایماهای پیش‌بینی کنندهٔ پاداش پاولفی هستند. چرا که تصمیمات‌شان به سمت انتخاب مسیری که شامل ایمای نزدیک‌شوندگی پاولفی است بایاس شده است درحالی‌که مسیر سبز و قرمز از لحاظ طول با یکدیگر برابر هستند. یافتهٔ دوم این است که احتمال انتخاب مسیر سبز در کلاس رو به جلو بیشتر از کلاس رو به عقب است (P<۰/۰۵) این نتیجه بیان می‌کند که ایمای نزدیک شوندگی پاولفی در برنامه‌ریزی رو به جلو در مقایسه با برنامه‌ریزی رو به عقب بایاس قوی‌تری روی انتخاب افراد اعمال می‌کند. یافتهٔ سوم این است که احتمال انتخاب مسیر سبز در برنامه‌ریزی دوجهته حتی از برنامه‌ریزی رو به جلو نیز بیشتر است (P<۰/۰۵)،که نشان می‌دهد افراد به صورت دوجهته به تحلیل مسیرهای موجود پرداخته و در نهایت تحت تأثیر بایاس نزدیک‌شوندگی پاولفی مسیری را انتخاب نمودند.

طراحی الگوریتم برنامه‌ریزی دوجهته: از منظر رایانشی، تئوری کنترل بهینه تطبیقی برای یافتن بهترین توالی تصمیمات برای دستیابی به یک هدف خاص را می‌توان تحت چارچوب یادگیری تقویتی مطالعه کرد. یادگیری تقویتی[14] از دو واحد به نام‌های عامل[15] و محیط[16] تشکیل شده است. عامل واحد یادگیرنده می‌باشد که با انتخاب عمل‌هایی[17] با محیط پیرامون خود به تعامل پرداخته و با آنالیز فیدبک‌های دریافتی از طرف محیط در طول زمان بهترین توالی تصمیمات جهت دست‌یابی به هدف را می‌آموزد (۲۰). بسته به در دسترس بودن یا نبودن مدل محیط، مسالهٔ یادگیری تقویتی به دو دستهٔ مبتنی بر مدل و بدون نیاز[18] به مدل تقسیم‌بندی می‌شود. اولی جهت مدل‌سازی یادگیری ابزاری از نوع هدف‌گرا، و دومی به منظور مدل‌سازی یادگیری عادتی[19] در مقالات مورد استفاده قرار گرفته است (۲۳-۲۱). در این مقاله، ما فقط بر یادگیری مبتنی بر مدل به‌عنوان یادگیری ابزاری هدف‌گرا تمرکز می‌کنیم زیرا که در آزمایش طراحی شده نقشهٔ بازی به‌عنوان مدل محیط در دسترس شرکت‌کنندگان قرار دارد. در این مقاله ما از الگوریتم تکرار ارزش[20] که یک الگوریتم یادگیری تقویتی مبتنی بر مدل است برای توسعهٔ الگوریتم دوجهته تحت بایاس نزدیک شوندگی پاولفی استفاده می‌کنیم (۲۰). قانون به‌روزرسانی این الگوریتم در زیر آمده است، جایی‌که، s حالت محیط، a عمل انتخابی عامل در حالت s، و r پاداش دریافتی را نشان می‌دهد. همچنین $\pi(a \mid s)$ سیاست رفتاری عامل می‌باشد که تعیین می‌کند در هر حالت s چه عملی را عامل انجام دهد. $Q^{\pi}(s,a)$ تابع ارزش حالت- عمل می‌باشد که امید ریاضی مجموع پاداش بلند- مدت دریافتی توسط عامل وقتی که وی با انتخاب عمل aدر حالت s شروع می‌کند و مطابق با سیاست رفتاری π عمل می‌کند را نشان می‌دهد.

$$Q^{\pi}(s,a)=r+\gamma V^{\pi}(\acute{s})$$

در این رابطه $V\pi(\acute{s})$ که به صورت زیر تعریف می‌شود میزان تابع ارزش در حالت بعدی ś در تکرار قبل می‌باشد.

