UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE TECNOLOGIA E GEOCIÊNCIAS
DEPARTAMENTO DE ENGENHARIA DE PRODUÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE PRODUÇÃO

ADRYENNE CRISTINNI DE OLIVEIRA ANDRADE

PROCESSOS DE APRENDIZAGEM EM MODELOS *AGENT-BASED*: OS ALGORITMOS *REINFORCEMENT LEARNING* APLICADOS A TEORIA DOS JOGOS

Recife
2019

ADRYENNE CRISTINNI DE OLIVEIRA ANDRADE

PROCESSOS DE APRENDIZAGEM EM MODELOS *AGENT-BASED*: OS ALGORITMOS *REINFORCEMENT LEARNING* APLICADOS A TEORIA DOS JOGOS

Dissertação apresentada ao Programa de Pós-graduação em Engenharia de Produção da Universidade Federal de Pernambuco como parte dos requisitos parciais para obtenção do título de mestre em Engenharia de Produção.

Área de concentração: Pesquisa Operacional

Orientador: Profº. Dr. Francisco de Sousa Ramos

Recife
2019

Catalogação na fonte
Bibliotecária Maria Luiza de Moura Ferreira, CRB-4 / 1469

A553p Andrade, Adryenne Cristinni de Oliveira.
Processos de aprendizagem em modelos *agent-based*: os algoritmos *Reinforcement Learning* aplicados a teoria dos jogos / Adryenne Cristinni de Oliveira Andrade. - 2019.
71 folhas, il., tab.

Orientador: Prof. Dr. Francisco de Sousa Ramos.

Dissertação (Mestrado) – Universidade Federal de Pernambuco. CTG. Programa de Pós-Graduação em Engenharia de Produção, 2019.
Inclui Referências e Apêndices.

1. Engenharia de Produção. 2. Modelagem baseada em agentes. 3. Algoritmos de aprendizagem. 4. Teoria dos jogos. 5. Comportamento estratégico. I. Ramos, Francisco Souza (Orientador). II. Título.

UFPE

658.5 CDD (22. ed.) BCTG/2019-232

ADRYENNE CRISTINNI DE OLIVEIRA ANDRADE

PROCESSOS DE APRENDIZAGEM EM MODELOS *AGENT-BASED*: os algoritmos *Reinforcement Learning* aplicados a teoria dos jogos

Dissertação apresentada ao Programa de Pós-graduação em Engenharia de Produção da Universidade Federal de Pernambuco como parte dos requisitos parciais para obtenção do título de mestre em Engenharia de Produção.

Área de concentração: Pesquisa Operacional

Aprovada em: 12/04/2019.

BANCA EXAMINADORA

Prof. Dr. Francisco de Sousa Ramos (Orientador)

Prof. Dra. Isis Didier Lins (Examinadora Interna)

Prof. Dr. Tiago Alessandro Espinola Ferreira (Examinador Externo)

AGRADECIMENTOS

Agradeço à Deus por me permitir tê-lo como base durante as minhas caminhadas, essa, sem dúvidas, foi uma das mais importantes e difíceis. Nesses dois anos, adquiri mais que o conhecimento disponibilizado pelo programa de pós-graduação, me tornei mais forte e conheci pessoas que vou levar para sempre comigo. Eu agradeço, especialmente, à minha família, que tem me apoiado a cada novo desafio, aos meus pais, Marley e Geraldo, ao meu padrasto João e aos meus irmãos Ramon, Gabriella, Antonella e Sofia. Meus sinceros agradecimentos ao meu orientador, Francisco Ramos, por ser exemplo de comprometimento e dedicação. Eu espero, um dia, sentir pelo meu trabalho o mesmo entusiasmo que o senhor sente pelo seu e assim poder fazê-lo tão bem quanto. Meus agradecimentos aos membros da banca examinadora, por terem se disponibilizado a avaliar e contribuir com o meu trabalho e aos professores do PPGEP pelo conhecimento compartilhado durante esses dois anos de mestrado. Ao meu amigo, Dione, por me apoiar e sempre me incentivar a ser melhor. Aos amigos de longa data que souberam ser pacientes durante os momentos de estresse e ausência, especialmente, aos meus amigos Victor e João, que são sempre portos seguros durante meus dias difíceis. Aos amigos do PPGEP e também do PIMES, Patrícia, Emanuelly, Amanda, Olivier, Darío, Daniel, Diogo, Junior, Risomário e Flavius, por todos os momentos compartilhados, pelas conversas, conselhos e incentivos. Agradeço em especial à Carol, por todo o suporte nessa reta final. Agradeço, também, à Tereza, que sempre nos recebe com um sorriso, na secretaria, e pronta para nos ajudar, muito obrigada. Levo cada um de vocês para a vida, levo também todos os ensinamentos diários que esses dois anos me proporcionaram. O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001, agradeço, também, à CAPES por ter tornado isso possível.

RESUMO

A partir da modelagem baseada em agentes, contextos em diferentes níveis de complexidade podem ser simulados. Esta técnica de análise que é principalmente desenvolvida levando em consideração a heterogeneidade dos indivíduos, quando utilizada em conjunto com algoritmos de aprendizagem por reforço, possibilita resultados precisos e mais próximos dos encontrados em contextos reais. Isso ocorre pois o comportamento estratégico é introduzido ao modelo de simulação por meio dos algoritmos de aprendizagem, possibilitando que o agente atue de forma a maximizar sua utilidade e satisfação. Ao aplicar estas abordagens ao estudo de problemas-padrão da teoria dos jogos, que apresentam equilíbrios pautados em racionalidade ilimitada, verificar-se-á a influência dos processos de aprendizagem tanto no comportamento individual do agente, quanto no resultado do jogo como um todo. Os algoritmos de aprendizagem por reforço, *Roth-Erev RL* (RE), *Modified Roth-Erev RL* (MRE) e *Variant Roth-Erev - RL* (VRE) foram incorporados ao comportamento de apenas um dos agentes que compõem a situação de conflito, com o objetivo de avaliar a capacidade de mapeamento de resposta, proporcionada por tais algoritmos, uma vez que o agente que não aprende apresenta dois diferentes comportamentos: fixo ou aleatório. Os parâmetros de experimentação e esquecimento, vieses psicológicos presentes nos algoritmos, sofreram variações buscando identificar possíveis influências nos processos de aprendizagem. Com isso, o objetivo do presente estudo é identificar possíveis alterações nos resultados canônicos conhecidos para os jogos do Dilema dos Prisioneiros, Batalha dos Sexos e *Chicken Game*, diante dos processos de aprendizagem incorporados ao modelo de simulação bem como da suposição de racionalidade limitada. Os três algoritmos foram capazes de proporcionar comportamento estratégico, ao agente que aprende, nos cenários em que os parâmetros de experimentação e esquecimento não foram considerados. Ao atribuir valores positivos a ambos os parâmetros, variações nos comportamentos puderam ser observadas. De um modo geral, o algoritmo *Roth-Erev RL* demonstrou maior robustez, quando incorporado a este tipo de estudo, ao confirmar os resultados canônicos determinados para cada um dos jogos clássicos testados, mesmo em resposta às variações de ambos os parâmetros. Já os algoritmos MRE e VRE demonstraram-se sensíveis às variações feitas no parâmetro de experimentação, resultando em comportamentos não correspondentes com o melhor cenário que poderia ser alcançado na situação de conflito, impossibilitando que o agente dotado de aprendizado realizasse o mapeamento das ações do agente oponente. Constatou-se que há uma escassez de trabalhos, na literatura, utilizando em conjunto, a Modelagem Baseada em Agentes, os algoritmos de aprendizagem e a teoria dos jogos, para estudar, sob diferentes perspectivas, o comportamento estratégico em ambiente de simulação, demonstrando dessa forma a contribuição deste estudo e uma área com alto potencial de exploração.

Palavras-chave: Modelagem baseada em agentes. Algoritmos de aprendizagem. Teoria dos jogos. Comportamento estratégico.

ABSTRACT

From agent-based modeling, contexts at different levels of complexity can be simulated. This technique of analysis that is mainly developed taking into account the heterogeneity of the individuals, when used in conjunction with reinforcement learning algorithms, allows accurate results and closer to those found in real contexts. This is because strategic behavior is introduced to the simulation model through the learning algorithms, enabling the agent to act in a way to maximize its utility and satisfaction. In applying these approaches to the study of problems of game theory, which present equilibria based on unlimited rationality, the influence of learning processes will be verified both on the individual behavior of the agent and on the outcome of the game as a whole. The reinforcement learning algorithms, Roth-Erev RL (RE), Modified Roth-Erev RL (MRE) and Variant Roth-Erev RL (VRE) were incorporated into the behavior of only one of the agents that compose the conflict situation, with the objective of evaluating the response mapping capability provided by such algorithms, since the non-learning agent presents two different behaviors: fixed or random. The experimentation and recency parameters, psychological bias present in the algorithms, suffered variations in order to identify possible influences in the learning processes. The objective of the present study is to identify possible changes in the canonical results known for the games of Prisoners' Dilemma, Battle of the Sexes and Chicken Game, in view of the learning processes incorporated into the simulation model as well as the assumption of limited rationality. The three algorithms were able to provide strategic behavior, to the learning agent, in the scenarios in which the parameters of experimentation and recency were not considered. By assigning positive values to both parameters, variations in behaviors could be observed. In general, the Roth-Erev RL algorithm demonstrated greater robustness, when incorporated to this type of study, in confirming the canonical results determined for each of the classic games tested, even in response to the variations of both parameters. However, the MRE and VRE algorithms proved to be sensitive to the variations made in the experimentation parameter, resulting in non-corresponding behaviors with the best scenario that could be reached in the conflict situation, making it impossible for the agent with learning to map agent actions opponent. It was found that there is a shortage of works, in the literature, using together, Agent-Based Modeling, learning algorithms and game theory, to study, from different perspectives, the strategic behavior in a simulation environment, demonstrating this the contribution of this study and an area with high exploration potential.

