# استفاده از آتوماتونهای یادگیری در یادگیری مهارتهای فوتبال روباتها

**محمد رضا میبدی** meybodi@ce.aku.ac.ir
**حمیدرضا نصیری** hnasiri@aeoi.org.ir
**دانشگاه صنعتی امیرکبیر**
**دانشکده مهندسی کامپیوتر**

## چکیده

*آتوماتونهای یادگیری روشی برای آموختن بر اساس نیاز به تصمیم‌گیری سازگارپذیر در محیطهایی با خصوصیت تصادفی و عدم قطعیت بالا می‌باشند. مهارتهای فردی یکی از عوامل موفقیت تیمهای دارنده این بازیکنان می باشند. تمام تیمهای موفق در رقابتهای لیگ شبیه سازی روباتها دارای بازیکنانی با مهارتهای فردی – دریافت توپ ، پاس دادن توپ ، دریبل ، و گل زدن – آگاهانه بوده انـد. در پیاده سازی این مهارتها باید به کارآیی آنها در عمل در محیطی پویا و غیر قطعی، همچون فوتبال روباتها ، توجه شود و بـهترین راه بـرای حل این مسئله استفاده از یادگیری برای این مهارتها می باشد*یکی از مهمترین چالشها در سیستمهای چند عامله مسئله هماهنگی و بهینه سازی آن می باشد. عوامل یک تیم برای هماهنگی باید به گونه ای در اجرای یک طرح مشترک با یکدیگر توافق و اقدام نمایند. راه حل این امر در شکل طبیعی از طریق تبادل دانش فردی عوامل (درک وضعیت فعلی، طرح مورد نظر) باید انجام شود. اما تعامل امری پر هزینـه و نامطمئن در محیطهای پیچیده سیستمهای چند عامله است. یکی از مهمترین کاربردهای یادگیری در سیستمهای چند عاملـه کمـک بـه بهینه سازی هماهنگی و کاهش هزینه های آن می باشد. در این مقاله به ارائه کاربرد *آتوماتونهای یادگیری در پیاده سازی مهارتهای پایه و پیچیده ، مثل دریافت توپ و دریبل ، برای عامل بازیگر فوتبال پرداخته ایم.* همچنین به ارائه روشی برای حـل مسـئله همـاهنگی عوامـل فوتبالیست برای یادگیری رفتاری تیمی،اتخاذ سیاست بهینه دفاعی در برابر حمله تیم رقیب، می پردازیم.

**واژه های کلیدی** : آتوماتون یادگیری، فوتبال روباتها، مهارت تیمی، دریبل، دریافت توپ، سیاست دفاعی، عامل، سیستم چند عامله

## ۱–مقدمه

لیگ شبیه سازی جام روباتها از سال ۱۹۹۶ میلادی آغاز شد و از آن تاریخ تا کنون هر ساله در کنار سایر لیگهای ربوکاپ برگزار گردیده است . این لیگ برای محققینی است که ممکن است منابع سـاخت روباتـها واقعـی را نداشته باشند اما بشدت مشتاق پیامدهای یادگیری و استدلال چند عاملـه پیچیـده هسـتند . تیمـهای لیـگ شـبیه سازی با سه چالش تحقیقاتی استراتژیک روبرو بوده اند: یادگـیری چند عامله،کارتیمی،مدلسازی عامل [2].

فرمانهای پایه که در اختیار یک عامل بازیگر فوتبال قرار دارند برای برآوردن اهداف او کفـایت نمـی کننـد اول بدلیل اینکه تاثیر این فرمانها در محیط قطعی نیست و در اعمال آنها باید عدم قطعیت و اغتشـاش محیـط را در نظر گرفت و دوم بدلیل اینکه حضور عوامل دیگر ، همکار یا رقیب ، در رسیدن عامل بازیگر به هدفش موثر است. به همین دلیل تیمهای شرکت کننده در جام روباتها پیش از کـار روی طرحـها و اسـتراتژیهای تیمـی مجبـور بـه حـل مسئله مهارتهای فردی عوامل هستند. برخی از این مهارتها در حد تنظیم پارامترهای یک فرمان خاص بـرای بـازیگر می باشند که می توان آنها را مهارتهای پایه نامید و معمولا برای بازیگران همه تیمـها ضـروری مـی باشـند. برخـی دیگر از مهارتها مجموعه ای از فرمانها و مهارتهای پایه را با ترتیبی از پیش طراحی شده به کار می گیرند تـا هدفـی خاص را برای یک عامل برآورده سازند این مهارتها را می توان مهارتهای پیچیده نامید . در دست داشتن آنـها بـرای بازیگران یک تیم یک مزیت است نه ضرورت و برخی تیمها روی آنها کار می کنند تا در طراحی و اجرای سیاستهای چند عامله امکانات بیشتری داشته باشند . طراحی و پیاده سازی مهارتهای پیچیده برای عامل فوتبالیست وظیفه ای سخت است و علت آن علاوه بر عدم قطعیت محیط و دخالت عوامل دیگر ، جنبه ها و حالتهایی است که تغییر آنـها

در عملکرد مهارتی بخصوص تاثیر دارد . در طراحی یک مهارت پیچیده تا حد ممکن باید مهارت را مستقل از جنبه ها و حالتهای خاص تعریف کرد در غیر اینصورت باید برای حالتهای نامتناهی یا زیاد ، که در عملکرد یک مهارت تاثیر گذارند ، فکری کرد زیرا نه می توان همه آنها را پیش بینی کرد و نه در کد گنجاند . یک راه برای حل مسئله حالتهای پیش بینی نشده استفاده از روشهای یادگیری می باشد . اکثر تیمها از روش یادگیری نظارتی [7] و یادگیری تقویتی [3] به عنوان روش یادگیری و آموزش عامل خود استفاده کرده اند.

یکی از پیچیده ترین مسائل در سیستمهای چند عامله مسئله هماهنگی و بهینه سازی آن می باشد [4]. عوامل یک تیم برای هماهنگی باید به گونه ای در اجرای یک طرح مشترک با یکدیگر توافق و اقدام نمایند. راه حل این امر در شکل طبیعی از طریق تبادل دانش فردی عوامل (درک وضعیت فعلی، طرح مورد نظر) باید انجام شود. اما تعامل امری پر هزینه و نامطمئن در محیطهای پیچیده سیستمهای چند عامله است. یکی از مهمترین کاربردهای یادگیری در سیستمهای چند عامله کمک به بهینه سازی هماهنگی و کاهش هزینه های آن می باشد.

در معماریهای لایه ای عوامل بازیگر فوتبال [3] رفتار عامل با توجه به تصمیم گیری در لایه های مختلف آن تعیین می شود. به این معنی که در مقام اتخاذ هر وظیفه پیچیده وظیفه را به شکل پایین به بالا تجزیه نموده و اجرا می نماید. همانطور که در انجام مهارتهای فردی نیازمند به آموزش و داشتن روشهای هوشمند هستیم در مهارتهای گروهی نیز داشتن روشهایی برای بهبود و سازگاری رفتار و مهارتهای گروهی عامل در بهبود کارایی تیم بسیار موثر است.

یک چالش مهم در ارزیابی رفتارهای گروهی مسئله ارزیابی رفتار بلند مدت است [3]. به طور مثال در ارزیابی یک سیاست دفاعی به کار گرفته شده قضاوت به محض انجام عمل میسر نیست و در مجموع یک سلسله از اعمال نهایتا با حصول به هدفی خاص یا عدم حصول آن هدف مجموع کل اعمال انجام شده موفق یا ناموفق ارزیابی می شود. علاوه بر مسئله ارزیابی عملکرد در بلند مدت چالش بزرگتر پیش رو مسئله نگاشت مسئولیت پاداش/جریمه به اعمال جزء سیاست اتخاذ شده است.

در مسئله یافتن سیاست بهتر، که نمونه ای از رفتارهای تیمی است، مسائل اساسی سیستم های چند عامله بیشتر مطرح می شود. مسائل همکاری، ارتباط، و تعامل بین عوامل از مسائلی هستند که طراح سیستم باید برای دست یابی به رفتار بهینه آنها را حل کند. هدف و اولویت اصلی در اینجا بهبود رفتار تیمی عوامل یک سیستم چند عامله، فوتبال روباتها در اینجا، می باشد.

در این مقاله ما به ارائه امکان کاربرد آتوماتونهای یادگیری قطعی به منظور یادگیری رفتاری تیمی، اتخاذ سیاست دفاعی، می پردازیم. روش ما را به دلیل استفاده از آتوماتون یادگیری ، روشی جدید و ابتکاری می توان به حساب آورد. کاهش هزینه های ارتباط و افزایش اطمینان در هماهنگی از خصوصیات برجسته این روش می باشند .

## ۲- آتوماتون یادگیری[4]

وضعیت کیفی آتوماتون یادگیرنده (LA) ممکن است به صورت زیر بیان شود : تعدادی متناهی از اعمال می‌توانند در محیطی اتفاقی انجام شوند ، وقتیکه یک عمل مشخص انجام شد محیط یک پاسخ اتفاقی را فراهم می‌کند که ممکن است مطلوب یا نامطلوب باشد . هدف در طراحی آتوماتون یادگیرنده تعیین این است که چطور انتخاب یک عمل در هر مرحله به وسیله اعمال و پاسخ‌های قبلی باید جهت داده شود (یا راهنمائی شود) . نکته مهم در اینجاست که تصمیمات باید با دانشی ناچیز با توجه به طبیعت محیط گرفته شود . به خاطر خصوصیات متغیر محیط و یا قرار گرفتن در ساختار سلسله مراتبی در هر حال آتوماتون یادگیری باید به گونه‌ای طراحی شده باشد که کارائی عملگری کلی را بهبود بخشد.

یک آتوماتون یک پنج تائی $\{\phi,\alpha,\beta,F(0/0),H(0/0)\}$ است که درآن $\phi$ یک مجموعه از وضعیت‌های داخلی است، $\beta$ یک مجموعه از اعمال ورودی، و $\alpha$ یک مجموعه از خروجی‌ها می باشد همچنین داریم :

تابع وضعیت بعدی : $F(0,0):\phi * \beta \to \phi$

تابع عمل بعدی : $H(0,0):\phi * \beta \to \alpha$

یک آتوماتون را متناهی گویند اگر مجموعه های مذکور متناهی باشند . یک آتوماتون متناهی و state- output را آتوماتون سریالی اتفاقی (stochastic) نیز گویند .آنها بطور دقیق با مولفه‌های زیر قابل بیان می‌باشند :

وضعیت در هر نمونه n : $\phi(n)$ — (i) $\phi = \{\phi_1,\phi_2,...,\phi_s\}$

خروجی در هر نمونه n : $\alpha(n)$ — (ii) $\alpha = \{\alpha_1,\alpha_2,...,\alpha_s\}$

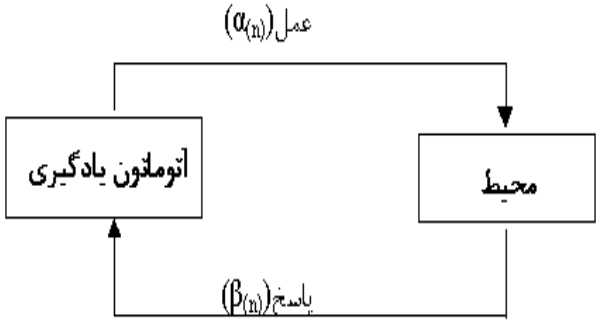
ورودی در نمونه n : $\beta(n)$ — (iii) $\beta = \{\beta_1,\beta_2,...,\beta_m\}$ یا $\beta = \{(a,b)\}$

F می‌تواند قطعی یا تصادفی باشد. — (iv) $\phi(n+1) = F[\phi(n),\beta(n)]$

G می‌تواند قطعی یا تصادفی باشد. — (v) $\alpha(n) = G[\phi(n)]$

آتوماتون را قطعی گویند اگر هر دوی H و F قطعی باشند .در غیر اینصورت آتوماتون را تصادفی گویند.