$$V^{\pi}(\acute{s}) = \max_{\acute{a}} Q^{\pi}(\acute{s},\acute{a})$$

قانون به‌روزرسانی در این الگوریتم یک قانون یک گام به جلو بوده، و همچنین فاقد تأثیرات پاولفی می‌باشد. در نتیجه، به‌منظور مدل کردن برنامه‌ریزی‌ها تحت تأثیر بایاس پاولفی، نیازمند آنیم که این الگوریتم را تغییر دهیم. هنگامی‌که هیچ بایاس پاولفی یا راهبرد هرس دیگری وجود نداشته باشد، انتظار می‌رود که عامل درخت تصمیم را با عمق یکسان در تمام شاخه‌های موجود گسترش دهد. اما بایاس پاولفی با تشویق عامل به گرفتن تصمیماتی که وی را به سمت دریافت پاداش نزدیک می‌کند به کنترل تصمیمات هدف‌گرای عامل می‌پردازد. در نتیجه به‌منظور مدل کردن این تأثیر ما پیشنهاد می‌کنیم که عامل درخت تصمیم را با عمق مختلف در زیرشاخه‌های متفاوت گسترش می‌دهد و گسترش درخت در شاخه‌ای که شامل ایمای پاولفی است عمیق‌تر از شاخه‌ای است که شامل این ایما نیست. به‌عنوان مثال، تصویر ۳-آ را در نظر بگیرید که شامل دو مسیر از نقطهٔ شروع به نقطه هدف است و مشابه نقشه‌هایی است که شرکت‌کنندگان در آزمایش با آن روبرو بوده‌اند. در چارچوب پیشنهادی ما، عامل درخت تصمیم را تا عمق $d_{green}^{forward}$ در مسیر سبز گسترش می‌دهد، و تا عمق $d_{red}^{forward}$ در مسیر قرمز، به‌طوری که $d_{green}^{forward} > d_{red}^{forward}$ بوده و مجموع آن‌ها باید کمتر- مساوی ظرفیت کلی عامل در بسط درخت تصمیم به صورت

$$Q^{forward}(s_{start},a_x)=r_0+\gamma r_1+\ldots+\gamma^{d_x^{forward}-1} r_{d_x^{forward}-1}+\gamma^{d_x^{forward}} V^{\pi}\left(s_{d_x^{forward}}\right),\ x\in\{green, red\}$$

ارزش را به صورت زیر پیشنهاد می‌دهیم: به طور مشابه، در برنامه‌ریزی روبه‌عقب عامل پاداش حالت هدف را به حالت‌های همسایهٔ خود و به طرف حالت شروع پس‌انتشار می‌دهد. ما فرض می‌کنیم که ظرفیت انتشار رو به عقب نیز محدود است. همچنین پیشنهاد می‌کنیم که عامل می‌تواند ارزش

[14] Reinforcement Learning
[15] Agent
[16] Environment
[17] Actions
[18] Model- Based and Model-Free
[19] Habitual
[20] Value Iteration

[DOI: 10.52547/shefa.9.4.51]

حالت نهایی را در مسیرهای مختلف با عمق‌های متفاوت پس‌انتشار دهد. این بدان معناست که عامل می‌تواند پاداش را با عمق $d_{green}^{backward}$ در مسیر سبز و با عمق $d_{red}^{backward}$ در مسیر قرمز پس‌انتشار دهد. به‌طوری که، $d_{green}^{backward} > d_{red}^{backward}$

و مجموع آن‌ها باید کمتر- مساوی ظرفیت بسط درخت در وضعیت پس‌انتشار باشد. مشابه برنامه‌ریزی روبه‌جلو که در آن پاداش‌های آینده کاهیده می‌شوند، در برنامه‌ریزی روبه‌عقب نیز ما پاداش حالت هدف را هنگام پس‌انتشار به سمت نقطهٔ شروع کاهش می‌دهیم. در نتیجه خواهیم داشت:

$$V_x^{backward}(s_m)=\gamma^m r_{goal}\ ,\quad m=1,2,\ldots,d_x^{backward}\ ,\ x\in\{green, red\}$$

در این‌جا m عمق بسط درخت به صورت عقب‌گرد را نشان می‌دهد. در برنامه‌ریزی دوجهته، ترکیب دو وضعیت فوق را خواهیم داشت. از یک طرف عامل درختان تصمیم را با عمق‌های $d_{green}^{forward}$ و $d_{red}^{forward}$ به صورت روبه‌جلو در دو مسیر سبز و قرمز گسترش می‌دهد. و از طرف دیگر، عامل ارزش حالت هدف را با عمق‌های $d_{red}^{backward}$ و $d_{green}^{backward}$ در دو مسیر سبز و قرمز به صورت عقب‌گرد انتشار می‌دهد. میزان کلی بسط درخت تصمیم باید کوچکتر مساوی ظرفیت کلی بسط درخت باشد:

$$d_{green}^{forward}+d_{red}^{forward}+d_{green}^{backward}+d_{red}^{backward}\le d^{total}$$