Keywords: Agent-based modeling. Learning algorithm. Game theory. Strategic behavior.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

A suposição de racionalidade ilimitada é comum aos estudos que consideram o comportamento dos indivíduos, principalmente, em áreas voltadas para a teoria da escolha do consumidor e a organização industrial, como relatam Urbina e Ruiz-Villaverde (2019), por exemplo. Essa suposição defende um comportamento padrão e sem falhas, como o estabelecido pelo conceito de *homo economicus* (STEINGRABER; FERNANDEZ, 2013). Novos estudos têm sido desenvolvidos com foco no comportamento, dando origem à racionalidade limitada, sendo, portanto, uma alternativa ao modelo de comportamento desenvolvido inicialmente (MELO; FUCIDJI, 2016).

A incorporação deste aspecto comportamental nos modelos de simulação recebeu um grande incentivo devido ao surgimento da denominada Modelagem Baseada em Agentes (ABM). Esses modelos têm como filosofia, a modelagem de sistemas complexos compostos por agentes autônomos e interativos. Esse tipo de modelagem mantém especial atenção sobre a heterogeneidade dos indivíduos, pois se considera que a interação dos diferentes comportamentos é responsável por originar o comportamento global, contradizendo assim o conceito de *homo economicus*. Essa modelagem é caracterizada por uma abordagem *bottom-up*, adotada a partir dos agentes individuais, permitindo capturar propriedades emergentes. A partir da identificação do comportamento individual, considerando as diferenças existentes no comportamento dos indivíduos, se torna possível explicar o comportamento global emergente dos comportamentos individuais (ABAR et al., 2017).

De acordo com Macal e North (2010), o comportamento dos agentes sofre influência tanto do ambiente, decorrente da interação com os demais agentes, quanto das regras de comportamento introduzidas no modelo. Ringler, Keles e Fichtner (2016), evidenciam uma coleção de características que podem ser incorporadas ao modelo de simulação como, por exemplo, informações assimétricas, incerteza, interação estratégica, normas sociais, custos de transação, externalidades e até mesmo a aprendizagem.

No que concerne a aprendizagem, alguns algoritmos têm surgido na literatura, sendo os mais utilizados os de aprendizagem por reforço (*Reinforcemente Learning* - RL) como, por exemplo, o *Q-Learning* e o *Roth-Erev RL*. Esses algoritmos são incorporados aos modelos de simulação com o objetivo de fazer com que o agente responda estrategicamente aos estímulos gerados por meio da interação com os demais agentes e com o ambiente. De acordo com Radhakrishnan et al. (2015), a aprendizagem por reforço torna a representação mais realista e proporciona melhores resultados. O processo adaptativo que compõe os algoritmos de aprendiza-

gem por reforço, consiste em um ciclo de ações que se repetem e são alimentadas com novas informações, obtidas ao longo das rodadas de simulação, que respaldam o agente no processo de tomada de decisão, tornando-o mais preciso.

Em cenários que envolvam agentes tomadores de decisão em situações de conflito, a aprendizagem se torna uma valiosa aliada na captação de resultados mais realistas. É de conhecimento que o acesso dos indivíduos a um conjunto de informações sobre os demais e sobre o ambiente torna possível que eles adotem comportamento estratégico, tendo em vista que quanto maior a quantidade de informação disponível, melhor ele pode se postar estrategicamente para a sua tomada de decisão (RINGLER; KELES; FICHTNER, 2016).

A teoria dos jogos, por sua vez, é uma ferramenta frequentemente utilizada na modelagem de situações de conflito. Ela pode ser utilizada para estudar assuntos recorrentes e relevantes, capaz de orientar o comportamento em situações de negociação e fornecer informações sobre como agir colaborativamente diante das condições do jogo (ARAUJO; LEONETI, 2018). Estudos empregando a teoria dos jogos vêm sendo desenvolvidos com ênfase na abordagem da concorrência e posicionamento estratégico de firmas (SANCHEZ-CARTAS, 2018), determinação de preço (HAFEZALKOTOB et al., 2018), decisões relacionadas à adoção de tecnologias (LI et al., 2018) e estratégias de investimento (MANTOVI; SCHIANCHI, 2018).

Este estudo, tem o objetivo de analisar qual o impacto decorrente da suposição de racionalidade limitada, considerando que os agentes são diferentes e que apenas um possui capacidade de aprendizado, sobre os equilíbrios teóricos conhecidos para os jogos clássicos do Dilema dos Prisioneiros, Batalha dos Sexos e *Chicken Game*. Ao incorporar a aprendizagem por meio dos algoritmos *Roth-Erev RL* (RE), *Modified Roth-Erev RL* (MRE) e *Variant Roth-Erev RL* (VRE), de forma a atribuir vantagem ao agente que tem capacidade de mapear o comportamento do outro, busca-se comparar o desempenho dos algoritmos utilizados, bem como identificar as implicações que eles atribuem ao resultado do jogo, tendo em vista que em jogos matriciais o resultado não depende única e exclusivamente da sua própria decisão, mas da combinação das decisões de ambos.

1.1 **Justificativa**

A escolha pela utilização da Modelagem Baseada em Agentes, foi embasada, principalmente, pelo diferencial obtido na representação dos agentes com características diferentes (ZHOU; CHAN; CHOW, 2009). Essa modelagem apresenta uma nova perspectiva sobre o conceito de racionalidade, empregando a premissa de racionalidade limitada proposta por Simon

(1955). A visão neoclássica da racionalidade trata a incerteza de maneira estrutural, ou seja, assume-se que somente há incerteza na probabilidade de que os eventos ocorram, admitindo a completa capacidade de se antecipar os resultados decorrentes das ações dos indivíduos. Ao abordar a racionalidade limitada, se aceita como verdade que o sistema é dinâmico, nesse sentido tanto o ambiente em que os indivíduos interagem podem apresentar mudanças, conforme o tempo passa, quanto os próprios indivíduos podem ter seus comportamentos modificados. A racionalidade limitada propõe ainda que os indivíduos não têm acesso a todas as informações relevantes ao processo de tomada de decisão, devido a complexidade de um sistema dinâmico. Dessa forma, os argumentos da racionalidade limitada se contrapõem aos defendidos pela racionalidade ilimitada ou substantiva, quando considera a incerteza como fator decisivo para os resultados futuros, do sistema como um todo, e não somente de forma estrutural (MELO; FUCIDJI, 2016).

Com o auxílio de algoritmos de aprendizagem, o comportamento do indivíduo é moldado e direcionado para decisões que maximizem o seu objetivo e, portanto, a sua satisfação, bem como a sua utilidade, considerando para isso a dinâmica do ambiente e as informações que lhes são disponibilizadas. Diferentes tipos de algoritmos de aprendizagem por reforço, podem ser utilizados para simular a inteligência dos agentes em um modelo de simulação. Os mais populares encontrados na literatura, dentro da classe de algoritmos de aprendizagem por reforço, incluem os algoritmos *Q-Learning* e *Roth-Erev RL*, bem como versões modificadas de ambos (ALIABADI; KAYA; SAHIN, 2017).

O presente estudo aplica conceitos da teoria dos jogos, bastante repercutida por matemáticos e economistas, tendo destaque no estudo de problemas econômicos devido às suas contribuições. Os jogos escolhidos para serem testados foram: Dilema dos Prisioneiros (DP), Batalha dos Sexos (BS) e *Chicken Game* (CG), por se tratarem de jogos clássicos com resultados canônicos fundamentados e equilíbrios que possibilitam identificar a alteração do comportamento em resposta ao aprendizado.

Trabalhos que analisam o processo de tomada de decisão, incorporando ao comportamento do agente um algoritmo de aprendizagem, foram desenvolvidos usando não somente o Dilema dos Prisioneiros bem como outros jogos de matriz 2 x 2 tão importantes quanto (ZSCHACHE, 2016).

Estudar esse tipo de problema sob a ótica da Modelagem Baseada em Agentes apoiada por algoritmos de aprendizagem, permite identificar se e como ocorrem as alterações comportamentais do agente diante do processo decisório e como os resultados são afetados. O estudo

em questão foi desenvolvido considerando três principais abordagens, sendo elas, os algoritmos de aprendizagem, eixo principal, a Modelagem Baseada em Agentes e a teoria dos jogos, eixos secundários, assim como mostra a Figura 1. A junção destas abordagens permite estudar a aprendizagem incorporada ao comportamento do decisor face à racionalidade limitada.