شکل ۱ نمایی کلی از یک آتوماتون یادگیری را نشان می دهد. شروع و ترتیب عمل آتوماتون به صورت $\beta_{(0)} \leftarrow \alpha_{(0)} \leftarrow \phi_{(0)}$ می باشد . یعنی آتوماتون ابتدا تعیین وضعیت شده سپس به انجام عمل مبادرت می نماید و پس از آن پاسخ محیط را دریافت می کند.




**شکل ۱: اتصال بازخوردی آتوماتون یادگیری و محیط**

## ۲-۱-انواع آتوماتونهای یادگیری قطعی

آتوماتون $L_{2N,2}$ دارای 2N وضعیت، N وضعیت برای هر عمل، و دو عمل است. اصطلاحا می گویند این آتوماتون دارای حافظه به عمق N است. تابع وضعیت بعدی در شکل ۲ به طور صوری توضیح داده شده است (پاسخ نامطلوب با $\beta$ =1 نشان داده شده است). تابع عمل بعدی نیز به صورت زیر تعریف می شود :

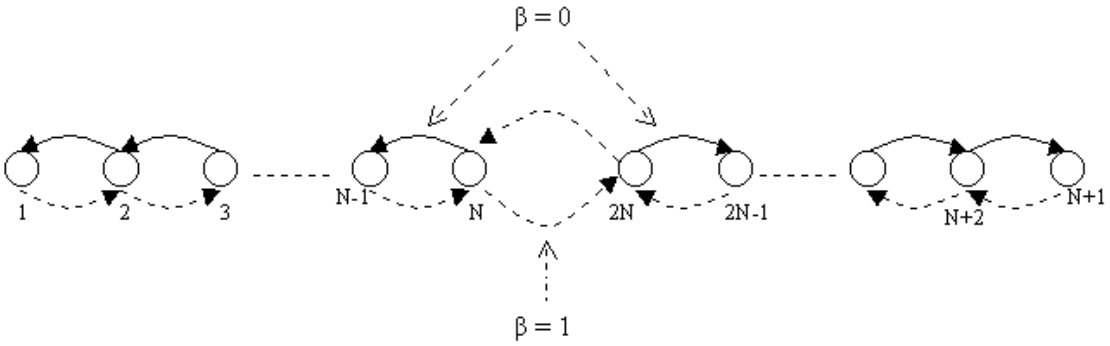
when $1 \le i \le N : G[\varphi_i] = \alpha_1$

when $N < i \le 2N : G[\varphi_i] = \alpha_2$

آتوماتون $G_{2N,2}$ از نظر تعداد وضعیتها مثل آتوماتون $L_{2N,2}$ ، دارای 2N وضعیت، می باشد ولی تابع وضعیت بعدی آن این فرق را دارد که که در صورت دریافت پاسخ نامطلوب از $\varphi_N$ به $\varphi_{N+1}$ می رویم و از $\varphi_{2N}$ به $\varphi_1$ می رویم (شکل ۳).

اما تابع عمل بعدی آتوماتون $G_{2N,2}$ مثل آتوماتون $L_{2N,2}$ تعریف شده است.

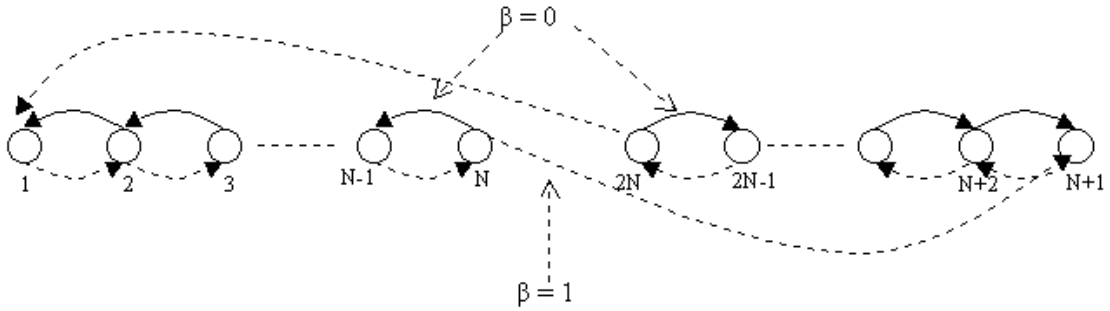
در آتوماتون **Krinsky (K1)** مجدداً تابع وضعیت بعدی به صورت نشان داده شده در شکل ۴ نمایش داده می شود و با آتوماتون $L_{2N,2}$ متفاوت است ولی تابع عمل بعدی همان تعریف ارائه شده برای آتوماتون $L_{2N,2}$ را دارد.




شکل ۲ : آتوماتون $L_{2N,2}$

برای اینکه بتوان از آتوماتون یادگیری برای حل یک مسئله یادگیری در یک سیستم استفاده کرد لازم است که طبیعت مسئله دارای خصوصیات فرآیند های مارکف باشد . کاری که طراح باید انجام دهد این است که قدمهای زیر را در رابطه با مسئله انجام دهد :

- انتخاب مدل نمادین یا تحلیلی برای بازنمایی سیستم یا مسئله.
- تعیین طبیعت اطلاعاتی که باید بدست آورده شود.
- تعیین ساختار فضای تصمیم گیری.
- تعیین ملاک کارآیی که باید ، توسط آتوماتون یادگیری ، بهینه شود.




شکل ۳ : آتوماتون $G_{2N,2}$

یک فرآیند زنجیره ای مارکف(MDP) عبارتست از:

یک مجموعه از وضعیتهای نمایش داده شده با S.

مجموعه ای از اعمال A .

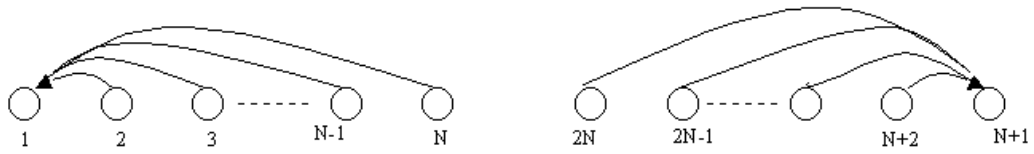
یک تابع پاداش $R : S \times A \rightarrow R$ .

یک تابع انتقال $T : S \times A \rightarrow \pi(S)$ هرعضو $\pi(S)$ یک مقدار احتمالی روی اعضای S است. و از نماد برای نمایش احتمال انتقال از S به با عمل a استفاده می شود.

فرآیند مارکف است اگر تابع انتقال مستقل از وضعیتها و اعمال قبلی باشد .

$\beta = 1$ :


$\beta = 0$ :




**شکل ۴ : آتوماتون Krinsky (K1)**

## ۲-۲-آتوماتونهای تصادفی

آتوماتونهای تصادفی یا غیر قطعی نمونه ای از طرحهای تقویتی[3] عمومی می باشند . در اینگونه آتوماتونها عمل بعدی بر اساس بردار احتمال اعمال انتخاب می شود ، نه بر اساس وضعیت آتوماتون ، و با دریافت پاسخ محیط بردار احتمال عمل به روز رسانی شده و عمل بعدی بر اساس بردار جدید انتخاب می گردد.

معادله عمومی برای به روز رسانی بردار احتمال عمل طرح تقویتی به صورت زیر بیان میشود :

If $\alpha(n) = \alpha_i$

When $\beta=0$ :

$$P_j(n+1) = (1-a) \cdot p_j(n) \quad , \forall\ j \# i$$

$$P_i(n+1) = p_i(n) + a\,[\,1 - p_i(n)\,]$$

When $\beta=1$ :

$$P_j(n+1) = b/(r-1) + (1-b) \cdot p_j(n) \quad , \forall\ j \# i$$

$$P_i(n+1) = (1-b) \cdot p_i(n)$$

$0 < a , b < 1$

این آتوماتونها را به طور کلی $L_{RP}$ می گویند.اگر $b = 0$ بگیریم ، آتوماتون را $L_{RI}$ می گویند.

# ۳-دریافت توپ

در این قسمت به ارائه امکان کاربرد آتوماتونهای یادگیری به عنوان روشی برای حل مسئله سازگاری در محیطی غیرقطعی و پویا نظیر فوتبال روباتها اقدام نموده ایم و یادگیری مهارتی فردی ، همچون دریافت توپ در حال حرکت را ، به وسیله آتوماتون یادگیری بررسی نموده ایم . روش ما را به دلیل استفاده از آتوماتون یادگیری ، روشی جدید و ابتکاری می توان به حساب آورد . ما از یک آتوماتون یادگیری برای یادگیری مهارت دریافت توپ در حال حرکت استفاده کرده ایم . این آتوماتون جهت مناسب را برای حرکت به سمت توپ در حال حرکت به سمت عامل را تعیین می کند . خصوصیت روش ما این است که عامل فوتبالیست را ابتدا با مقدار مناسبی آموزش غیر برخط آماده می کنیم در حالیکه عامل به طور بر خط و در زمان بازی واقعی نیز همچنان به بهبود کارآیی خود و سازگاری با تغییرات محیط قادر است .

## ۳-۱-کارهای قبلی

برای پیاده سازی مهارتهای فردی اکثراً روشهای ابتکاری و برخی یادگیری نظارتی را به صورت غیر بر خط به کار برده اند . نمونه ای از مهارتهای فردی حائز اهمیت مهارت دریافت توپ در حال حرکت می باشد که بر خلاف

تصور مهارتی پیچیده و مهم است [6] . در این زمینه اقای پیتر استون در تز دکترای خود به حل این مسئله پایه و اساسی برای عامل فوتبالیست پرداخته اند . روش استفاده شده توسط ایشان استفاده از شبکه عصبی و آموزش تجربی و غیر بر خط عامل برای دریافت توپ بوده است . ایشان خاطر نشان کرده اند که روش ایشان یک روش ابتکاری می باشد و ممکن است از روشهای دیگری نیز بتوان برای حل این مسئله استفاده کرد [2,6].