نکتهٔ حائز اهمیت در این‌جا این است که درختان تصمیم در حین گسترش رو به جلو و رو به عقب ممکن است در بین مسیر از حالت مشترکی گذر کنند. در این صورت دو اقدام در جهت تطبیق دو تخمین رو به جلو و رو به عقب انجام می‌دهیم. نخست این‌که، در وضعیت‌هایی که دو درخت تصمیم رو به جلو و رو به عقب یکدیگر را در حالتی در بین مسیر ملاقات کردند، ما گسترش درختان تصمیم را از آن حالت به بعد متوقف می‌کنیم. اقدام دوم این‌که مقدار ارزش حالتی که توسط برنامه‌ریزی رو به عقب تخمین زده شده است را در تخمین برنامه‌ریزی رو به جلو دخیل می‌کنیم. این امر در روابط زیر نشان داده شده است:

$$Q^{bidirectional}(s_{start}a_{green}) = r_0+\gamma r_1+\cdots+\gamma_{green}^{forkwardToMeet-1} r_{d_{green}^{forkwardToMeet}-1}+\gamma^{a_{green}^{forkwardToMeet}} V_{max}$$

$$V_{max}=\max\{V^{backward}(s_{green}^{backwardToMeer}).V^{forward}(s_{green}^{forwardToMeer}\}$$

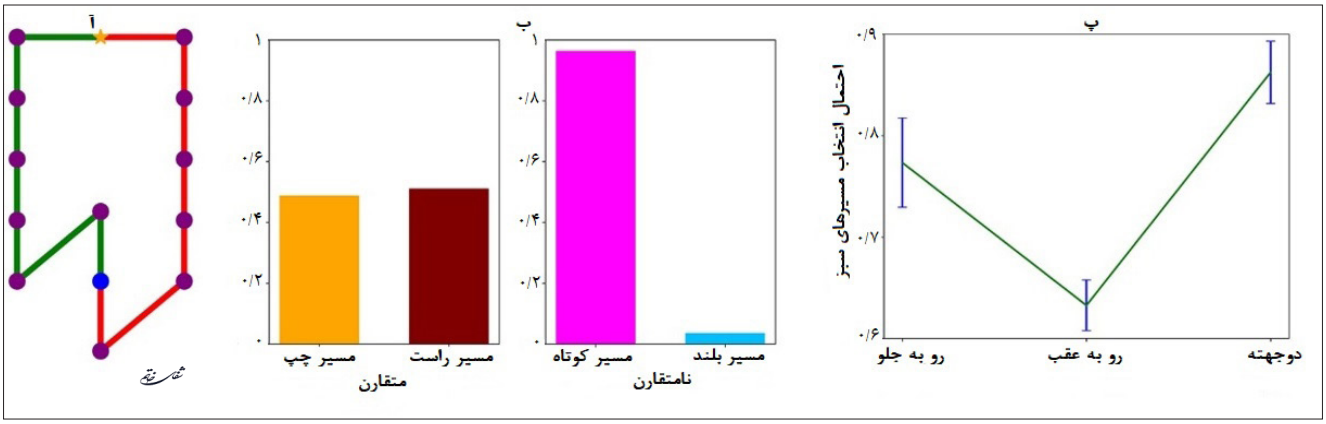
$$V^{backward}(s_{green}^{backwardToMeer})=\gamma^{backwardToMeet} r_{goalState}$$

همچنین برای در نظر گرفتن تأثیر بایاس پاولفی در مدل، پیشنهاد می‌کنیم که احتمال این‌که عامل درختی عمیق‌تر در مسیری که شامل ایماهای پاولفی است گسترش دهد بیشتر از احتمال گسترش درخت تصمیم عمیق در مسیری است که شامل این ایماها نیست.

$$\pi(s_{start},a_{green})>\pi(s_{start},a_{red})$$

در نقشه‌های رو به جلو که ایمای پاولفی در نزدیکی شروع قرار دارد، هنگامی که عامل برنامه‌ریزی را آغاز می‌کند، احتمال این‌که عامل عملی که منجر به گام نهادن در مسیر سبز می‌شود انجام دهد بیشتر از احتمال انتخاب عملی است که منجر به رفتن در مسیر قرمز می‌شود.