Figura 1 – Pilares do estudo

Fonte: A autora (2019)

Os ambientes simulados, situações de conflito representadas pelos jogos matriciais, foram implementados por meio da ABM, que possibilitou a construção dos agentes considerando seus respectivos processos de tomada de decisão. Já os algoritmos de aprendizagem nortearam o comportamento e permitiram a obtenção de informações acerca da influência da aprendizagem nos resultados globais para os três jogos.

Os processos de aprendizagem, incorporados na ação do agente, neste estudo, foram obtidos a partir do modelo desenvolvido por Pentapalli (2008) para analisar o comportamento resultante da interação entre vendedores e compradores empregando os algoritmos de aprendizagem RE, MRE e VRE.

1.2 **Objetivo geral**

O presente estudo tem como objetivo identificar quais as implicações da suposição de racionalidade limitada e, principalmente, dos processos de aprendizagem, nos equilíbrios teóricos dos jogos clássicos do Dilema dos Prisioneiros, Batalha dos Sexos e *Chicken Game*.

1.3 Objetivos específicos

- Implementar um modelo de simulação baseado em agentes incorporando os algoritmos de aprendizagem: *Roth-Erev RL Modified Roth-Erev RL* e *Variant Roth-Erev RL*;
- Identificar mudanças de comportamento emergentes da interação de ambos os agentes face aos algoritmos de aprendizagem por reforço;
- Comparar o desempenho dos algoritmos de aprendizagem utilizados no modelo de simulação.

1.4 Estrutura do trabalho

Este trabalho está organizado em seis capítulos, como segue:

- O Capítulo 1 - Introdução - é composto pela justificativa, objetivos sendo eles geral e específicos, bem como a descrição da estrutura do trabalho.
- O Capítulo 2 - Modelagem Baseada em Agentes - tem como objetivo fornecer ao leitor a base conceitual sobre a modelagem baseada em agentes e apresentar-lhe trabalhos da literatura que respaldem-no sobre a relevância do tema de estudo.
- O Capítulo 3 - Algoritmos de Aprendizagem - tem como objetivo fornecer ao leitor a base conceitual sobre os algoritmos que foram implementados no modelo de simulação bem como descrevê-los quantitativamente.
- O Capítulo 4 - Modelo de Simulação - descreve tanto a forma com que se deu a construção do modelo quanto o que se espera alcançar, evidenciando os resultados canônicos obtidos para os jogos clássicos.
- O Capítulo 5 - Resultados - apresenta de forma ilustrativa e descritiva os principais resultados obtidos por meio da simulação.
- O Capítulo 6 - Conclusão - contém as principais considerações sobre o trabalho que foi realizado.

2 MODELAGEM BASEADA EM AGENTES

A representatividade é um conceito muito comum, utilizado por pesquisadores em estudos que envolvam um grande número de agentes. Os agentes representativos (RA) refletem as mesmas características e por consequência desempenham o mesmo comportamento. Dilaver, Jump e Levine (2018) ressaltam que a ideia proporcionada pela suposição RA é a de que um único agente seja capaz de representar todo um setor econômico, ignorando qualquer viés de agregação ou heterogeneidade presente nas preferências dos agentes. É bastante comum que as interpretações do mercado sejam feitas a partir do comportamento do sistema como um todo e repassadas aos agentes que o compõem.

De uma maneira geral, se pode dizer que tal suposição traz simplificação à análise do estudo e que em muitos casos isto acaba não interferindo no objetivo proposto. No entanto, quando se busca um resultado que seja condizente com a realidade, é preferível abordar a heterogeneidade dos agentes assim como empregado nos modelos baseados em agentes. De acordo com Grimm et al. (2006) o acesso a esse tipo de modelo proporciona aos formuladores de políticas um ambiente de experimentação que permite avaliar o efeito de ações e políticas governamentais distintas.

Conforme ressaltado por Zhou, Chan e Chow (2009), a ABM é construída em torno de três elementos básicos: os agentes, o ambiente e as regras. Este último elemento é definido tanto para a interação entre os agentes como para a interação entre o agente e o ambiente. Por meio de interações repetidas, os agentes se modificam e modificam suas ações em resposta às ações dos demais agentes e do ambiente. Novos padrões de comportamento emergem deste aprendizado, daí a abordagem ser do tipo *bottom-up*, pois é a partir da avaliação dos comportamentos das unidades individuais e de como estes comportamentos se modificam mediante as interações que se torna possível obter o comportamento do sistema como um todo (BUSCH et al., 2017).

Corroborando com a crescente aplicação dessa técnica de análise em áreas diversas, novas ferramentas são desenvolvidas. Zhou, Chan e Chow (2009) e Abar et al. (2017) se encarregaram de realizar a classificação das ferramentas e softwares, respectivamente, utilizadas para a construção e execução dos modelos baseados em agentes, facilitando o desenvolvimento de estudos posteriores, considerando a diversidade de abordagens que são desenvolvidas e que acabam dificultando a visão geral do campo de pesquisa.

Em uma Modelagem Baseada em Agentes algumas características principais são definidas em relação aos agentes e ao ambiente. Esse tipo de modelagem é popularmente conhecida por ser capaz de simular um grande número de agentes heterogêneos e capturar o resultado

oriundo de suas interações. O termo baseado em agentes, implica simplesmente que o modelo de simulação é composto por agentes ou objetos com comportamento autônomo. Toda a simulação é voltada para o comportamento dos agentes ou objetos simulados, seja ele individual ou do sistema de uma maneira geral. Os agentes têm conhecimento sobre o ambiente com o qual interagem e essa interação é decorrente das regras iniciais que são definidas internamente.

As regras internas estimulam a interação e estão interligadas ao processo de tomada de decisão, movimento e ação dos agentes. Esse tipo de modelo tem sido indicado em situações contendo um grande número de indivíduos como, por exemplo, pessoas em multidões, agentes em mercados financeiros, humanos e máquinas em campos de batalha, personagens artificiais em jogos de computador e até mesmo situações cotidianas como a interação entre veículos e pedestres no trânsito. (SANCHEZ; LUCAS, 2002).

Com o objetivo de esclarecer como se dá a determinação e o funcionamento das regras de comportamento, em um modelo baseado em agentes, é considerada, como exemplo, uma representação utópica da interação entre pedestres e veículos. Enquanto os agentes representam os pedestres, o ambiente é composto por uma rua com faixa de pedestres e sinal de trânsito. O objetivo da simulação é levar o agente até o outro lado da rua. As regras internas para essa situação definem que o agente somente pode atravessar quando o sinal de trânsito estiver indicando a luz verde, a regra interna diz, então, a maneira com que o agente deve se comportar, ou seja, ele somente poderá se movimentar quando a luz indicada pelo semáforo for verde. Suponhamos que a regra geral se modifique, e que os agentes também possam atravessar quando o fluxo de carros diminuir, para qualquer cor indicada pelo semáforo, no entanto somente atravessarão os agentes com níveis altos de propensão ao risco, enquanto que os agentes com aversão ao risco permanecerão reagindo somente à primeira regra.

Nessa situação, o comportamento face ao risco é incorporado como uma característica dos agentes, mas em níveis diferentes, o que faz com que cada um, de acordo com o seu comportamento individual, responda diferentemente aos estímulos. Observe que a regra interna é estabelecida a todos os agentes, mas devido às diferenças presentes em suas características, o comportamento individual não é o mesmo para todos. Esta suposição explica o porquê dos comportamentos individuais serem distintos mesmo que as regras gerais sejam as mesmas.

Um exemplo um pouco mais completo do impacto que as regras de comportamento têm sobre os agentes é dado por Shiflet e Shiflet (2014). Eles enfatizam que o comportamento de um consumidor diante da decisão de comprar ou não um determinado bem depende de uma gama de informações. A decisão pode ser baseada, principalmente, em necessidade, no

entanto também pode levar em consideração o otimismo em relação a economia ou inflação, caso o bem seja de alto valor. A situação financeira dele também é levada em consideração, bem como a percepção que ele tem sobre o negócio que está prestes a fazer. Logo, é possível entender que as regras de comportamento podem ser tão simples quanto abrangentes e atribuir maior complexidade ao processo decisório. Nesse sentido, a ação de um agente depende de sua situação, do ambiente em que o mesmo está e de suas regras de comportamento. Portanto, os comportamentos previstos pela economia real em que o comportamento do mercado emerge dos comportamentos individuais dos agentes que interagem passam a ser refletidos com maior precisão (MACAL; NORTH, 2010).

O desenvolvimento de categorias comportamentais e o escalonamento delas para toda a população de agentes, segundo Smajgl et al. (2011), são duas etapas de fundamental importância para a construção de um modelo baseado em agentes que simula respostas comportamentais humanas. Para isso, dados empíricos são necessários. Bell (2017) enfatiza que incorporar o processo de tomada de decisão por meio das regras de comportamento pode ser visto tanto como a maior força desse tipo de modelo quanto como a maior fragilidade. A segunda afirmação se deve ao fato de que a representação, com detalhes, dos processos de decisão pode se tornar um desafio quando se trata de informação empírica.