## ۳-۲-مسئله دریافت توپ

مسئله دریافت توپ در حالت عمومی عبارت از حرکت به سمت توپ در حال حرکت و در اختیار گرفتن توپ می باشد . ما در اینجا به تعریف مسئله از دید سیستم آتوماتون یادگیری می پردازیم .

در این مسئله محیط آتوماتون را خدمتگزار فوتبال تشکیل می دهد . اعمال آتوماتون در این محیط تاثیر مستقیم دارد ولی با یک واسطه ، که عامل باشد . به این معنی که عامل از بین تعداد معینی از اعمال یکی را انتخاب می کند بدون اینکه بداند اصلا ماهیت عمل چیست . در واقع عامل انتخاب آتوماتون را تفسیر و تبدیل به عملی با معنی در محیط خدمتگزار فوتبال می کند و سپس در انتهای هر سعی پس از ارزیابی عمل با معیار تعریف شده خودش ، فقط آتوماتون را از مطلوب یا نامطلوب بودن عمل آگاه می سازد .

در سناریوی یادگیری از یک عامل منتقد نیز استفاده کرده ایم که شروع ، خاتمه ، و نتیجه عمل را در محیط اعلان می کند . عامل ما در هر بار دریافت اطلاعات دیداری از محیط مقادیر فاصله توپ از عامل ($BallDist_t$) ، و زاویه توپ با عامل ($BallDir_t$) را به عنوان اطلاعات ورودی مطلوب و مورد نیاز دریافت می کند و از طریق ارزیابی این اطلاعات وظیفه بعدی را تعیین می کند . در نهایت عامل منتقد با در نظر گرفتن معیارهای موفقیت از پیش تعریف شده برای آن موفقیت یا عدم موفقیت را اعلان می کند و عامل حاصل این ارزیابی را به آتوماتون یادگیری اعلان می دارد .

ساختار فضای تصمیم گیری ما ساختاری پیوسته است ولی با تدبیری که در اینجا به کار برده ایم فضای تصمیم گیری را به صورت تابعی گسسته و متناهی از اعداد صحیح در آورده ایم . اعمال واقعی ، عملی که عامل در موضوع حاضر می خواهد یادگیری کند ، عبارت است از زاویه مناسب برای چرخش در دستور (turn direction) به منظور دریافت توپ در حال حرکت به طرف عامل . از آنجاییکه زاویه چرخش مناسب مقداری پیوسته می باشد ما این فاصله را به بازه هایی با طول یکسان تقسیم کرده ایم تا فضای اعمال یا جواب یک فضای گسسته شود .

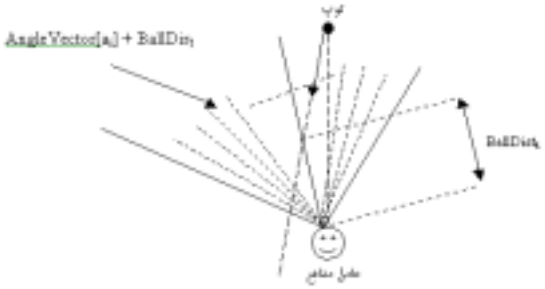
آتوماتون، شاخص (اندیس)، عملی را از بین تعدادی متناهی از اعمال انتخاب و این انتخاب به آرایه ای از بازه های حقیقی نگاشت می شود و عامل در آن بازه مقداری را به تصادف انتخاب و به کار می گیرد . هر چه طول بازه ها کمتر باشد دقت بیشتر ولی سربار هم بیشتر می شود . در این مسئله عمل مطلوب ، در عمل و زمان بازی واقعی ، نمی تواند هر مقداری باشد بلکه یک یا چند بازه در واقع جواب مسئله هستند بنابراین پس از یادگیری انتظار ما این است که احتمال اتخاذ برخی از بازه ها در حد ناممکن باشد (شکل ۵).

در بازی واقعی از نتایج حاصله از یادگیری برای مقدار دهی اولیه آتوماتون استفاده می کنیم و همانطور که ذکر شد اتخاذ برخی اعمال در آن هنگام تقریبا ممکن نخواهد بود و نباید باشد. این مطلب سرنخی برای انتخاب آتوماتون مناسب برای حل این مسئله یادگیری به ما می دهد .

## ۳-۳-آموختن دریافت توپ

در هر تکرار از جلسه آموزش دریافت توپ ابتدا عامل مربی ، توپ و شوت کننده را در فاصله ۲۰ الی ۳۰ واحدی عامل دریافت کننده توپ قرار می دهد و مدافع را در فاصله ۴ واحدی جلو دروازه و رو به وسط زمین قرار می دهد و سپس به عوامل شروع تکرار بعدی را اعلان می کند.

با دریافت اعلان شروع از طرف عامل مربی ، عامل شوت کننده توپ را با حداکثر قدرت و با انحراف زاویه خواسته شده از آن به طرف دروازه شوت می کند . عامل مدافع توپ را تا رسیدن به محدوده مشخص فقط مشا هده و به طرف آن رو می گرداند و پس از رسیدن به محدوده مشخص شده ، عامل یک بازه از بردار فاصله های زاویه ای را با مشورت با آتوماتون یادگیری انتخاب و به آخرین جهت توپ که مشاهده شده اضافه و سپس حرکت می کند. سپس به طرف توپ رو گردانده و به توپ در جهت مناسب برای دفع آن ضربه می زند .

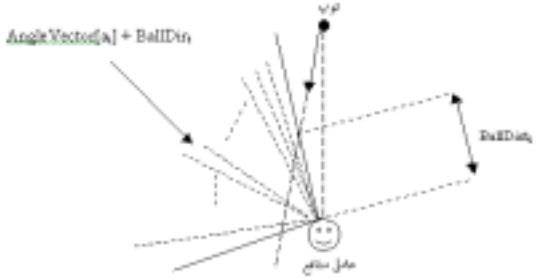



**شکل ۵ : نمونه ای عمومی از وضعیت دریافت توپ**

ما دو روش برای تبدیل عمل آتوماتون به زاویه چرخش مناسب را در پیاده سازی عامل به کار گرفته ایم : یک روش معادلی از روش تجربی با شبکه های عصبی است که در آن زاویه انتخابی برای چرخش ، که توسط آتوماتون انتخاب شده ، بین ۴۵- تا ۴۵ درجه می باشد (شکل ۶) . در روش دیگر زاویه انتخابی برای چرخش ، که توسط آتوماتون انتخاب می شود ، بین ۰ الی ۹۰ درجه می باشد که معادل روش تحلیلی می باشد(شکل ۶).

با تفاسیر ذکر شده ما از آتوماتون مناسب آموزش عامل یک آتوماتون P-Model و از نوع آتوماتون های $L_{RP}$ می باشد . تعداد اعمال آتوماتون به وسیله عامل اعلان می شود و در حین تنظیم عامل برای یادگیری به طور پویا قابل تنظیم است . در زمان یادگیری بردار احتمال اعمال آتوماتون در شروع با مقادیر برابر مقداردهی می شود این بردار در طول یادگیری به روز رسانی شده و بردار نهایی ثبت می شود . در زمان بازی عامل از بهترین بردار ثبت شده برای مقداردهی اولیه آتوماتون داخلی خود استفاده می کند.

## ۳-۴-نتایج




**شکل۶ : وضعیت دریافت توپ معادل روش تحلیلی**

ما عامل را با روشها و پیکره بندیهای متفاوتی از آتوماتون در محیط تحت یادگیری قرار دادیم که در این قسمت نتایج این آزمایشات ارائه گردیده است . در محاسبات انجام شده از سیستم های کامپیوتر شخصی Pentium III 500 با ۲۵۶ مگابایت حافظه ، استفاده شده است .

برای محاسبه هر سلول از جداول ارائه شده در این بخش با استفاده از یک سیستم از نوع ذکر شده ۲۰ الی ۲۲ دقیقه زمان صرف می شود . توجه داشته باشید که کلیه آزمایشهای این بخش با همان سناریوی ذکر شده درباره موقعیت عوامل که توضیح داده شد انجام شده است .

### ۳-۴-۱- کارآیی $L_{RI}$

در این قسمت نتایج کارایی عامل دریافت کننده توپ، مجهز به آتوماتون $L_{RI}$ ارائه گردیده است . ما این آتوماتون را با تعداد اعمال متفاوت ، مقادیر متفاوت پاداش ، و در دو بازه ۰ الی ۹۰ درجه و ۴۵- تا ۴۵ درجه تحت آموزش قرار دادیم و در نهایت بهترین تعداد اعمال در هر دو بازه ۴۰ و با پاداش ۰/۱ بود. در جدول۱ کارآیی بهترین آتوماتونهای $L_{RI}$ یافته شده در تکرارهای متفاوت ارائه گردیده است جدول ۲ کارآیی همین آتوماتونها ، آموزش دیده با بهترین بردارهای مربوطه در جدول ۱ ، را در محیط خدمتگزار فوتبال با مقادیر متفاوت اغتشاش ارائه

می کنند:

**جدول۱ : کارآیی دریافت توپ با آتوماتون $L_{RI}$با پاداش ۰/۱ در تکراهای متفاوت**

| تعداد تکرار / بازه چرخش | ۱۰۰ | ۲۰۰ | ۳۰۰ | ۴۰۰ | ۵۰۰ |
|---|---|---|---|---|---|
| ۰ الی ۹۰ درجه | ٪۸۶ | ٪۹۱ | ٪۹۰ | ٪۹۰ | ٪۹۰ |
| ۴۵- تا ۴۵ درجه | ٪۸۵ | ٪۹۴ | ٪۹۱ | ٪۹۳ | ٪۹۰ |

**جدول ۲ : کارآیی $L_{RI}$ در دریافت توپ با مقادیر متفاوت اغتشاش**

| اغتشاش / بازه چرخش | ۰ | ۰/۰۵ | ۰/۰۶ | ۰/۰۷ | ۰/۰۸ | ۰/۰۹ | ۰/۱ |
|---|---|---|---|---|---|---|---|
| ۰ الی ۹۰ درجه | ٪۹۳ | ٪۹۱ | ٪۸۰ | ٪۷۲ | ٪۷۰ | ٪۷۰ | ٪۷۸ |
| ۴۵- تا ۴۵ درجه | ٪۹۶ | ٪۹۴ | ٪۸۲ | ٪۷۴ | ٪۷۳ | ٪۷۴ | ٪۸۰ |