نتایج شبیه‌سازی الگوریتم پیشنهادی: در این بخش چند نقشه، به‌عنوان مثال‌های موردی در نظر گرفته، و از الگوریتم پیشنهادی جهت مسیریابی در آن‌ها استفاده می‌کنیم تا بتوانیم نتایج عمل‌کرد الگوریتم را با داده‌های جمع‌آوری شده از شرکت‌کنندگان در آزمایش مقایسه کنیم. نتیجهٔ اجرای الگوریتم پیشنهادی برای نقشه‌های خنثی در تصویر ۳-ب آمده است. همان‌طور که انتظار می‌رفت، در نقشه‌های متقارن احتمال انتخاب دو مسیر یکسان می‌باشد، در حالی‌که در نقشهٔ نامتقارن احتمال انتخاب مسیر کوتاه‌تر نزدیک به یک است. روشن است که این روند تصمیم‌گیری



تصویر ۳- نتایج شبیه‌سازی الگوریتم برنامه‌ریزی دوجهتهٔ پیشنهادی تحت بایاس نزدیک‌شوندگی پاولفی. تصویر ب نشان می‌دهد در زمانی که ایمای پاولفی وجود ندارد، و راهبرد رفتاری سافتمکس عامل وزن یکسانی به انتخاب دو مسیر اختصاص می‌دهد، عامل‌ها توسط یادگیری تقویتی مبتنی بر مدل هدایت شده و مسیر منطقی را انتخاب می‌کنند. در زمانی که دو مسیر یکسان هستند، الگوریتم این عدم تفاوت را درک کرده و ارزش دو عمل که به مسیرهای چپ و راست منتهی می‌شوند را یکسان در نظر می‌گیرد. همچنین در زمانی که یکی از دو مسیر کوتاه‌تر از دیگری است مجدداً الگوریتم به طور مناسبی می‌آموزد که کوتاه‌ترین مسیر کدام است و آن را در اکثر موارد انتخاب می‌کند. تصویر پ نشان می‌دهد با این‌که مسیرها در هر سه کلاس رو به جلو، رو به عقب، و دوجهته برابرند اما ایمای پاولفی که به صورت اختصاص وزن بیشتر در راهبرد سافتمکس مدل شده است، عامل را به انتخاب مسیری که شامل این ایما هست تشویق می‌کند. همچنین از آنجا که ظرفیت گسترش درخت تصمیم در حالت روبه‌عقب کوچک‌تر از روبه‌جلو در نظر گرفته شده است، تأثیر وزن‌دهی روی عمل‌های انتخابی عامل نیز کمتر خواهد بود. از طرف دیگر، از آن‌جا که در حالت رو به جلو امکان ملاقات درختان تصمیمی که به صورت رو به جلو و رو به عقب گسترش داده شده‌اند وجود دارد، امکان تأثیرپذیری عمل‌ها از ایمای پاولفی نیز افزایش می‌یابد. این بدان معناست که هر چقدر که عامل ظرفیت بیش‌تری در گسترش درخت‌های تصمیم داشته باشد، احتمال آن‌که توسط بایاس پاولفی متأثر شود نیز بیش‌تر است. دلیل این امر آن است که هر چقدر ظرفیت گسترش درخت تصمیم در حالت رو به جلو (یا رو به عقب) بیش‌تر باشد احتمال اینکه درخت گسترش داده شده نقطهٔ هدف (یا شروع) را مشاهده کند نیز بیش‌تر می‌شود. به طور مشابه، احتمال این‌که دو درخت گسترش داده شده در برنامه‌ریزی دوجهته جایی در بین مسیر یکدیگر را ملاقات کنند نیز بالاتر می‌رود، در نتیجه عامل تحت تأثیر بایاس پاولفی خود را متمایل‌تر به انتخاب مسیر سبز می‌بینند.

[DOI: 10.52547/shefa.9.4.51]

تنها تحت کنترل یادگیری ابزاری هدف‌گرا انجام شده است. این نتایج منطقی کمترین انتظاری است که از الگوریتم پیشنهادی‌مان می‌توانیم داشته باشیم. نتایج شبیه‌سازی الگوریتم بر روی نقشه‌هایی که شامل ایمای پاولفی هستند در تصویر ۳-پ آمده است که نشان می‌دهند سیاست رفتاری عامل به‌گونه‌ای است که احتمال انتخاب عملی که به مسیر سبز منجر می‌شود بیشتر از احتمال انتخاب عملی است که موجب گام نهادن در مسیر قرمز می‌شود. همچنین احتمال گسترش درختان عمیق‌تر در هر دو برنامه‌ریزی رو به جلو و رو به عقب در مسیر سبز بیشتر از مسیر قرمز می‌باشد. نتایج شبیه‌سازی الگوریتم پیشنهادی، با نتایج به دست آمده از تحلیل داده‌های جمع‌آوری شده از شرکت‌کنندگان در آزمایش که در تصویر ۲ نشان داده شده بودند، سازگاری دارد که بر درستی الگوریتم پیشنهادی دلالت می‌کند.