Quanto às ferramentas úteis na coleta de dados para informar e dar suporte aos processos de decisão em um ABM, pode-se contar com uma variedade delas, com vantagens e desvantagens intrínsecas a cada uma. A escolha pelo tipo de ferramenta pode refletir preocupações com as metas propostas pelo estudo, custo, tempo, entre outros (BELL, 2017). Não há para isso uma ferramenta e uma abordagem padrão que dê o suporte empírico necessário. Por outro lado, uma série de métodos incluem conhecimento especializado, pesquisas, entrevistas e até mesmo observação, que auxiliam na caracterização empírica e parametrização dos processos decisórios, podendo haver, ainda, a combinação de mais de um método para o alcance dos objetivos (SMAJGL et al., 2011).

Dentro da classe de modelos baseados em agentes, encontram-se aqueles utilizados para fim de prova de conceito - neles prevalecem a ausência de apoio empírico e a simplicidade nas regras, e aqueles utilizados para análise de eventos específicos - que apresentam uma demanda maior de detalhes (SUN et al., 2016). Os ABM's de maior popularidade são muito simples e têm como objetivo fornecer informações sobre padrões gerais. O nível de complexidade de um ABM de acordo com Grimm et al. (2006) tem relação com o detalhamento da estrutura do modelo e sua composição: número e tipos de entidades, processos e interações. Quanto maior a

necessidade de detalhe para a composição do modelo e, principalmente, para a composição do processo decisório dos agentes, mais complexidade é atribuída ao mesmo.

Modelos que exigem uma grande quantidade de dados empíricos contradizem afirmações que destacam os ABM's como de prova de conceito apenas. Essa classe de modelos foi impulsionada, principalmente, pela demanda de tomadores de decisão apoiado pelo crescente avanço da computação. No entanto uma significativa parcela do sucesso desses modelos é justificada pela sua capacidade de representar sistemas emergentes por meio de regras simples e especificadas a nível individual (SUN et al., 2016). Um especial foco tem chamado a atenção dos autores, para a aplicação de modelos baseados em agentes, sendo ele o mercado de energia elétrica, que abrange problemáticas existentes desde a geração até a comercialização, assim como demonstrado pelas revisões críticas realizadas por Sensfub et al. (2007), Weidlich e Veit (2008) e Guerci, Rastegar e Cincotti (2010).

Uma das principais qualidades desta técnica de análise é a flexibilidade em relação a sua aplicabilidade em áreas de estudo completamente diferentes. É comum que os modelos baseados em agentes sejam aplicados em contextos econômicos que envolvam o processo decisório entre compradores e vendedores ou simulando comportamentos referentes ao consumo em um determinado mercado. No entanto também são encontrados trabalhos que consideram um contexto econômico com foco no comportamento social, como, por exemplo, as desigualdades sociais resultantes de um sistema econômico. Blok, Lenthe e Vlas (2018) propuseram um modelo que pudesse ser usado como prova de conceito para o estudo do impacto de medidas preventivas e intervenções à crescente desigualdade financeira que atinge a participação esportiva de um sistema como um todo.

Aplicações desses modelos, também podem ser encontradas em estudos voltados para a área da saúde (TRACY; CERDA; KEYES, 2018), com foco inclusive na saúde pública, bem como podem ser utilizados para estudar fenômenos como a migração humana (THOBER; SCHWARZ; HERMANS, 2018), o terrorismo (MOYA et al., 2017), a disseminação de epidemias (CLIFF et al., 2018) e o desenvolvimento biológico (ALEDO et al., 2018).

Dentro da grande massa de trabalhos encontrados na literatura aplicando os conceitos de Modelagem Baseada em Agentes, os mais comuns são aqueles voltados para áreas da ciência da computação ou tecnologia da informação. Exemplos desse tipo de aplicação, oriundos de produção nacional, são apresentados por Girardi, Marinho e Oliveira (2005) que estudaram a aplicabilidade dos modelos AB no estudo de padrões de softwares. Na mesma linha de pesquisa Boukerche et al. (2007) abordaram a segurança em redes de computadores, enquanto Farias e

Santos (2005) optaram por desenvolver um sistema de informação geográfica.

A segunda área de estudo mais comum, entre os trabalhos encontrados na literatura que utilizam esses modelos, é a área econômica que aborda conceitos políticos e da Teoria dos Jogos. Dentro dessa área, motivados pela situação de crise financeira e econômica brasileira, Streit e Borenstein (2009) desenvolveram um estudo utilizando ABM para avaliar o Sistema Financeiro Brasileiro dado que de acordo com os autores uma das principais causas da crise está ligada às deficientes estruturas de governança. Mais recentemente, Alexandre e Lima (2017) combinaram, por meio dos modelos AB, conceitos de política monetária e regulamentação prudencial estudando situações de estabilidade financeira. Por sua vez, Crepaldi, Ferreira e Rodrigues (2012) abordaram o comportamento de agentes vendedores e compradores no mercado financeiro por meio do Jogo da Minoria.

Durante a pesquisa por trabalhos que tenham utilizado em conjunto as características da Modelagem Baseada em Agentes e processos de aprendizagem, percebeu-se que a participação nacional é pequena, evidenciando que não se tratou de uma pesquisa exaustiva. Quando a busca por trabalhos envolveu o termo Teoria dos Jogos, além dos termos mencionados anteriormente, nenhum trabalho de produção nacional pôde ser encontrado. Ao utilizar a base de dados *Web of Science* para procurar por trabalhos com o termo *Agent-based*, em nível mundial, mais de 14 mil trabalhos foram encontrados, quando o termo *Reinforcement learning* foi adicionado à pesquisa, o resultado foi reduzido para 163 trabalhos: destes, 41,10% correspondem aos últimos cinco anos (2014 - 2018), se tratando, portanto, de um campo com vastas oportunidades de estudos a serem exploradas.

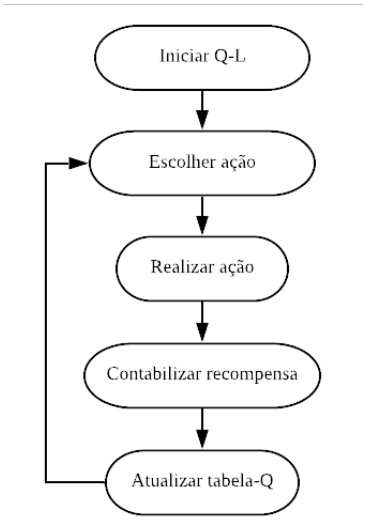
3 ALGORITMOS DE APRENDIZAGEM

Os principais algoritmos de aprendizagem por reforço, encontrados na literatura, são os denominados *Q-Learning* e *Roth-Erev RL*. A aprendizagem por reforço recebe esse nome por ser baseada em informações que são reforçadas a cada rodada de simulação, atribuindo maior força aos posicionamentos que devem ser adotados em cada situação. Apesar de pertencerem à mesma classe de algoritmos, o aprendizado se dá diferentemente em cada um deles. Enquanto o algoritmo *Roth-Erev RL*, tem a aprendizagem baseada em um valor de probabilidade, o *Q-Learning*, é atualizado com base em estados e ações que compõem a função Q-valor, atualizada a cada rodada de simulação, sendo assim responsável por direcionar o comportamento do agente.

Uma das principais desvantagens do algoritmo *Q-Learning*, está na determinação dos estados e ações, pois acaba sendo um processo complexo a depender do objetivo do estudo. Weidlich e Veit (2008), evidenciam que na maioria dos estudos em que esse algoritmo foi implementado não houve qualquer justificativa acerca de como os estados e ações foram definidos.

Neste estudo, somente o algoritmo *Roth-Erev RL* e suas duas versões modificadas, foram utilizados para simular o comportamento inteligente do agente. A Figura 2, apresenta a sequência lógica de execução do algoritmo *Q-Learning* que tem como ponto central a tabela-Q, composta pelos valores de recompensas correspondentes a cada combinação de estados e ações.

Figura 2 – Sequência de execução do algoritmo *Q-Learning*

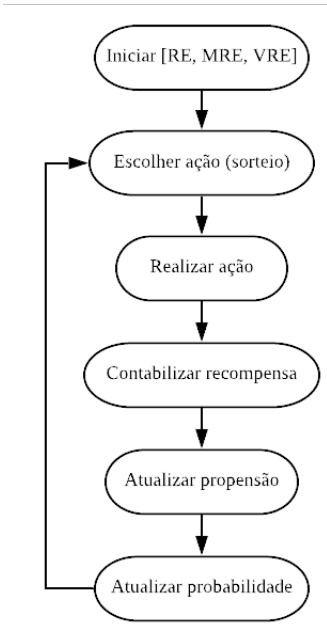


Fonte: A autora (2019)

Assim como mencionado anteriormente, o algoritmo *Roth-Erev RL* implementa o processo decisório baseado em um valor de probabilidade, da mesma forma é feito para os

algoritmos *Modified Roth-Erev RL* e *Variant Roth-Erev RL*. Todas as ações disponíveis ao agente são consideradas no cálculo de atualização da propensão e por consequência no cálculo da probabilidade. A lógica de funcionamento desses algoritmos é apresentada na Figura 3.