### ۳-۴-۲- کارآیی $L_{RP}$

در این قسمت نتایج کارایی عامل دریافت کننده توپ، مجهز به آتوماتون $L_{RP}$ ارائه گردیده است . ما این آتوماتون را با تعداد اعمال متفاوت ، مقادیر متفاوت پاداش ، و در دو بازه ۰ الی ۹۰ درجه و ۴۵- تا ۴۵ درجه تحت آموزش قرار دادیم و در نهایت بهترین تعداد اعمال در هر دو بازه ۱۰ و با پارامترهای (پاداش،جزا) =(۰/۲،۰/۱) بود. در جدول۳ کارآیی بهترین آتوماتونهای $L_{RP}$ یافته شده در تکرارهای متفاوت ارائه گردیده است جدول ۴ کارآیی همین آتوماتونها ، آموزش دیده با بهترین بردارهای مربوطه در جدول ۳ ، را در محیط خدمتگزار فوتبال با مقادیر متفاوت اغتشاش ارائه می کنند:

**جدول۳ : کارآیی دریافت توپ با آتوماتون $L_{RP}$ با پارامتر (پاداش،جزا)=(۰/۲،۰/۱) در تکراهای متفاوت**

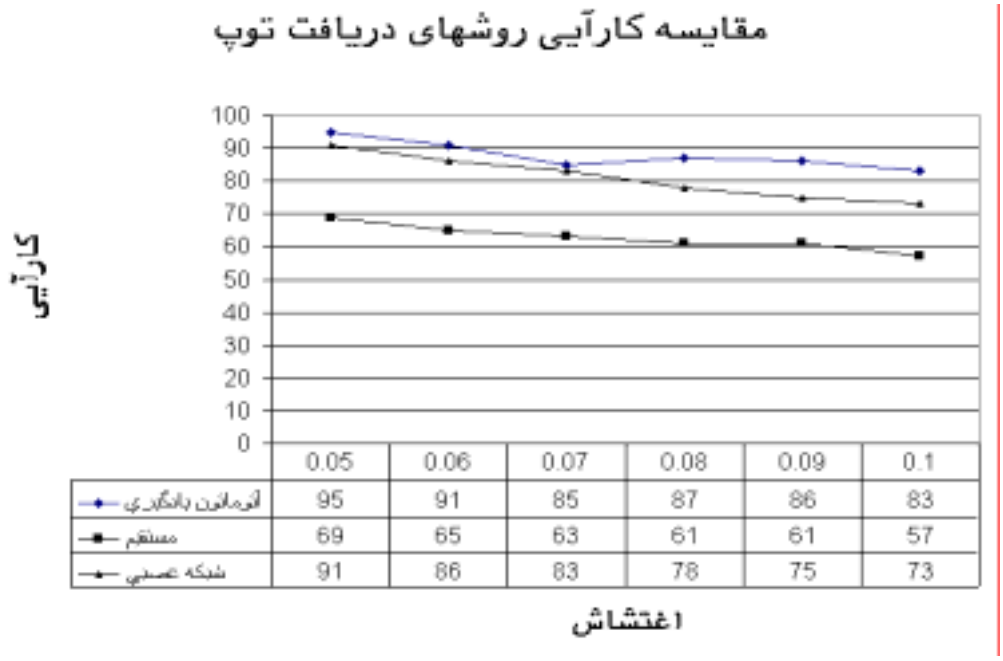
| تعداد تکرار / بازه چرخش | ۱۰۰ | ۲۰۰ | ۳۰۰ | ۴۰۰ | ۵۰۰ |
|---|---|---|---|---|---|
| ۰ الی ۹۰ درجه | ٪۹۰ | ٪۸۵ | ٪۸۷ | ٪۹۱ | ٪۹۵ |
| ۴۵- تا ۴۵ درجه | ٪۹۳ | ٪۸۶ | ٪۹۲ | ٪۸۵ | ٪۹۰ |

**جدول ۴ : کارآیی $L_{RP}$ در دریافت توپ با مقادیر متفاوت اغتشاش**

| اغتشاش / بازه چرخش | ۰ | ۰/۰۵ | ۰/۰۶ | ۰/۰۷ | ۰/۰۸ | ۰/۰۹ | ۰/۱ |
|---|---|---|---|---|---|---|---|
| ۰ تا ۹۰ درجه | ٪۹۶ | ٪۹۵ | ٪۹۱ | ٪۸۵ | ٪۸۷ | ٪۸۶ | ٪۸۳ |
| ۴۵- تا ۴۵ درجه | ٪۹۶ | ٪۹۳ | ٪۸۸ | ٪۸۴ | ٪۸۵ | ٪۸۳ | ٪۸۱ |

## ۳-۵-نتیجه گیری

در این مقاله درباره مهارت دریافت توپ و اهمیت آن ، به عنوان مهارتی که در موفقیت سیاستهای پیچیده نقش اساسی دارد ، توضیح داده شد و بـه بررسـی مـهارت دریـافت تـوپ در فوتبـال روباتـها ، لیـگ شـبیه سـازی ، پرداختیم . ابتدا مسئله دریافت توپ را به صورتی مشخص که قـابل تحلیـل باشـد تعریـف نمودیـم . و سـپس مـدل سیستم آتوماتون یادگیری متناسب با تعریف ارائه شده را طراحی نمودیم . توضیح داده شد که بر طبق مدل تعریف شده از مسئله دریافت توپ ما باید پارامتر جهت را در دستور (turn direction) تعییـن و بهینـه سـازی نمـاییم . هدف ما نشان دادن قابلیت به کارگیری آتوماتونها یـادگیـری در حـل و پیـاده سـازی مسـائل یـادگیـری و پیچیـده سیستمهای چند عامله بود و نتایج حاصله ضمن اثبات این امکان از نظر کارآیی نیز دلگرم کننده بودند . در شکل ۴ نمودار مقایسه سه روش آتوماتون یادگیری ، روش مستقیم ، و شبکه عصبی ارائه گردیده است .




**شکل ۴ : مقایسه کارآیی دریافت توپ با روشهای مختلف یادگیری و مقادیر متفاوت اغتشاش**

از بین دو آتوماتون $L_{RP}$ و $L_{RI}$ هر چند که در آمار های گرفته شده می توان آتومـاتون هـایی بـا کـارآیی یکسان یافت اما به دو دلیل مهم آتوماتون بهینه $L_{RP}$ ، که همان آتومـاتون $L_{RP}$ بـا جفـت پـارامتر (پـاداش،جـزا) = (۰/۱،۰/۲) می باشد ، از آتوماتون $L_{RI}$ مناسبتر است : اول اینکه آتوماتون $L_{RP}$ قدرت تغییر رفتـار خـود بـا افزایـش تغییرات پیش بینی نشده (مثلا میزان باد) و یا اغتشاش را دارد و این بـه خـاطر پـارامتر جـزای آن مـی باشـد ولـی آتوماتون $L_{RI}$ نمی تواند عمل یا اعمال بهینه خود را به عمل جدیدی که احتمال انتخاب ان صفر اسـت تغییـر دهـد مگر اینکه از بردار احتمالی برای مقدار دهی آغازین آن استفاده کنیم که این محدودیت را ندارد. دومیـن دلیـل ایـن است که آتوماتون $L_{RP}$ یافته شده دارای ۱۰ عمل می باشد و به روز رسانی آن از آتوماتون $L_{RI}$ با ۴۰ عمل سریعتر و مستلزم محاسبات کمتر است .

## ۴-یادگیری مهارت چند عامله

مهارتهای یا رفتارهای چند عامله از منظر گروه [7] مهارتهایی هستند در حصول به نتیجه چندین عامل، همکار یا رقیب، درگیر هستند. از منظر عامل منفرد فوتبالیست برخی از مهارتها مجموعه ای از فرمانها و مهارتهای پایه را با ترتیبی از پیش طراحی شده به کار می گیرند تا هدفی خاص را برای یک عامل برآورده سازند این مهارتها را می توان مهارتهای پیچیده نامید . در دست داشتن آنها برای بازیگران یک تیم یک مزیت است نه ضرورت و برخی تیمها روی آنها کار می کنند تا در طراحی و اجرای سیاستهای چند عامله امکانات بیشتری داشته باشند . طراحی و پیاده سازی مهارتهای پیچیده برای عامل فوتبالیست وظیفه ای سخت است و علت آن علاوه بر عدم قطعیت محیط و دخالت عوامل دیگر ، جنبه ها و حالتهایی است که تغییر آنها در عملکرد مهارتی بخصوص تاثیر دارد . در طراحی یک مهارت پیچیده تا حد ممکن باید مهارت را مستقل از جنبه ها و حالتهای خاص تعریف کرد در غیر اینصورت باید برای حالتهای نامتناهی یا زیاد ، که در عملکرد یک مهارت تاثیر گذارند ، فکری کرد زیرا نه می توان همه آنها را پیش بینی کرد و نه در کد گنجاند . یک راه برای حل مسئله حالتهای پیش بینی نشده استفاده از روشهای یادگیری می باشد . اکثر تیمها از روش یادگیری نظارتی [6] و یادگیری تقویتی [3] به عنوان روش یادگیری و آموزش عامل خود استفاده کرده اند .

در این قسمت ما به ارائه کاربرد آتوماتونهای یادگیری به عنوان روشی برای یادگیری مهارتی پیچیده ، دریبل ، می پردازیم. مسئله دریبل از منظر گروه یک مسئله چند عامله است زیرا در موفقیت یا عدم موفقیت بیش از یک عامل نقش دارد. ما از دو آتوماتون یادگیری برای یادگیری مهارت دریبل استفاده کرده ایم . این دو آتوماتون به طور همزمان عمل کرده و مقدار قدرت و جهت مناسب را برای انجام عمل دریبل تعیین می کنند . خصوصیت دیگر روش ما این است که عامل فوتبالیست را ابتدا با مقدار مناسبی آموزش غیر برخط آماده می کنیم در عین حالیکه عامل به طور بر خط و در زمان بازی واقعی نیز همچنان به بهبود کارآیی خود و سازگاری با تغییرات محیط قادر است .

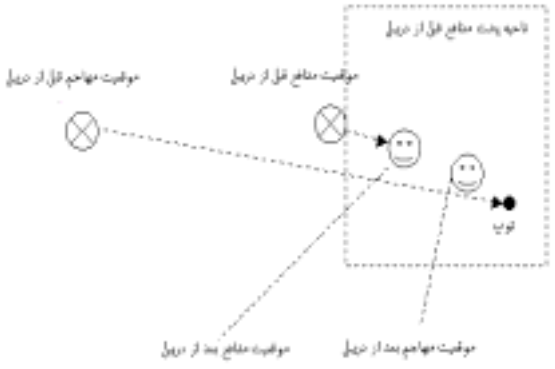
### ۴-۲-مسئله دریبل

دریبل کردن یک مهارت پیچیده است که در پیاده سازی مهارتهای گروهی دیگر می تواند نقش موثری داشته باشد . هنگامیکه یک عامل در موقعیتی از بین انتخاب های پاس، شوت،ویا عبور از حریف آخری را انتخاب می کند قابلیت و کارایی دریبل قابلیتی تعیین کننده است ، مهارتی پیچیده که در فوتبال انسانی نیز از عهده هر بازیکنی برنمی آید. دریبل قابلیتی نسبی است . بدین معنی که ممکن است شیوه دریبل پیاده سازی شده برای عامل ما در برابر بازیکنان یک تیم موفق باشد و در برابر بازیکنان تیم دیگری ناموفق باشد (همچون فوتبال واقعی).