بحث و نتیجه‌گیری

این مقاله بر مبنای دو مفهوم کلیدی برنامه‌ریزی دوجهته و تأثیرات ایمای پاولفی نزدیک‌شوندگی روی آن بنا نهاده شده است. از آن‌جا که مطالعهٔ همزمان آن‌ها کمتر مورد توجه پژوهش‌گران پیشین واقع شده است، برآن شدیم تا با طراحی یک آزمایش مسیریابی بررسی نماییم که آیا افراد از برنامه‌ریزی رو به جلو و رو به عقب به طور همزمان استفاده می‌کنند یا خیر؟ و این‌که آیا این برنامه‌ریزی‌های سه‌گانه توسط ایماهای پاولفی بایاس می‌شوند یا خیر؟ استفاده از برنامه‌ریزی به‌منظور ارزیابی درختان تصمیم جهت یافتن بهترین شاخه‌ها (توالی عمل‌ها)، ضروری به نظر می‌رسد. اما محدودیت منابع شناختی گسترش درختان تصمیم عمیق در دو جهت رو به جلو و رو به عقب را با تنگنا مواجه می‌کند، چرا که رشد درختان تصمیم به صورت‌نمایی افزایش می‌یابد. حال آن‌که اگر درختان تصمیم به صورت همزمان در دو جهت گسترش داده شوند و این دو درخت جایی در بین مسیر یکدیگر را ملاقات کنند، ارزیابی رو به عقب می‌تواند دانش خود را به تخمین‌های رو به جلو منتقل کند و بدین ترتیب به برنامه‌ریزی کارآمدتری منجر شود، چرا که بجای گسترش درخت تصمیم عمیق رو به جلو (یا رو به عقب) درختان کم عمق دوجهته خواهیم داشت که دیگر از رشد نمایی رنج نخواهند برد (۳). این برنامه‌ریزی دوجهته سطحی از کنترل در گسترش درختان تصمیم اعمال می‌کند که فرآیند تصمیم‌گیری را کارآمدتر می‌نماید. تحلیل آماری داده‌های جمع‌آوری شده از آزمایش نشان می‌دهد که شرکت‌کنندگان به انتخاب مسیری که شامل ایمای نزدیک‌شوندگی پاولفی است تمایل بیشتری دارند در مقایسه با مسیری که فاقد آن است. این امر در حالی رخ می‌دهد که طول دو مسیر موجود با هم برابر بوده و از نقطه نظر کنترل ابزاری نباید تفاوتی بین دو مسیر وجود داشته باشد. در نتیجه نیرویی که شرکت‌کننده را به انتخاب مسیر سبز متمایل می‌کند ایماهای نزدیک‌شوندگی پاولفی می‌باشند. همچنین میزان تمایل شرکت‌کنندگان به انتخاب مسیر سبز در نقشه‌های مختلف بسته به جایگاه ایمای پاولفی متفاوت می‌باشد به طوری که میزان بایاس در وضعیتی که ایما تنها در نزدیکی نقطهٔ هدف قرار دارد، نسبت به وضعیتی که در نزدیکی نقطهٔ شروع قرار دارد کمتر، و این میزان بایاس در هر دوی آن‌ها نسبت به زمانی که ایما هم در نزدیکی نقطهٔ شروع و هم در نزدیکی نقطهٔ هدف قرارداد کمتر است. در نتیجه افراد در این آزمایش نه تنها از برنامه‌ریزی رو به جلو، بلکه از برنامه‌ریزی دوجهته نیز استفاده می‌کنند. به‌منظور شبیه‌سازی رفتار شرکت‌کنندگان، الگوریتم برنامه‌ریزی دوجهته‌ای را تحت قالب یادگیری تقویتی مبتنی بر مدل توسعه دادیم و تأثیر نزدیک‌شوندگی پاولفی را به‌عنوان وزن‌های دخیل در سیاست رفتاری سافتمکس مدل نمودیم. این امر به نوبهٔ خود منجر به گسترش درختان تصمیم عمیق‌تر در شاخه‌هایی شد که شامل ایماهای پاولفی بودند. همچنین نتایج شبیه‌سازی با نتایج برآمده از آزمایش مشابه بوده که بر درستی الگوریتم پیشنهادی صحه می‌گذارد.