Figura 3 – Sequência de execução dos algoritmos RE, MRE e VRE



Fonte: A autora (2019)

Apesar de serem algoritmos diferentes, a lógica de funcionamento permanece a mesma para as três versões do *Roth-Erev*. As alterações que originaram as versões modificadas do RE são referentes aos termos que compõem as equações de atualização da propensão e de atualização da probabilidade de escolha. No entanto, a sequência com que se dá o processo de tomada de decisão bem como o surgimento da aprendizagem mantém-se para as três versões.

Os algoritmos de aprendizagem podem ser incorporados ao comportamento dos agentes em qualquer que seja o ambiente de simulação. Ao esclarecer as regras de comportamento, o algoritmo determinará como os agentes devem se adaptar e tomar decisões, proporcionando o alcance de seus objetivos (WEIDLICH; VEIT, 2008).

Os processos de aprendizagem mencionados anteriormente, têm sido aplicados, principalmente, para analisar o mercado de energia elétrica, devido a complexidade presente nessa área de estudo. Calabria, Saraiva e Rocha (2018), por exemplo, empregaram o algoritmo *Q-Learning* na simulação empírica do mercado de energia elétrica brasileiro. Já Gaivoronskaia e Tsyplakov (2018) utilizaram o algoritmo de aprendizagem *Roth-Erev RL* no modelo baseado em agentes desenvolvido para analisar o mercado russo de energia elétrica.

O mercado de energia elétrica pode ser estudado sob diferentes perspectivas, a partir

da combinação entre ABM e algoritmos de aprendizagem, como, por exemplo, abordando a negociação de contratos de energia (RODRIGUEZ-FERNANDEZ et al., 2018), o impacto das energias renováveis no preço e no equilíbrio dos mercados de energia elétrica (MUREDDU; MEYER-ORTMANNS, 2018), bem como o processo de licitação de Companhias Geradoras (*GenCo's*) (RADHAKRISHNAN et al., 2015). Enquanto o primeiro trabalho mencionado empregou o algoritmo *Q-Learning*, os outros dois utilizaram versões modificadas do algoritmo *Roth-Erev RL*.

As seções 3.1, 3.2 e 3.3, respectivamente, descrevem mais detalhadamente as equações que dão origem aos processos de aprendizagem dos algoritmos *Roth-Erev RL*, *Modified Roth-Erev RL* e *Variant Roth-Erev RL*.

3.1 **Roth-Erev RL**

Criado em 1998 por Roth e Erev, esse algoritmo emprega três parâmetros diferentes em sua composição: experimentação (ϵ), esquecimento (ϕ) e propensão (q_{nj}) do agente n escolher a ação j. O último parâmetro mencionado, é atualizado conforme mostra a Equação (1).

$$\mathrm{q}_{nj}(t+1) = \begin{cases} (1-\phi)q_{nj}(t) + R(j)(1-\epsilon) & se \quad j = k; \\ (1-\phi)q_{nj}(t) + R(j)\frac{\epsilon}{M-1} & se \quad j \neq k. \end{cases} \tag{1}$$

em que

- n representa o n-ésimo jogador;
- j é a ação a ser escolhida;
- t é o tempo atual;
- $R(j)$ é a recompensa correspondente a ação escolhida;
- M é o número total de escolhas possíveis.

A Equação (1) apresenta a propensão do jogador n escolher a ação j no tempo $(t+1)$. Dizer que a ação escolhida j é igual a k é o mesmo que dizer que durante aquela rodada de simulação a ação j foi adotada pelo agente que aprende. Essa propensão pode ser atualizada de duas maneiras: i) se a ação j escolhida pelo jogador n for igual a k então a propensão é atualizada de acordo com a equação que se dá pela soma entre a propensão no tempo atual, ponderada pelo parâmetro de esquecimento, e a recompensa obtida com a ação j, ponderada pelo parâmetro de

experimentação; ii) se a ação j do jogador n for diferente de k, ou seja, a ação j não foi adotada naquela rodada de simulação, a propensão se atualiza de acordo com a equação dada pela soma entre a propensão no tempo atual, ponderada pelo parâmetro de esquecimento, e a recompensa, ponderada pela divisão entre o parâmetro de experimentação e o número de escolhas possíveis subtraída de 1. A Equação (1) é formulada em função de uma única ação denominada como j, no entanto ao considerar mais ações disponíveis ao agente, o mesmo cálculo deve ser feito para todas as demais.

A capacidade de resposta do algoritmo é influenciada por meio dos parâmetros de experimentação e esquecimento. O primeiro permite que tanto as escolhas bem sucedidas quanto aquelas similares que ocorrem com frequência sejam escolhidas, possibilitando, portanto, a experimentação no processo decisório, evitando que o agente se fixe na primeira ação que lhe proporcione recompensa positiva. Já o parâmetro de esquecimento fornece maior peso à experiências recentes em comparação a experiências passadas, dessa forma decisões tomadas recentemente tem maior probabilidade de serem repetidas (EREV; ROTH, 1998). Como pode ser observado na Equação (1), a maneira como a recompensa é ponderada sofre alteração quando a ação não é adotada, isso ocorre para que o valor de recompensa seja reduzido e com isso as chances de que essa ação seja adotada novamente sejam, também, menores. No entanto, caso a recompensa seja igual a 0, a determinação da ação que será adotada reflete a força da propensão, associada a cada ação, no período atual (t), já que os demais termos assumirão valor igual a 0.

A cada nova rodada de simulação uma nova propensão é calculada e duas variáveis principais desempenham um papel importante no cálculo desta. A primeira variável é a propensão associada à decisão tomada no período passado e a outra é a recompensa correspondente à ação adotada nesse mesmo período. Os parâmetros mencionados anteriormente atuam na ponderação desses valores, dando maior ou menor peso de acordo com o desejado. Foi definido para este estudo, com base no estudo de Pentapalli (2008), que o intervalo de variação empregado para os parâmetros de experimentação e esquecimento, será de 0 a 0,1.

No tempo $t = 1$ cada agente tem uma propensão inicial de adotar uma estratégia, quando o aprendizado é incorporado ao comportamento de mais de um agente, eles recebem igual valor de propensão inicial. Tendo em vista que o comportamento de escolha, de acordo com esse algoritmo, é probabilístico, é assumido, que na primeira rodada da simulação, a probabilidade de escolha associada às estratégias é igual para todos os agentes. A probabilidade de escolha

($p_{nk}(t)$) é dada pela Equação (2).

$$p_{nk}(t) = \frac{q_{nk}(t)}{\sum_{j=1}^{M} q_{nj}(t)} \tag{2}$$

em que

- $q_{nk}(t)$ é a propensão do n-ésimo jogador escolher a ação k no tempo t;
- $\sum_{j=1}^{M} q_{nj}(t)$ é o somatório das propensões associadas às ações disponíveis ao jogador n, no tempo t.

De forma a esclarecer como o aprendizado ocorre, o jogo DP é tomado como exemplo. Por se tratar de um jogo de matriz 2 x 2, o agente dispõe de apenas duas ações: Não Delatar (ND) e Delatar (D), considerando essas decisões, o seguinte questionamento é feito: a ação ND é igual a k? Uma resposta positiva para esse questionamento indica que a ação que está sendo adotada naquela rodada de simulação é ND, dessa forma a propensão para essa ação deve ser atualizada levando em consideração que $j = k$. Caso contrário, se a resposta for não, indicando que a ação que está sendo adotada na rodada de simulação é D e não ND, a propensão correspondente à ND será atualizada considerando $j \neq k$, do mesmo modo é feito para a ação D.

Feito isso, as propensões correspondentes a cada uma das ações são obtidas, tendo em vista que uma foi adotada e a outra não. As probabilidades são então calculadas com base nos valores de propensão por meio da divisão entre a propensão atual e o somatório das propensões atribuídas às duas ações. A ação que tiver o maior valor de probabilidade terá então maior chance de ser adotada na próxima rodada de simulação, por se tratar de um processo de aprendizagem probabilístico um sorteio é feito a cada nova rodada para determinar a ação que deve ser adotada.

A cada rodada de simulação, os questionamentos são refeitos e novos valores de propensão são atribuídos a cada ação, consequentemente novas probabilidades são calculadas, gerando então o que é chamado de aprendizagem por reforço, já que à medida que as ações são repetidas, mais fortemente uma ação pode ser determinada como preferível recebendo assim maiores valores de probabilidade a cada rodada de simulação, tornando mais claro ao agente qual decisão deve ser adotada. O princípio de aprendizagem descrito se mantém para as demais versões do algoritmo, que serão apresentadas na sequência.