اگر نگاهی دقیق به مهارت دریبل در فوتبال انسانها داشته باشیم – و یا اگر حتی بهتر ، خودمان تجربه ای در این زمینه داشته باشیم – مشاهده می شود که در ساده ترن حالت دریبل یک حرکت سریع در مقابل بازیکن حریف است که بازیگر توپ را ناگهان به مسیر یا نقطه ای خارج از دسترس بازیکن رقیب می رساند . اگر این ایده را بخواهیم در اینجا به زبان ساده تری و در دنیای روباتها شبیه سازی کنیم عامل مورد نظر ما باید توپ و خودش را در یک حرکت انفجاری (با حداکثر سرعت ممکنه ) به نقطه ایی برساند ، در ورای عامل رقیب ، که در آن نقطه عامل رقیب نتواند توپ را بزند ولی خودش قادر به این کار باشد . یعنی آن نقطه در ناحیه قابل ضربه زدن عامل رقیب نباشد . مسئله اصلی که مطرح است این است که عامل ما هنگامیکه می خواهد توپ را از نقطه فعلی به ناحیه پشت عامل رقیب برساند به توپ در چه جهتی و با چه قدرتی باید ضربه بزند تا ، در حالیکه خودش با حداکثر

سرعت حرکت می کند و گرنه از رقیب باز می ماند ، توپ در تمام یا حداقل در نقطه پایـان مسیر هنـوز در ناحیـه قابل ضربه زدن[1] او باشد ولی در ناحیه قابل ضربه زدن عامل رقیب نباشد.(شکل ۵)

## ۴-۳-آموختن دریبل




**شکل۵ : موقعیت اشیاء بعد از عمل دریبل موفق**

محیـط آتومـاتون یـادگیـری در مسـئله دریبـل میـدان فوتبـال و خدمتـگزار فوتبـــال (SoccerServer) بـا حضور مربی ناظر و عامل رقیب که تلاش می کند مانع از عبور توپ شود می باشد. ایــن محیـط توسـط عـامل حس شده و نتیجه حرکت و یا به عبارتی ارزیابی عمـل انتخاب شده توسط آتوماتون یادگیری به وسیله عـامل به آتوماتون اعلان می شود. آتوماتون بـاید عـامل را در تعیین ضربه مناسب برای عبور از مدافع راهنمایی کند.

از آنجاییکه ما باید مقادیر مناسب دو پارامتر را بیابیم تا عملی به طور مناسب انجام شود محیط ما یک محیـط چنـد آتوماتونی همزمان می باشد. آتوماتونهای یادگیری ما از ماهیت عمل بی خبر هستند و تنـها از بیـن تعـدادی اعمـال ممکنــه یکــی را انتخـاب میکنــد . مقــدار پارامترهــای قــدرت و جــهت ضربــه بــه تــوپ در دســتور (kick pow dir) اعمــال واقعـی هسـتند کـه انتخـاب آتومـاتون آنـها را تعییـن مـی کنـد. مقـدار پـارامتر قـدرت می تواند عددی بین ۰ تا ۱۰۰ باشد و مقدار پارامتر جهت نیزعددی بین ۱۸۰- تا ۱۸۰ می تواند باشد . برای نگاشت فضای پیوسته تصمیم گیری به فضایی گسسته ، ما مقادیر ممکن برای هر یک از این پارامترها را بــه بـازه هـایی بـا طول یکسان تقسیم کرده و از آتوماتون انتظار داریم مناسب ترین بازه برای پارامتر مربوطه را ، به طوریکــه احتمـال موفقیت و در نتیجه دریافت پاداش بیشتر باشد ، بیابد . این پارامترها تعیین کننده زمان و مکـان رسـیدن تـوپ بـه نقطه ای در پشت عامل مدافع می باشند . با روشی ابتکاری می توانیم فضای انتخاب را محدودتر بنماییم. وقتــی بـه مسئله دقیقا نگاه شود معلوم می شود که عمل مطلوب آتوماتون یعنی مقدار پارامترهای قدرت و جهت ضربه به توپ در دستور (kick pow dir) در این مسئله باید مقادیری در بازه ای با حدود تعریف شده بخصوصی باشــند . بـرای پارامتر جهت این بازه یکی از بازه های تعریف شده در فاصله ۱۰ تا ۵۰ درجه را ما در نظر گرفته ایم ایــن انتخـابی ابتکاری[2] برای مسئله ما می باشد که با توجه به تعریف و انتظاراتی که ما از عمل دریبل داریم تعیین شــده اسـت . برای پارامتر قدرت ما بازه بین ۲۰ الی ۱۰۰ را انتخاب کرده ایـم زیـرا قـدرت کمـتر از ۲۰ تـوپ را چنـدان بـه جلـو نمی برد که بتوان دریبل موفقی را انجام داد . در زمان بازی این بازه مقادیر برای پارامترها عامل ما ثابت اسـت ولـی بسته به شرایط در یکی از قسمتهای تعیین شده داخل این بازه ، و حداکثر در قسمتهای همسایه قسمت یافته شده در زمان یادگیری ، تغییر می کند . بنابراین در صورتیکه ما آتوماتونها را بـه طـور مناسـبی تنظیـم و بـه روز رسـانی نماییم آنها باید به یک بردار همگرا شوند که در آن شانس انتخاب اعمال معدودی بسیار بیشتر است بـه طوریکـه در نهایت عمل مطلوب در بین یکی از چند بازه همسایه واقع می شود و تمام بـازه هـا نمـی تواننـد بـه عنـوان یکـی از انتخابهای ما در زمان واقعی بازی مورد استفاده قرار گیرند.

به نظر ما آتوماتونهای مورد استفاده برای این مسئله باید از نوع خطی باشند تا با همگـرا شدن به یک بازه مشخص احتمال انتخاب اعمال دیگر را در بازی واقعی بسیار کم کند . ضمن اینکه امکان تغییر به بازه های همسایه ، با تغییر شرایط ، محتمل باشد . ما از آتوماتون یادگیری از خانواده $L_{rp}$ برای عامل خودمان استفاده کرده ایم .

---

[1] *Kickable Area*

به منظور آموختن مهارت دریبل ، ما خدمتگزار را با مد فقط مربی اجرا می نماییم و عـامل مربـی را نـیز فعـال مـی کنیم . دو عامل را نیز برای بازی در این سناریو به کار می گیریم : اولی عامل دریـافت کننـده تـوپ و دومـی عـامل مورد نظر ما برای یادگیری مهارت دریبل است . در هر تلاش از جلسه دریبل عامل دریبل زننده که درست روبـروی مدافع قرار داده شده است سعی در عبور توپ از عامل مدافع و رساندن توپ به ناحیه پشت آنرا می کند (شکل ۵) و عامل مدافع تلاش می کند توپ را دفع نماید . علاوه بر این ما توپ را در ناحیه قابل ضربه زدن جلوی عامل مهاجم قرار می دهیم . عامل مربی در اینجا فقط نقش منتقد را بازی می کند و موفق یـا نـاموفق بـودن عمـل را تعییـن میکند هر چند که در این سناریو تعیین موفقیت عمل بسیار پیچیده تر از مهارتی مثل دریافت توپ اسـت و تعییـن علت عدم موفقیت عمل (نگاشت پس خورد) - به این معنی که علت عدم موفقیت انتخاب زاویـه نامناسـب ، قـدرت نامناسب، و یا هر دو بوده است - مسئله ای پیچیده و مهم است.

## ۴-۴-نتایج

در این مسئله به دلیل استفاده از دو آتوماتون برای حل مسئله فضای انتخاب ما بسیار وسیع تر از مسـئله دریافت توپ می باشد . از نظر ریاضی فضای انتخاب برای پارامترهای یک آتوماتون $L_{rp}$ برابر $R^2$ مـی باشـد و بـا در نظر گرفتن استفاده از دو آتوماتون یادگیری فضای انتخاب ما برابر $R^4$ خواهد بـود . در عمـل ترکیبـات متفـاوتی از آزمایشات را برای یافتن بهترین نتیجه باید بررسی کرد : استفاده از تعداد اعمال متفاوت برای هر یـک از آتومانـها و ارزیابی عملکرد ترکیب آنها با یکدیگر و سپس ارزیابی آنها در محیطهای دیگری با مقادیر متفـاوت اغتشـاش فضـای تنظیمات ممکنه را بسیار وسیع می سازد . روش ما برای انتخاب پارامترها بر اساس تجربه و ابتکاری بوده است .

ما ابتدا عامل دریبل زننده را در حالیکه عامل مدافع ثابت در روبروی او است و فقط اقـدام بـه دفـع تـوپ می کند به کار گرفتیم و با چند ترکیب ابتکاری از آتوماتونها ارزیـابی کردیـم(جـدول ۵) و سـپس از بـهترین نتـایج حاصله از آزمایشهای قبلی برای عبور از عامل مدافع متحرک و عکس العملی استفاده نمودیم که نتایج آن در جدول ۳ ارائه شده است .

**جدول ۵ : کارآیی دریبل با ترکیب آتوماتونهای یادگیرنده متفاوت (۱۰۰تکرار) ، مدافع ثابت**

| آتوماتون قدرت | | | | آتوماتون جهت | | | | کارآیی کل |
|---|---|---|---|---|---|---|---|---|
| نوع | تعداد عمل | (پ ،ج)[1] | کارآیی | نوع | تعداد عمل | (پ ،ج) | کارآیی | موفقیت دریبل |
| Lrp | ۲۰ | ۰/۱،۰/۳ | ٪۸۰ | Lrp | ۸ | ۰/۱،۰/۳ | ٪۸۳ | ٪۷۴ |
| Lrp | ۴۰ | ۰/۱،۰/۳ | ٪۹۲ | Lrp | ۲۰ | ۰/۱،۰/۳ | ٪۷۸ | ٪۷۷ |
| Lrp | ۱۰ | ۰/۱،۰/۳ | ٪۷۸ | Lrp | ۴ | ۰/۱،۰/۳ | ٪۶۱ | ٪۵۱ |
| Lrp | ۲۰ | ۰/۲،۰/۲ | ٪۹۰ | Lrp | ۸ | ۰/۲،۰/۲ | ٪۶۳ | ٪۵۶ |
| Lrp | ۲۰ | ۰/۱،۰/۱ | ٪۷۰ | Lrp | ۸ | ۰/۱،۰/۱ | ٪۴۱ | ٪۳۴ |
| Lrp* | ۸۰ | ۰/۱،۰/۳ | ٪۹۷ | Lrp | ۲۰ | ۰/۱،۰/۳ | ٪۸۵ | ٪۸۴ |
| Lrp | ۸۰ | ۰/۱،۰/۳ | ٪۹۴ | Lrp | ۴۰ | ۰/۱،۰/۳ | ٪۶۷ | ٪۶۴ |
| Lrp | ۴۰ | ۰/۱،۰/۳ | ٪۷۵ | Lrp | ۸ | ۰/۱،۰/۳ | ٪۷۵ | ٪۶۰ |
| Lri* | ۸۰ | ۰،۰/۳ | ٪۹۵ | Lri | ۴۰ | ۰،۰/۳ | ٪۹۲ | ٪۹۱ |
| Lri | ۸۰ | ۰،۰/۲ | ٪۹۳ | Lri | ۴۰ | ۰،۰/۲ | ٪۷۴ | ٪۷۲ |
| Lri | ۸۰ | ۰،۰/۱ | ٪۸۰ | Lri | ۴۰ | ۰،۰/۱ | ٪۳۴ | ٪۳۲ |