منابع

1. Simon DA, Daw ND. Neural correlates of forward planning in a spatial decision task in humans. Journal of Neuroscience. 2011; 31(14): 5526-39.

2. Russell SJ, Norvig P. Artificial Intelligence- A Modern Approach, Third Int. Edition. Pearson Education, Upper Saddle River, NJ, USA; 2010.

3. Afsardeir A, Keramati M. Behavioural signatures of backward planning in animals. European Journal of Neuroscience. 2018; 47(5): 479-87.

4. Khamassi M, Girard B. Modeling awake hippocampal reactivations with model- based bidirectional search. Biological Cybernetics (Modeling). 2020.

5. Huys QJ, Eshel N, O'Nions E, Sheridan L, Dayan P, Roiser JP. Bonsai trees in your head: how the pavlovian system sculpts goal- directed choices by pruning decision trees. PLoS computational biology. 2012; 8(3): e1002410.

6. Rescorla RA. Pavlovian conditioning: It's not what you think it is. American psychologist. 1988; 43(3): 151.

7. O'Doherty JP, Cockburn J, Pauli WM. Learning, reward, and decision making. Annual review of psychology. 2017; 68: 73-100.

[DOI: 10.52547/shefa.9.4.51]

8. Mogg K, Field M, Bradley BP. Attentional and approach biases for smoking cues in smokers: an investigation of competing theoretical views of addiction. Psychopharmacology. 2005; 180(2): 333-41.

9. Dayan P, Niv Y, Seymour B, Daw ND. The misbehavior of value and the discipline of the will. Neural networks. 2006; 19(8): 1153-60.

10. Balleine BW, Delgado MR, Hikosaka O. The role of the dorsal striatum in reward and decision- making. Journal of Neuroscience. 2007; 27(31): 8161-5.

11. Cartoni E, Balleine B, Baldassarre G. Appetitive Pavlovian- instrumental transfer: a review. Neuroscience & Biobehavioral Reviews. 2016; 71: 829-48.

12. Lloyd K, Dayan P. Pavlovian- instrumental interactions in active avoidance: The bark of neutral trials. Brain research. 2019; 1713: 52-61.

13. Pool E, Pauli W, Kress C, O'Doherty J. Behavioural evidence for parallel outcome-sensitive and outcome-insensitive Pavlovian learning systems in humans. Nature Human Behaviour, 3 (3), 284-96.

14. Dorfman HM, Gershman SJ. Controllability governs the balance between Pavlovian and instrumental action selection. Nature communications. 2019; 10(1): 1-8.

15. Watson P, De Wit S, Hommel B, Wiers RW. Motivational mechanisms and outcome expectancies underlying the approach bias toward addictive substances. Frontiers in psychology. 2012; 3: 440.

16. Hunt LT, Rutledge RB, Malalasekera WN, Kennerley SW, Dolan RJ. Approach-induced biases in human information sampling. PLoS biology. 2016; 14(11): e2000638.

17. Csifcsák G, Melsæter E, Mittner M. Intermittent absence of control during reinforcement learning interferes with Pavlovian bias in action selection. Journal of Cognitive Neuroscience. 2020; 32(4): 646-63.

18. Gureckis TM, Love BC. Computational reinforcement learning. The Oxford handbook of computational and mathematical psychology. 2015: 99-117.

19. Huys QJ, Cools R, Gölzer M, Friedel E, Heinz A, Dolan RJ, et al. Disentangling the roles of approach, activation and valence in instrumental and pavlovian responding. PLoS computational biology. 2011; 7(4): e1002028.

20. Sutton RS, Barto AG. Reinforcement learning: An introduction: MIT press; 2018.

21. Daw ND, Niv Y, Dayan P. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. Nature neuroscience. 2005; 8(12): 1704-711.

22. Dayan P, Berridge KC. Model-based and model-free Pavlovian reward learning: revaluation, revision, and revelation. Cognitive, Affective, & Behavioral Neuroscience. 2014; 14(2): 473-92.

23. Cushman F, Morris A. Habitual control of goal selection in humans. Proceedings of the National Academy of Sciences. 2015; 112(45): 13817-22.

[DOI: 10.52547/shefa.9.4.51]