3.2 **Modified Roth-Erev RL**

Ao utilizar o algoritmo *Roth-Erev RL* em um de seus estudos, Nicolaisen, Petrov e Tesfatsion (2001) mencionaram uma característica, denominada por eles como problemática,

presente no algoritmo original. A partir desta constatação os autores desenvolveram uma nova formulação para o algoritmo, agora chamado de *Modified Roth-Erev RL*. A crítica feita pelos autores consiste em ausência de atualização das propensões em casos de recompensa igual a 0, uma vez que restará apenas a propensão atual ponderada pelo parâmetro de esquecimento. Com base nisso os autores propuseram uma versão modificada do algoritmo original, como consta na Equação (3) (NICOLAISEN; PETROV; TESFATSION, 2001).

$$\mathrm{q}_{nj}(t+1) = \begin{cases} (1-\phi)q_{nj}(t) + R(j)(1-\epsilon) & se \quad j = k; \\ (1-\phi)q_{nj}(t) + q_{nj}(t)\frac{\epsilon}{M-1} & se \quad j \neq k. \end{cases} \tag{3}$$

O termo $R(j)$ é substituído por $q_{nj}(t)$ quando $j \neq k$. Por meio desta formulação, a atualização da propensão é percebida mesmo quando a ação escolhida resulta em recompensa nula. O cálculo da probabilidade de escolha, por sua vez, permanece de acordo com a Equação (2) apresentada na seção anterior.

3.3 **Variant Roth-Erev RL**

A segunda variação encontrada para o algoritmo original, desenvolvida por Sun e Tesfatsion (2007), propõe uma modificação no cálculo das probabilidades de escolha a partir da implementação da distribuição de *Boltzmann*, como mostra a Equação (4). Entretanto, mantém as mesmas características do MRE para a atualização da propensão.

$$p_{nk}(t) = \frac{e^{\frac{q_{nk}(t)}{T}}}{\sum_{j=1}^{M} e^{\frac{q_{nj}(t)}{T}}} \tag{4}$$

De acordo com Weidlich e Veit (2008), é comum que em vez da Equação (2) alguns pesquisadores utilizem a distribuição de *Boltzmann* para determinar a probabilidade de escolha, incluindo na equação um parâmetro de temperatura T também conhecido como parâmetro de resfriamento. Esse parâmetro determina o foco em ações com valores altos de propensão que correspondem, consequentemente, às ações com valores altos de recompensa.

4 MODELO DE SIMULAÇÃO

O modelo de simulação foi escrito em *NetLogo*, *toolkit* utilizada para simular fenômenos complexos, principalmente voltada para o desenvolvimento de modelos baseados em agentes (TISUE; WILENSKY, 2004). Os pacotes desenvolvidos para a modelagem baseada em agentes são divididos em duas classes: *toolkit* e *software*. O *NetLogo*, que surgiu em resposta às limitações do *StarLogo*, encontra-se dentro da classe denominada como *toolkit* responsável por fornecer bibliotecas e funções projetadas especificamente para sistemas baseados em agentes (ZHOU; CHAN; CHOW, 2009). Além de ser uma linguagem bastante intuitiva, a plataforma, tanto *online* quanto *offline*, possui uma interface que permite ilustrar o ambiente que está sendo modelado auxiliando na explicação do modelo.

De acordo com Macal e North (2010) os modelos encontrados nas bibliotecas de *toolkits*, como o *NetLogo*, facilitam a programação fazendo com que não seja necessário programar a rotina a partir do zero. Ao aproveitarem os modelos disponibilizados pela plataforma, os pesquisadores ganham tempo no desenvolvimento de seus modelos. Dentre as áreas de estudo nas quais os modelos desenvolvidos em *NetLogo* podem ser aplicados, são encontradas, por exemplo, a área econômica, de comportamento organizacional e empresarial, ciências sociais e naturais e até mesmo ecologia, assim como destaca Abar et al. (2017).

Para o estudo em questão, jogos de matriz 2 x 2 foram testados utilizando a modelagem baseada em agentes. Nestes jogos dois agentes interagem simultaneamente, levando em consideração que a decisão de um impacta no ganho do outro e vice versa. Cada agente tem apenas duas decisões, a depender de cada jogo. Os jogos simulados foram: Dilema dos Prisioneiros (DP), Batalha dos Sexos (BS) e *Chicken Game* (CG). A escolha de tais jogos se deu por serem jogos bastante conhecidos na Teoria dos Jogos e muito utilizados para estudar conflitos entre firmas competitivas, decisões estratégicas como, por exemplo, implementação de uma nova tecnologia e até mesmo dilemas envolvendo questões sociais.

O DP retrata o conflito entre duas pessoas que buscam a melhor satisfação, mas que por motivos egoístas acabam tomando uma decisão que as impede de obter o melhor resultado da matriz de incentivos. A Tabela 2 apresenta a matriz de incentivos correspondente a esse jogo, com as estratégias de Não Delatar (ND) e Delatar (D) e seus respectivos *payoffs*.

A matriz de incentivos foi montada levando em consideração que maiores valores representam o melhor resultado para os jogadores. O melhor cenário para ambos os jogadores é quando os dois tomam a mesma decisão de não delatar o crime, ou seja, estratégia ND_1ND_2. Ao atingirem essa combinação de estratégias, ambos os agentes recebem *payoff* igual a 5. Em

Tabela 1 – Matriz de incentivos do jogo Dilema dos Prisioneiros

	ND_2	D_2
ND_1	(5, 5)	(0, 10)
D_1	(10, 0)	(1, 1)

Fonte: A autora (2019)

contrapartida, um resultado que não representa um bom cenário está na combinação de estratégias D_1D_2, já que ambos optam por delatar o crime.

Nas outras duas combinações de estratégia, ND_1D_2 e D_1ND_2, pelo menos um dos jogadores recebe o maior valor de recompensa da matriz de incentivos, 10, enquanto o outro recebe o menor, 0. Esse resultado justifica o motivo pelo qual os agentes optam por delatar, pois, na grande maioria dos casos, em pelo menos um cenário resultante dessa estratégia um dos agentes obtém a maior recompensa da matriz de incentivos. O Equilíbrio de Nash (EN) neste jogo está justamente na combinação de estratégias: D_1D_2, pois a estratégia Delatar é uma estratégia dominante, o que faz com que ela seja preferível a qualquer outra.

Assim como o DP, o jogo conhecido como Batalha dos Sexos também é um clássico da Teoria dos Jogos. Esse jogo, representa uma situação de conflito em que ambos os agentes precisam entrar em acordo sobre suas decisões para que eles possam alcançar o melhor cenário da matriz de incentivos. O casal precisa escolher entre cinema (C) e futebol (F), por exemplo, e cada um dos dois tem preferência por uma das duas opções, discordando um do outro e ao mesmo tempo desejando que suas escolhas coincidam. A Tabela 3 apresenta a matriz de incentivos para este jogo contendo as estratégias e as recompensas correspondentes a elas.

Tabela 2 – Matriz de incentivos do jogo Batalha dos Sexos

	F_2	C_2
F_1	(2, 3)	(0, 0)
C_1	(0, 0)	(3, 2)

Fonte: A autora (2019)

Ao analisar a matriz de incentivos nota-se que as situações em que os jogadores discordam entre si, combinações de estratégias F_1C_2 e C_1F_2, não acrescentam valor algum aos mesmos. Nesse sentido, é melhor, para ambos os agentes, escolherem a opção de preferência do parceiro do que discordarem. Ao contrário do primeiro problema apresentado, o BS apresenta dois equilíbrios, correspondentes às combinações de estratégias F_1F_2 e C_1C_2, em que ambos recebem recompensas positivas, mas ao realizar a sua opção de lazer preferível o agente é melhor

recompensado.

O terceiro jogo, *Chicken Game*, retrata a competição entre dois indivíduos em que para que seja possível obter *payoffs* positivos, é necessário que os agentes tomem decisões contrárias. As decisões estratégicas consistem em desistir (D) e seguir (S).

Assim como o jogo representado anteriormente, este também possui dois equilíbrios. A Tabela 4 apresenta a matriz de incentivos correspondente ao CG.

Tabela 3 – Matriz de incentivos do jogo *Chicken Game*

	D_2	S_2
D_1	(3, 3)	(2, 4)
S_1	(4, 2)	(0, 0)

Fonte: A autora (2019)

Os equilíbrios para esse jogo estão nas combinações de estratégias S_1D_2 e D_1S_2, situações em que um dos jogadores desiste e o outro consegue a vitória ao fim da disputa. O cenário D_1D_2 indica que ambos desistem. Apesar de não ser a melhor situação para os jogadores, é benéfica a eles, sendo assim uma estratégia preferível à combinação de estratégias S_1S_2. Essa combinação resulta nos *payoffs* mais baixos da matriz de incentivos, não sendo vantajoso a nenhum dos jogadores envolvidos na situação de conflito.