[1] *پاداش ،جریمه*

در جدول ۵ بهترین کارآیی مربوط به ترکیبی ار آتوماتونهای $L_{ri}$ ، بـا ۹۱٪، کـارآیی مـی باشـد ولـی یـک مسئله در رابطه با یادگیری این آتوماتون در این محیط ما را نگران کرد و آن این که آتوماتونهای برتر یافته شده از نوع Lri ، به خاطر خصوصیت عدم دریافت جریمه ، قابلیت تغییر رفتار بد خود در برابر تغییر رفتـار محیـط ، و بـه خصوص عامل رقیب ، را ندارند زیرا بردار احتمال عمل آنها برای برخـی از اعمـال ، پـس از آمـوزش، مقـدار صفـر را تعیین کرده است. به همین دلیل ما بهترین ترکیب آتوماتونهای $L_{rp}$ را نیز انتخاب و این دو را در برابر مدافع فعال و متحرک آزمایش نمودیم . همچنانکه در جدول ۶ مشاهده می شود ، و انتظار هم می رفت ، ترکیب آتوماتونهای $L_{rp}$ کارآیی بهتری را در برابر مدافع فعال ، محیط پویا ، نشان دادند.

**جدول ۶: کارآیی دریبل با ترکیب آتوماتونهای آموزش دیده و برتر جدول ۵ ، مدافع فعال**

| آتوماتون قدرت | | | | آتوماتون جهت | | | | کارآیی کل |
|---|---|---|---|---|---|---|---|---|
| نوع | تعداد عمل | (پ ،ج) | کارآیی | نوع | تعداد عمل | (پ ،ج) | کارآیی | موفقیت دریبل |
| Lri | ۸۰ | ۰،۰/۳ | ٪۸۲ | Lri | ۴۰ | ۰،۰/۳ | ٪۵۱ | ٪۴۹ |
| Lrp* | ۸۰ | ۰/۱،۰/۳ | ٪۹۸ | Lrp | ۲۰ | ۰/۱،۰/۳ | ٪۷۱ | ٪۷۰ |

ما عامل دریبل زن آموزش دیده با ترکیب آتوماتونهای برتر جدول ۶ را در برابر مدافع فعال در تکرارهـای بیشتری سنجیدیم که نتایج جدول ۷ نشان دهنده ثبات کارآیی می باشند .

**جدول ۷ : کارآیی عامل دریبل زن با ترکیب آتوماتون آموزش دیده برتر جدول ۶ ، مدافع فعال**

| تکرار / کارآیی | ۱۰۰ | ۲۰۰ | ۳۰۰ | ۴۰۰ |
|---|---|---|---|---|
| آتوماتون قدرت | ٪۹۸ | ٪۹۶ | ٪۸۶ | ٪۵۸ |
| آتوماتون جهت | ٪۷۱ | ٪۶۷ | ٪۸۵ | ٪۶۴ |
| دریبل | ٪۷۰ | ٪۶۸ | ٪۷۵ | ٪۶۰ |

## ۴-۵-جمع بندی و کارهای آینده

در این قسمت درباره مهارت دریبل و اهمیت آن ، به عنوان مهارتی کـه در موفقیـت سیاسـتهای پیچیـده نقش مهمی دارد ، توضیح داده شد و به بررسی مهارت دریبل در فوتبال روباتـها ، لیـگ شـبیه سـازی ، پرداختیـم . ابتدا مسئله دریبل را به صورتی مشخص که قابل تحلیل باشـد تعریـف نمودیـم . و سـپس مـدل سیسـتم آتومـاتون یادگیری متناسب با تعریف ارائه شده را طراحی نمودیم . توضیح داده شد که بر طبق مـدل تعریـف شـده از مسـئله دریبل ما باید دو پارامتر قدرت و جهت را در دستور (kick power direction) به هنگام دریبل تعییـن و بهینـه سازی نماییم . هدف ما نشان دادن قابلیت به کارگیری آتوماتونها یادگیری در حل و پیاده سازی مسائل یادگیـری و پیچیده سیستمهای چند عامله بود و نتایج حاصله ضمن اثبات این امکان از نظر کارآیی نیز دلگرم کننده بودند .

کارآیی دریبل را می توان از این هم بیشتر کرد اگر که برای هر زاویه دریبـل یـک آتومـاتون قـدرت جـدا استفاده شود و برای آن زاویه بهینه گردد . این کار اگرچه سربار پیاده سازی را بیشتر می کند ولی احتمالا کارآیی را به طور محسوسی افزایش خواهد داد . استفاده از روش ابتکاری این مقاله در پیاده سازی مهارتهای پیچیـده دیـگر ، نظیر انتخاب پاس ، از کارهای آینده می تواند باشد .

## ۵-یافتن سیاست دفاعی بهتر

یکی از مهمترین مسائل در سیستمهای چند عامله مسئله هماهنگی و بهینه سازی آن می باشد [7]. عوامل یک تیم برای هماهنگی باید به گونه ای در اجرای یک طرح مشترک با یکدیگر توافق و اقدام نمایند. راه حل این امر در شکل طبیعی از طریق تبادل دانش فردی عوامل (درک وضعیت فعلی، طرح مورد نظر) باید انجام شود. اما تعامل امری پر هزینه و نامطمئن در محیطهای پیچیده سیستمهای چند عامله است. یکی از مهمترین کاربردهای یادگیری در سیستمهای چند عامله کمک به بهینه سازی هماهنگی و کاهش هزینه های آن می باشد.

در معماریهای لایه ای عوامل بازیگر فوتبال [6] رفتار عامل با توجه به تصمیم گیری در لایه های مختلف آن تعیین می شود. به این معنی که در مقام اتخاذ هر وظیفه پیچیده وظیفه را به شکل پایین به بالا تجزیه نموده و اجرا می نماید. همانطور که در انجام مهارتهای فردی نیازمند به آموزش و داشتن روشهای هوشمند هستیم در مهارتهای گروهی نیز داشتن روشهایی برای بهبود و سازگاری رفتار و مهارتهای گروهی عامل در بهبود کارایی تیم بسیار موثر است.

یک چالش مهم در ارزیابی رفتارهای گروهی مسئله ارزیابی رفتار بلند مدت است [6]. به طور مثال در ارزیابی یک سیاست دفاعی به کار گرفته شده قضاوت به محض انجام عمل میسر نیست و در مجموع یک سلسله از اعمال نهایتا با حصول به هدفی خاص یا عدم حصول آن هدف مجموع کل اعمال انجام شده موفق یا ناموفق ارزیابی می شود. علاوه بر مسئله ارزیابی عملکرد در بلند مدت چالش بزرگتر پیش رو مسئله نگاشت مسئولیت پاداش/جریمه به اعمال جزء سیاست اتخاذ شده است.

در مسئله یافتن سیاست بهتر، که نمونه ای از رفتارهای تیمی است، مسائل اساسی سیستم های چند عامله بیشتر مطرح می شود. مسائل همکاری، ارتباط، و تعامل بین عوامل از مسائلی هستند که طراح سیستم باید برای دست یابی به رفتار بهینه آنها را حل کند. هدف و اولویت اصلی در اینجا بهبود رفتار تیمی عوامل یک سیستم چند عامله، فوتبال روباتها در اینجا، می باشد.

در این قسمت ما به ارائه امکان کاربرد آتوماتونهای یادگیری قطعی به منظور یادگیری رفتاری تیمی، اتخاذ سیاست دفاعی، می پردازیم. کاهش هزینه های ارتباط و افزایش اطمینان در هماهنگی از خصوصیات برجسته این روش می باشند .

### ۵-۱-کارهای قبلی

اکثر تیمها در ابتدا به طراحی و پیاده سازی مهارتهای فردی عوامل فوتبالیست پرداخته اند و از این امر گریزی نیست زیرا بدون داشتن مهارتهای فردی هیچ طرح تیمی موفق نخواهد بود. در جام لیگ شبیه سازی سال ۱۹۹۶ بازیکنان تیمهای شرکت کننده تنها از مهارتهای فردی برخوردار بودند. در دوره های بعد تیمها به سمت طراحی و پیاده سازی مهارتهای گروهی و تیمی رفتند. طرحهای گروهی و تیمی در ابتدا بیشتر غیر هوشمند و از پیش برنامه ریزی شده بودند به همین دلیل نیازمند مقدار قابل توجهی تبادل پیام و تعامل بین عوامل بودند. در نمونه ای از طرحهای دفاعی که در مرجع [9] به آن اشاره شده سیاست دفاعی این تیم بر اساس هدایت مرکزی بازیکنان دفاع از نظر جایگیری و هماهنگی می باشد و این هدایت از طریق ارسال پیامهای شنیداری صورت می گیرد. در زمینه سیاست و طرح دفاعی تا سال ۲۰۰۰ کار خاصی انجام نشده است [10] بسیاری از تیمها روی طرحهای حمله بسیار پیشرفت کرده اند.

### ۵-۲-مسئله تعیین سیاست دفاعی

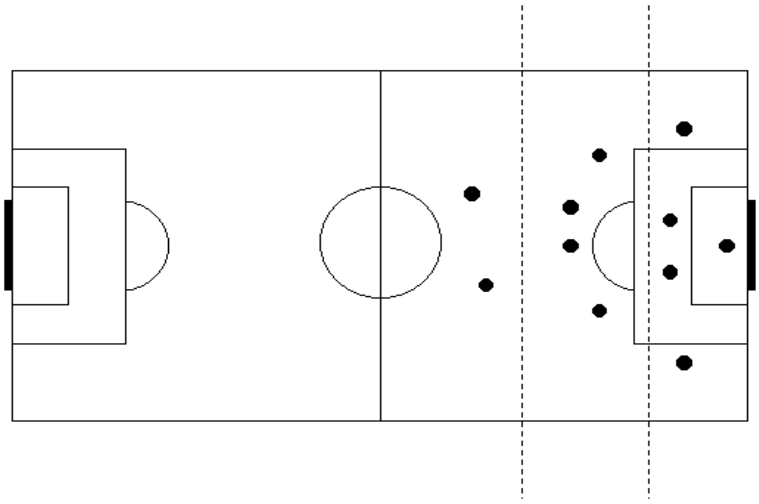
در هنگامیکه تیم می خواهد در برابر حمله تیم رقیب به دفاع بپردازد می توان از استراتژیهای دفاعی متفاوتی استفاده کرد. این استراتژیهای دفاعی ضمن داشتن هدف مشترک از نظر چگونگی موقعیت جاگیری نفرات تیم، یارگیری، و دفع توپ با یکدیگر می توانند متفاوت هستند. مناسب بودن یک شیوه دفاعی امری نسبی است به

این دلیل که ممکن است در برابر حمله یک تیم با استراتژی حمله به خصوصی موفق باشد و در برابر حمله تیم دیگری با استراتژی حمله متفاوت ناموفق باشد. مثلا تیمهایی هستند که توپ را با پاسگاری به موقعیت مناسبی برای شوت به سمت دروازه رسانده و اقدام به شوت می کنند ( مثل تیم CMUnited قهرمان سالهای ۱۹۹۸ و ۱۹۹۹ [7] ) و در مقابل تیمهایی هستند که توپ را به طرفین زمین برده و به محوطه جلوی دروازه سانتر می کنند تا هم تیمی های حاضر در آنجا اقدام به گل زنی نمایند (تیم FC Portugal قهرمان سال ۲۰۰۰ [10] ) . یکی از مسائلی که تا سال ۲۰۰۰ روی آن تیمها کار زیادی نکرده بودند استراتژی دفاعی بود. تیمها روی دروازه بان، روشهای حمله، و مهارتهای فردی پیشرفتهای زیادی داشتند ولی در زمینه دفاع کار متفاوتی دیده نمی شد [10].

دفاع تیمی یک مهارت گروهی است. به این معنی که در انجام آن گروهی از عوامل باید با یکدیگر همکاری کنند و بنابر این یکی از مسائلی که در اینجا مطرح می شود همکاری و هماهنگی مناسب عوامل است. اگر تیمی چند روش دفاعی را برای دفاع توانسته باشد طراحی و پیاده سازی نماید که هر یک برای دفاع در برابر تیمی با استراتژی حمله خاصی مناسب باشد مسئله مهم تصمیم گیری در لایه سیاست این است که تعیین شود در بازی در حال اجرا کدام سیاست دفاعی مناسب تر است و کارآیی بهتری دارد.

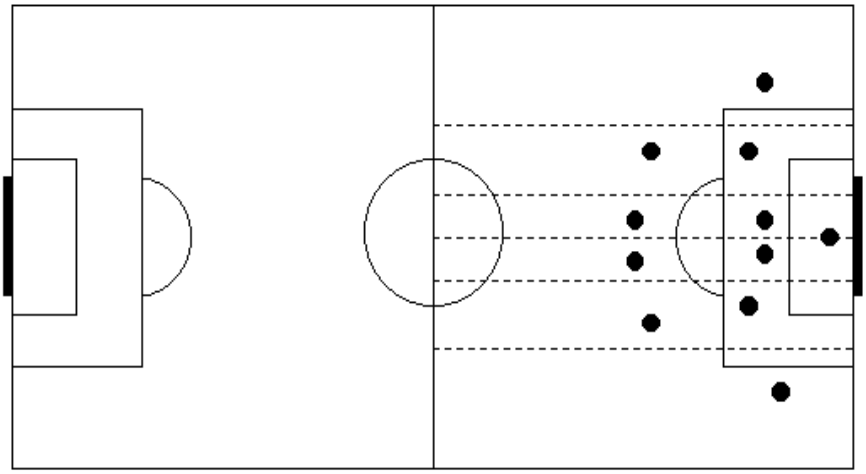
هرچند در این مقاله هدف ما طراحی یک روش دفاعی خاص نیست - زیرا این مسئله ای است که باید در لایه دیگر با استفاده از مهارتهای پیچیده جاگیری، آفساید گیری، یارگیری، ربودن یا دریافت توپ، و ... حل شود - ما دو روش دفاعی متفاوت را برای بکارگیری روشمان طراحی نمودیم :

- در یک روش بازیکنان تیم در هنگام دفاع در سه لایه اقدام به دفاع می کنند (شکل ۶) این لایه ها به موازات هم و به موازات خط عرضی زمین مسابقه قرار گرفته اند. یک قانون کلی برای هر عامل بازیکن این است که در ناحیه مشخصی فعالیت دفاعی انجام می دهد. بعلاوه هر بازیکن در منطقه دفاعی خود از قواعد مشخصی پیروی می کند.



**شکل ۶ : استراتژی دفاعی اول**

- در روش دوم بازیکنان تیم در هنگام دفاع در لایه های موازی با خط طولی میدان با آرایش متفاوتی اقدام به دفاع می کنند (شکل ۷). یک قانون کلی برای هر عامل بازیکن این است که بسته به نقش خودش در ناحیه مشخص شده آن نقش و بین خطوط موازی فعالیت دفاعی انجام می دهد. علاوه بر این هر بازیکن در منطقه دفاعی خود از قواعد مشخصی پیروی می کند.



**شکل۷ : استراتژی دفاعی دوم**

با داشتن چند روش دفاعی حال مسئله ما این است که کدام روش برای دفاع در بازی فعلی و در برابر تیم فعلی مناسب تر است. روش بهتر بستگی به نوع بازی تیم حریف دارد. اگر تیم ما با یک روش، به طور تصادفی و یا با نظر مربی، شروع کند باید در صورت مناسب بودن روش فعلی به آن ادامه دهد و روش دیگری را آزمایش نکند. در صورت مناسب نبودن روش دفاعی فعلی تیم باید به روش دفاعی دیگری تغییر وضعیت دهد. این کار ضمن اینکه حداقل رفتار تیمی ما را تغییر می دهد و احتمالا تیم رقیب را با مقداری مشکل روبرو می کند منجر به استفاده از بهترین روش در دسترس نیز می شود.

## ۵-۳-استفاده از آتوماتون یادگیری برای تعیین سیاست دفاعی

به منظور تعیین سیاست دفاعی، عوامل یا بازیگران تیم در زمان دفاع به طور جمعی بر سر یک طرح توافق و سپس به اجرای آن می پردازند بدون اینکه پیامی مبادله شود ویا هماهنگی متمرکزی در کار باشد. هر عامل به طور فردی سیاست دفاعی فعلی را تشخیص می دهد اما این تشخیص با تشخیص عوامل دیگر یکسان است. ایده ما اینست که برای حصول به این منظور هر عامل از یک آتوماتون قطعی از خانواده آتوماتونهای ثابت و دارای حافظه استفاده نماید. آتوماتونهای نامزد و قابل استفاده برای این منظور آتوماتونهای Krinsky , $G_{2N,2}$ , $L_{2N,2}$ هستند. چون ما فقط دو نوع سیاست دفاعی را طراحی کرده ایم از دو عمل برای آتوماتونهایمان استفاده کرده ایم اما در واقع به تعداد سیاستهای پیاده سازی شده در دسترس می توان برای این آتوماتونها عمل در نظر گرفت. ما نیز در پیاده سازی آتوماتونها را به گونه ای پیاده سازی نموده ایم که امکان توسعه اعمال آنها بدون تغییر در کد پیاده سازی آنها میسر می باشد. در اینجا رفتار سیستم را در صورت استفاده از هر یک از این آتوماتونها بررسی نموده ایم.

در آتوماتون $L_{2N,2}$ ما از بین دو عمل یکی را با توجه به وضعیت فعلی آتوماتون انتخاب می کنیم. برای هر عمل $N$ وضعیت داریم در ابتدا در اولین وضعیت هستیم و با هر عمل که پاسخ مثبت محیط را داشته باشد، دفاع موفق، به وضعیت بعدی می رویم - اگر در وضعیت آخر، $N$ ام، باشیم در همانجا می مانیم یعنی به همان سیاست دفاعی ادامه می دهیم. چنانچه از محیط پاسخ نا مطلوب دریافت شد به وضعیت قبلی می رویم ولی اگر در هنگام انجام عمل قبلی در وضعیت اول بوده ایم به اولین وضعیت عمل دیگر، سیاست دفاعی دیگر، می رویم و این منجر به تغییر سیاست دفاعی می شود. انتخاب عمل بعدی ما، در اینجا سیاست دفاعی، همان عملی است که در وضعیت آن هستیم. تفسیر این رفتار در بازی این گونه است که تیم ما تا زمانیکه یک سیاست دفاعی موفق عمل کند به آن ادامه می دهد ولی حداکثر پس از $N$ بار عدم موفقیت سیاست دفاعی خود را عوض می کند. ممکن است تغییر سیاست دفاعی حتی در یک بار عدم موفقیت رخ دهد و این در زمانی است که ما در اولین استفاده از یک سیاست دفاعی گل بخوریم. این مطلب نشان می دهد که در این روش، ما در شروع استفاده از هر سیاست دفاعی کمترین

اعتماد را به آن داریم و با هر بار عمل موفق دلگرم تر و با هر عمل ناموفق دلسرد می شویم.

در آتوماتون $G_{2N,2}$ ما از بین دو عمل یکی را با توجه به وضعیت فعلی آتوماتون انتخاب می کنیم. در اینجا هم برای هر عمل $N$ وضعیت داریم اما بر خلاف آتوماتون $L_{2N,2}$ در ابتدا در آخرین وضعیت هستیم. با هر عمل که پاسخ مثبت محیط را داشته باشد ، دفاع موفق، به وضعیت بعدی می رویم - اگر در وضعیت آخر، $N$ ام، باشیم در همانجا می مانیم و به همان سیاست دفاعی ادامه می دهیم. چنانچه از محیط پاسخ نامطلوب دریافت شد به وضعیت قبلی می رویم ولی اگر در هنگام انجام عمل قبلی در وضعیت اول بوده ایم به آخرین وضعیت عمل دیگر، سیاست دفاعی دیگر، می رویم و این منجر به تغییر سیاست دفاعی می شود. انتخاب عمل بعدی ما همان عملی است که در وضعیت آن عمل هستیم. تفسیر این رفتار در بازی این گونه است که تیم ما تا زمانیکه یک سیاست دفاعی موفق عمل کند به آن ادامه می دهد و حداقل پس ا ز $N$ بار عدم موفقیت سیاست دفاعی خود را عوض می کند. این مطلب نشان می دهد که ما در شروع استفاده از هر یک از سیاستهای دفاعی بیشترین اعتماد را به آن داریم و با هر بار عمل موفق دلگرم تر و با هر عمل ناموفق دلسرد می شویم. و برای تغییر سیاست دفاعی حداقل $N$ بار باید گل بخوریم.