No modelo de simulação as estratégias/ações disponíveis aos agentes foram nomeadas como Ação 0 e Ação 1. As ações 0 e 1 assumem estratégias distintas a depender do jogo que está sendo executado, a Tabela 5 indica a correspondência entre as ações 0 e 1 e as estratégias de cada jogo. Os algoritmos de aprendizagem foram incorporados ao comportamento de apenas um dos agentes, de forma a identificar se o posicionamento do agente que aprende se modifica em decorrência do aprendizado. Os agentes foram denominados como *Intelligent Player* (IP) e *Player* (P). O processo decisório do *Player* é escolhido pelo usuário na interface do modelo de simulação, Figura 4, podendo ser: somente Ação 0, somente Ação 1, ou Agir aleatoriamente, tendo em vista que a última consiste em um sorteio aleatório entre as opções disponíveis. Já o comportamento decisório do *Intelligent Player* é obtido por meio dos algoritmos aprendizagem: RE, MRE e VRE.

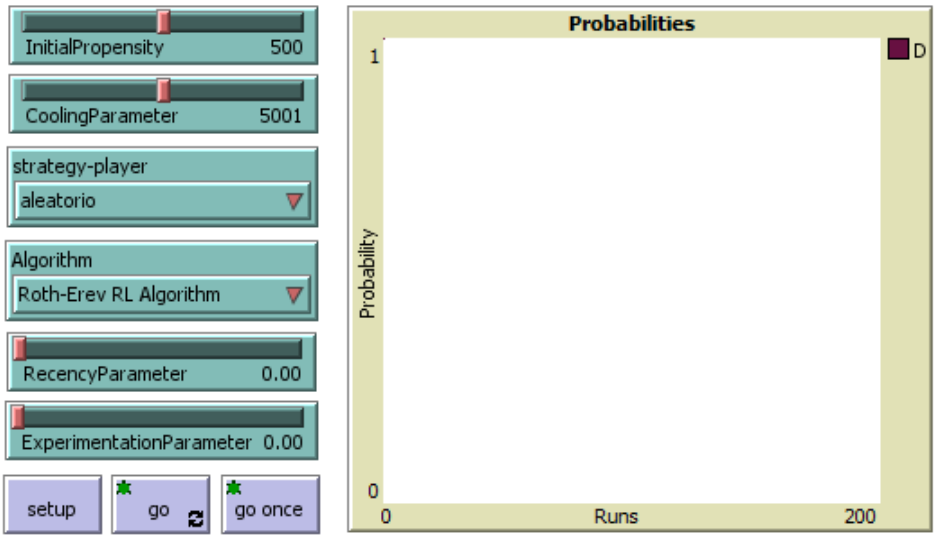
Tabela 4 – Determinações das ações 0 e 1

Ações/Estratégias	DP	BS	CG
Ação 0	Não Delatar	Futebol	Desistir
Ação 1	Delatar	Cinema	Seguir

Fonte: A autora (2019)

Também na interface do modelo, o usuário escolhe qual dos três algoritmos incorporar ao comportamento do agente que aprende. A partir de então o comportamento do agente se atualiza automaticamente, de acordo com as ações disponíveis. A partir do momento em que os agentes escolhem suas estratégias, o resultado da interação de ambos é obtido e as recompensas correspondentes às combinações de estratégias, alimentam as equações de atualização do algoritmo. A probabilidade de escolha associada a cada estratégia é projetada na interface do modelo, é por meio dela que o padrão de comportamento do agente que aprende pode ser interpretado.

Figura 4 – Interface do modelo de simulação desenvolvido em *NetLogo*



Fonte: A autora (2019)

Assim como o modo de ação dos agentes, os parâmetros de esquecimento, experimentação, resfriamento e propensão inicial, que compõem a equação dos algoritmos de aprendizagem, também são modificados a partir da interface do modelo, como mostra a Figura 4. Além dos botões que permitem alterações, a interface conta com botões que permitem restaurar as configurações do modelo a cada modificação, *setup*, e iniciá-lo, *go* e *go once*.

Tanto o parâmetro de experimentação quanto o parâmetro de esquecimento, foram definidos para variar de 0 a 0,1 sendo acrescidos de 0,01 a cada modificação. A propensão inicial varia de 0 a 1000 e é atualizada no decorrer das rodadas de simulação, sendo que quando é mantida em 0 não apresenta alteração na escolha do agente IP que é programado inicialmente para adotar a estratégia dita no modelo como ação 0. O parâmetro de resfriamento, por sua vez, assume valores maiores que 0 e menores que 10000. Os valores e limites de variação foram definidos de acordo com os utilizados por Pentapalli (2008).

A partir da determinação das combinações de estratégias, resultantes das ações adotadas

por ambos os agentes, foi possível determinar as recompensas correspondentes a cada resultado possível em cada um dos três jogos, a Figura 5 apresenta um recorte da rotina de programação em que as combinações são descritas.

Figura 5 – Combinações de estratégias

```
to DetermineCombinations
  if strategy = 0 and CurrentAction = Action 0 [set Combination "R"]
  if strategy = 0 and CurrentAction = Action 1 [set Combination "S"]
  if strategy = 1 and CurrentAction = Action 0 [set Combination "T"]
  if strategy = 1 and CurrentAction = Action 1 [set Combination "P"]
end
```

Fonte: A autora (2019)

As estratégias disponíveis ao agente IP foram denominadas como ações 0 e 1, enquanto as estratégias disponíveis ao agente P foram denominadas como estratégias 0 e 1. A forma com que a determinação das combinações foi escrita permitiu que ela fosse utilizada para os três jogos. A Figura 5 indica uma letra designada para cada combinação de estratégias: R, S, T e P. Estas letras foram utilizadas com o objetivo de proporcionar simplificação ao modelo, dessa forma, se necessário, elas podem ser substituídas por outras que apresentem maior adequação.

Posteriormente à determinação das combinações, os respectivos valores que compõem a matriz de incentivos foram considerados para determinar a recompensa correspondente a cada resultado do jogo, como mostra a Figura 6.

Figura 6 – Determinação das recompensas na rotina de programação

```
to DetermineProfits
  ask IntelligentPlayers [if Combination = "R" [set Profit 5]]
  ask IntelligentPlayers [if Combination = "S" [set Profit 10]]
  ask IntelligentPlayers [if Combination = "T" [set Profit 0]]
  ask IntelligentPlayers [if Combination = "P" [set Profit 1]]
end
```

Fonte: A autora (2019)

Os recortes da rotina de programação apresentados nas Figuras 5 e 6 são correspondentes ao modelo de simulação do jogo Dilema dos Prisioneiros, a forma com que a rotina foi escrita permitiu o escalonamento da mesma para os demais jogos por meio da alteração dos valores de recompensa oriundos de cada resultado. A rotina de programação, completa, pode ser encontrada no Apêndice A.

5 RESULTADOS

Os gráficos apresentados nas seções a seguir foram gerados na interface do *NetLogo*, como saída do modelo de simulação. Eles demonstram o comportamento do agente inteligente dado o comportamento de resposta do agente que não aprende. Nos eixos horizontal e vertical, estão representadas as rodadas de simulação e os valores de probabilidade, respectivamente. Sabe-se que em jogos de matriz 2 x 2 os agentes podem escolher entre duas ações, no entanto, uma única ação foi escolhida, a depender de cada jogo, para ser representada nos gráficos gerados para cada diferente cenário de simulação. Essa escolha trouxe simplicidade à representação gráfica, facilitando a interpretação.

As estratégias Futebol e Desistir, respectivamente, foram escolhidas para serem representadas nos gráficos gerados para os jogos BS e CG. Os gráficos gerados para o DP mostram o comportamento dos valores de probabilidade associados à estratégia "Delatar". Enquanto valores altos de probabilidade indicam que essa estratégia está sendo cada vez mais adotada no decorrer das rodadas de simulação, valores baixos representam a situação contrária.