در آتوماتون Krinsky ما از بین دو عمل یکی را با توجه به وضعیت فعلی آتوماتون انتخاب می کنیم. در اینجا ه هم برای هر عمل $N$ وضعیت داریم. بر خلاف آتوماتون $L_{2N,2}$ و مثل آتوماتون $G_{2N,2}$ در ابتدا در آخرین وضعیت هستیم. با هر عمل که پاسخ مثبت محیط را داشته باشد، دفاع موفق، به وضعیت آخر، $N$ ام، می رویم و به همان سیاست دفاعی ادامه می دهیم. چنانچه از محیط پاسخ نامطلوب دریافت شد به وضعیت قبلی می رویم ولی اگر در هنگام انجام عمل قبلی در وضعیت اول بوده ایم به آخرین وضعیت عمل دیگر، سیاست دفاعی دیگر، می رویم و این منجر به تغییر سیاست دفاعی می شود. انتخاب عمل بعدی ما همان عملی است که در وضعیت آن هستیم. تفسیر این رفتار در بازی این گونه است که تیم ما تا زمانیکه یک سیاست دفاعی موفق عمل کند به آن ادامه می دهد و همیشه پس از $n$ بار عدم موفقیت سیاست دفاعی خود را عوض می کند. این مطلب نشان می دهد که ما در شروع استفاده از هر سیاست دفاعی بیشترین اعتماد را به آن داریم و با هر بار عمل موفق اعتماد اولیه را مجددا به آن سیاست پیدا می کنیم و با هر عمل ناموفق دلسرد می شویم. و برای تغییر سیاست دفاعی حتما باید $N$ بار پشت سر هم در دفاع شکست بخوریم.

## ۵-۴-بحث و مقایسه

در این قسمت به شرح ایده به کار گیری آتوماتون یادگیری برای یافتن و عمل به بهترین سیاست دفاعی در برابر حریف پرداختیم. این کاربردی از آتوماتون یادگیری در تصمیم گیری در لایه سیاست را نشان می دهد. ما دو سیاست دفاعی طراحی نموده ایم و به طور برخط و در هنگام بازی سیاست بهتر را انتخاب می کنیم. به دلایل ذکر شده در قسمت ۶ (نیاز به تصمیم یکسان بدون تبادل پیام) از آتوماتونهای خطی و غیر ثابت نمی شد استفاده کرد. آتوماتون یادگیری مورد استفاده ما می توانست یکی از آتوماتونهای Krinsky , $G_{2N,2}$ , $L_{2N,2}$ باشد. رفتار دفاعی تیم در استفاده از هر یک از این آتوماتونها بررسی شد. با توجه به رفتار توضیح داده شده به نظر می رسد مناسب ترین آتوماتون، آتوماتون $G_{2N,2}$ برای یک تیم می باشد البته با عمق حافظه کم، مثلا $N \leq 4$، تا علاوه بر فرصت دادن به یک سیاست برای اثبات کارآیی و چشم پوشی متناسب از عملکرد احتمالا بد دیگران، پیش از دیر شدن و خیلی گل خوردن، به تغییر سیاست دفاعی در صورت نیاز بپردازد. آتوماتون $L_{2N,2}$ بدبین است و ممکن است سیاست دفاعی مناسبی را خیلی زود تغییر دهد و دائما به تغییر سیاست بپردازد. آتوماتون Krinsky هم خیلی خوش بین است و ممکن است باعث عدم تغییر سیاست دفاعی نامناسب و یا تغییر خیلی دیر آن شود.

در مقایسه با روشهای دیگر روش ما دارای خصوصیات برجسته ای است که در جدول زیر به آن اشاره شده است:

**جدول ۸ : مقایسه روشهای تعیین سیاست دفاعی**

| روش / خصوصیت | آتوماتون یادگیری | روش هدایت متمرکز |
|---|---|---|
| تصمیم گیری | توزیع شده | متمرکز |
| تبادل پیام | ندارد | زیاد |
| اطمینان | زیاد | نسبی/کم |
| انعطاف پذیری | معین | دارد |

## ۶-جمع بندی

به کارگیری آتوماتون یادگیری در جام روباتها سابقه قبلی ندارد ، هر چند از یادگیری تقویتی استفاده شده است . در برخی از کارهای انجام شده توسط گروه های علمی دانشگاهها یا موسسات تحقیقاتی ، هدف تحقیقات تنها امکان به کار گیری روشی خاص برای حل مسائل مطرح در جام روباتها بوده است و با به کارگیری توان فکری و محاسباتی زیادی نتایج مطلوبی نیز ، بر طبق اظهار خودشان ، بدست آورده اند . چنانچه فقط از این منظر به پژوهش انجام شده در این مقاله بنگریم می توانیم اظهار کنیم که نتایج بدست آمده حاکی از موفقیت و امکان به کار گیری آتوماتونهای یادگیری در حل مسائل جام روباتها ، و سیستمهای چند عامله ، داشته اند . اما نتایج حاصله چیزی بیش از این را نشان می دهد :

۱. در مسئله دریافت توپ روش به کار گرفته شده ما به کارآیی مطلوب و قابل رقابت با روشهای ارائه شده ، و بسیار مطرح ، دست یافته است و توضیح داده شد که در مقایسه با روش شبکه های عصبی که توسط آقای پیتر استون برای حل این مسئله استفاده شده بود دارای مزیتهایی از قبیل انعطاف پذیری در زمان واقعی ، تحلیلی بودن ، و داشتن پایه تئوریک می باشد .

۲. در زمینه استفاده از آتوماتون یادگیری در مهارتهای سطح بالا نظیر تعیین سیاستهای کلان بازی ، ما از آتوماتون یادگیری برای یافتن مناسب ترین سیاست دفاعی موجود در برابر حریف استفاده کردیم . خصوصیت روش ما امکان پیاده سازی مهارتی تیمی در ضمن داشتن کمترین ارتباط مرکزی یا سلسله مراتبی برای حصول هماهنگی و همکاری تیمی می باشد .

۳. در زمینه پیاده سازی مهارتهای پیچیده ما به چالش با مسئله دریبل ، که یک مسئله چند عامله است به این معنی که چند عامل در انجام کاری با یکدیگر تعامل دارند ، با استفاده از آتوماتونهای یادگیری پرداختیم . نتایج بدست آمده ضمن تایید امکان به کارگیری آتوماتون یادگیری در این مسئله ، کارآیی کاملا قابل قبولی را از نظر موفقیت در یادگیری ( ۷۰٪ موفقیت در عمل دریبل ) و زمان لازم برای یادگیری این مهارت ، به معنی واقعی یادگیری به گونه ای که یادگیری تقویتی وعده داده است ، بدست دادند .

خصوصیات مهم روش آتوماتون یادگیری ذکر شود:

- از آنجاییکه تئوری آتوماتون یادگیری دارای پایه های قوی تحلیل ریاضی می باشد هر سیستمی که با استفاده از این روش مدل سازی شود از این پایه تحلیلی سود خواهد برد و در واقع درباره عملکرد آن می توان مطمئن تر اظهار نظر کرد .
- این روش قابلیت سازگاری با محیط جدید را دارد و بنابر این انعطاف پذیری که از خصوصیات مهم عوامل در محیط سیستم های چند عامله است را برآورده می کند.
- از منظر هوش مصنوعی آتوماتون یادگیری به نحو مطلوبی توازن را در مسئله "بهره برداری در مقابل

اکتشاف " برقرار می کند و استفاده از آن به طور ذاتی راه حلی برای این مسئله را در بر دارد . هر چند میزان کارآیی این راه حل ذاتی بستگی به نوع آتوماتون مورد استفاده دارد.

- طراحی مدل نمادین سیستم آتوماتون یادگیری مهمترین بخش استفاده از این روش یادگیری می باشد که به عهده طراح سیستم می باشد . علاوه بر طراحی مدل نمادین سیستم ، طراحی صحیح ارزیابی پاسخ محیط نقشی تعیین کننده در موفقیت روش دارد.

از نظر مهندسی نرم افزار قابلیت استفاده مجدد[1] آتوماتون یادگیری حسن مهمی از این روش می باشد. همانطور که ملاحظه شد عملکرد آتوماتون یادگیری مستقل از ماهیت عملی بود که برای آن به کار گرفته شد ( مثلا بازه مناسب راویه چرخش ) . بنابر این اگر طراحی مناسبی انجام شود می توان از آن برای سازگاری اعمال دیگر هم استفاده کرد بدون اینکه نیاز باشد در پیاده سازی آن تغییری دهیم. ما این را در پیاده سازی راه حلهایمان تجربه نمودیم . از نظر پیاده سازی ساختمان و کد کلاس آتوماتون یادگیری به کار گرفته شده ، در تمامی مسائل حل شده ، یکی است .

چالش مهم پیش روی آتوماتون یادگیری ( مخصوصا آتوماتونهای از نوع متغیر ) تعیین بهترین مقادیر پارامترهای آن است که روش ما در اینجا برای این منظور روشی ابتکاری و شهودی بود. حل این چالش مستلزم به صرف زمان و قدرت محاسباتی عظیمی می باشد. همانطور که در تهیه و تکمیل این مقاله مهمترین و زمانگیرترین موضوع ، بعد از طراحی سیستم و مدل نمادین مسئله ، در پیش روی ما انجام آزمایشات مورد نیاز برای بدست آوردن جواب مطلوب بود. شاید یکی از زمینه های مناسب تحقیق بررسی روشی برای تعیین بهترین مقادیر ممکنه در فضای برداری (پاداش،جریمه) برای یک مسئله ، یا حتی یک مسئله خاص ، باشد و یا حتی پاسخ به این سئوال که شرایط لازم و کافی برای اینکه کارآیی به خصوصی (منظور درصد کارآیی است) حاصل شود چیست.

## ۷-مراجع


[1] Flantge ,F. , C. Meyer , B. Schappel , Th. Uthmann , Enhancing the Adaptive Abilities , Department of computer science,Johannes Gutehberg-university Mainz D-55099Mainz , Germany , 2001.

[2] Stone ,
Melon University, Technical Report, Nov 1995.

[3] Kaelbling, L. P., Littman, M. L., and Moore, A.W. (1996). Reinforcement Learning: A Survey, *Journal of Artificial Intelligence Research* Volume 4, pp. 237-285.

[4] Narendra,K.S.,M.A.LThathachar,LearningAutomata,Prentice_Hall Inc. , Newyork , 1989.

[5] Noda ,I
Agent 96 Carins , Australia , pp.570-579, Aug,1996.

[6] Stone ,P
Melon University , USA , 1998.

[7] Weiss,G. , Multi Agent Systems , The MIT Press , USA , 1999.

[8] Gupta ,A
Robot Soccer,Department of Computer Science and Engineering,Indian Institute of Technology,Kanpur,India,2001.

[9] Dorer K., Robocup-99 urg
magmaFreiburg ,
http://www.ep.liu.se/ea/cis/1999/007/17/.

[10] Asada, M.,A.Birk,E. Pagllo,Progressing in Robocup Soccer Research in 2000.


[1] *Reuseability*