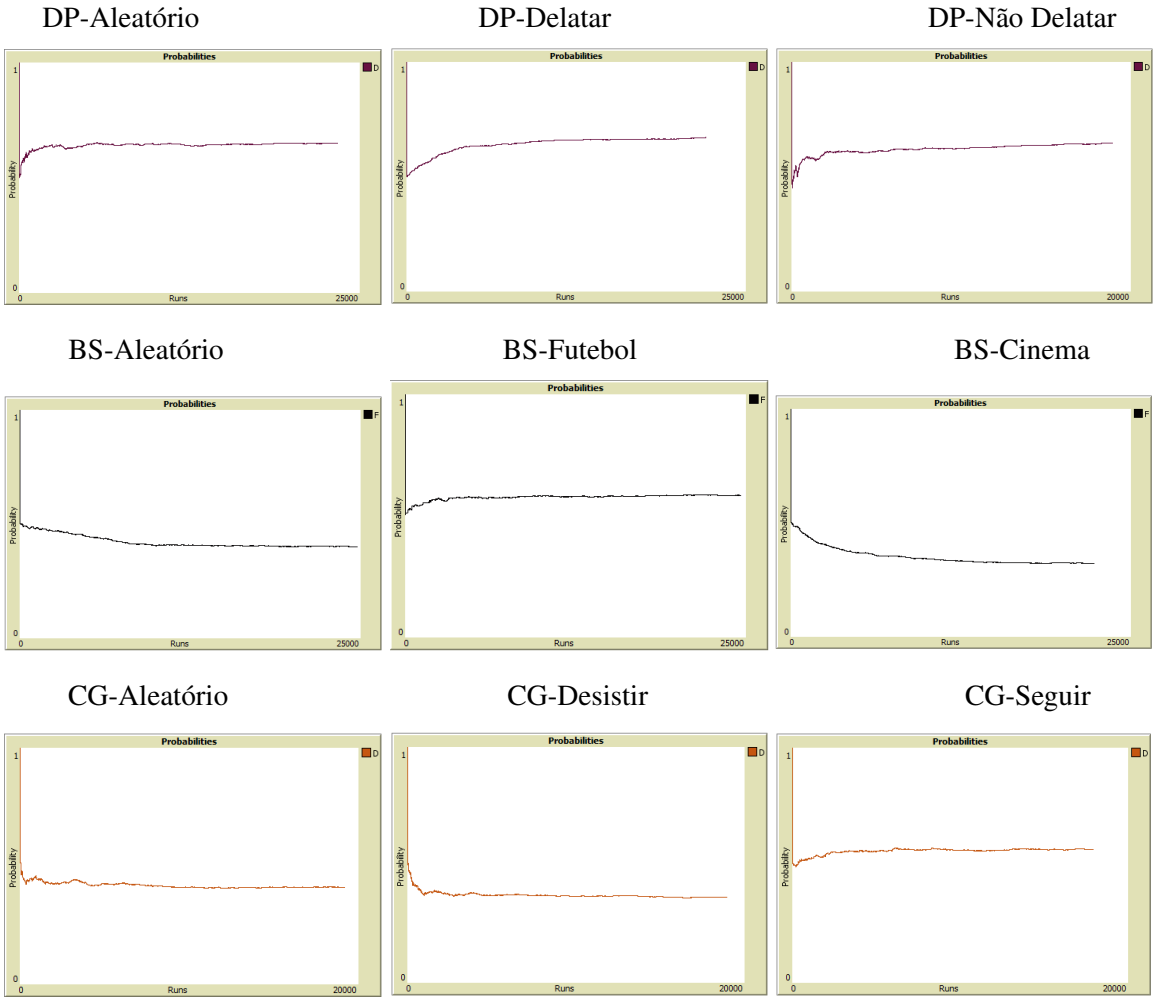
A fim de identificar as influências dos parâmetros de experimentação e esquecimento, nos três processos de aprendizagem, os seguintes cenários foram considerados durante as simulações: $(\phi; \epsilon)$ = (0; 0), (0,02; 0), (0,04; 0), (0,08; 0), (0; 0,02), (0; 0,04), (0; 0,08), (0,03; 0,02) e (0,09; 0,08).

Tendo em vista que o agente P foi programado para ter três posicionamentos diferentes, dois fixos e um aleatório, buscou-se observar como o comportamento de resposta do agente IP se modifica a depender do modo de ação do agente P. Cada um dos cenários apresentados foram analisados para os três comportamentos de resposta do agente que não aprende, de forma a avaliar a capacidade do agente inteligente em mapear o comportamento do seu oponente. Os resultados das simulações foram divididos de acordo com cada algoritmo de aprendizagem, como indicam as seções 5.1, 5.2 e 5.3. De forma a obter maior confiabilidade nos resultados, vinte simulações foram realizadas para cada cenário e algoritmo de aprendizagem. Essas repetições permitiram obter valores médios de rodadas necessárias para que o agente inteligente apresente um comportamento estável ou com a menor variação possível, Tabelas 5, 6 e 7. Foi estabelecido um percentual de variação de 5%, como regra para a determinação do ponto em que os valores de probabilidade se estabilizam. As probabilidades médias atingidas durante o comportamento estável podem ser observadas nas Tabelas 8, 9 e 10.

5.1 **Roth-Erev RL**

O modelo de simulação determina como regra inicial que o agente IP adote a ação 0 até que os valores de probabilidade sejam capazes de determinar uma nova estratégia. Em todas as simulações realizadas, foi possível perceber dois comportamentos iniciais: o agente IP optou pela ação 0 nas cinco primeiras rodadas, ou nas quatro primeiras. Posteriormente, as probabilidades associadas às duas ações apresentaram variações até o ponto em que se estabilizaram. A Figura 7 apresenta os resultados da simulação, considerando o cenário em que os parâmetros ϕ e ϵ são mantidos nulos durante a execução dos três jogos clássicos.

Figura 7 – Resultados das simulações dos três jogos, considerando ϕ e ϵ iguais a 0, para as três possibilidades de resposta do agente P - Agente IP aprende por meio do RE



Fonte: A autora (2019)

As respectivas legendas adotadas para cada gráfico, indicam o jogo que foi testado bem

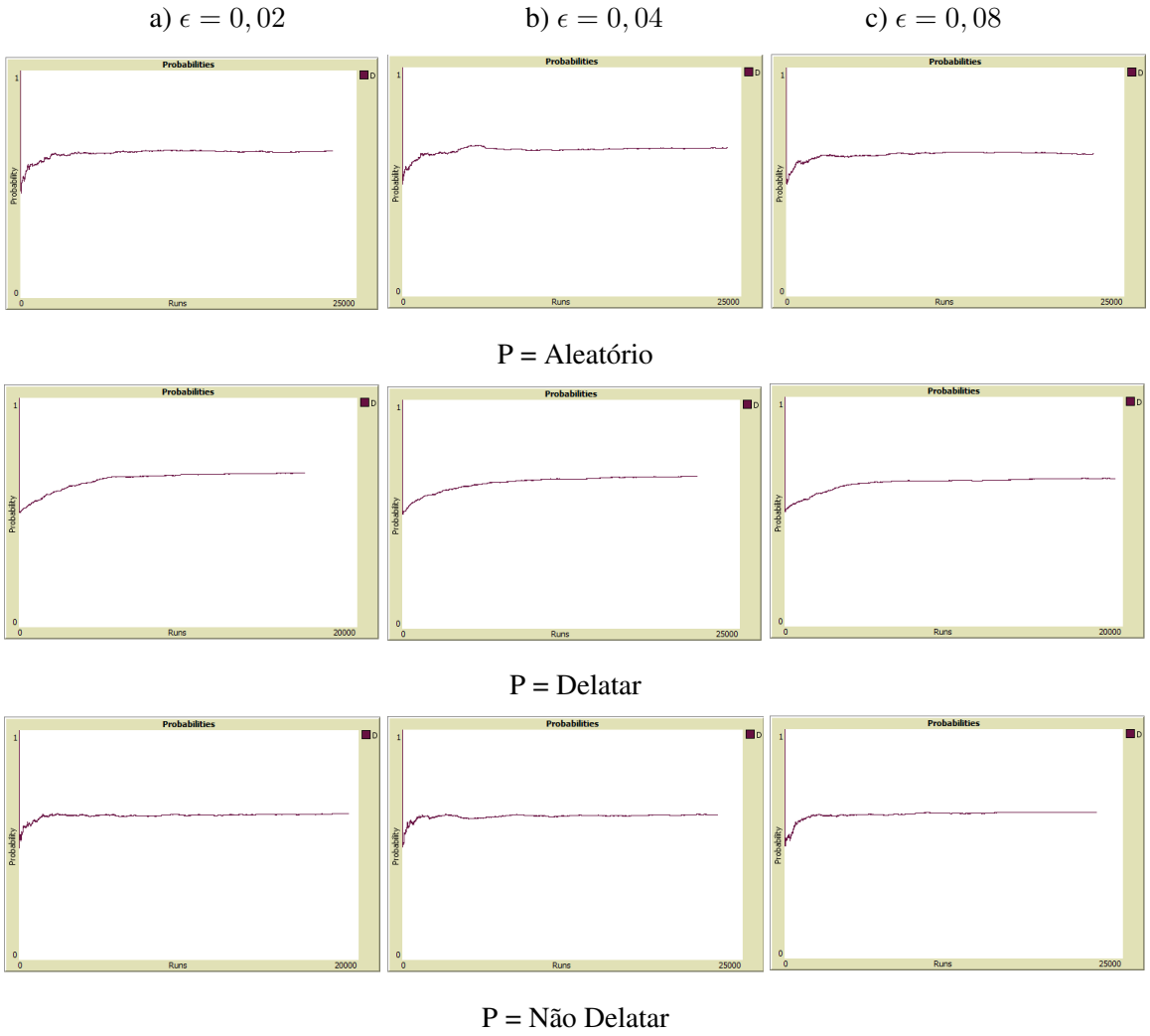
como a estratégia adotada pelo agente P. Já as curvas de probabilidade, indicam o comportamento do agente IP em resposta ao seu oponente. Ao incorporar o RE no comportamento de IP, todos os equilíbrios canônicos conhecidos, para os três jogos clássicos, foram obtidos. O agente inteligente foi capaz de se adaptar ao comportamento de resposta do agente que não aprende, adotando, dessa forma, as estratégias preferíveis a cada cenário. Foi possível identificar, a partir das repetidas simulações, que o agente inteligente estabiliza seu comportamento de resposta mais rapidamente no DP em comparação aos demais jogos, considerando o cenário em que os parâmetros de ϵ e ϕ são iguais a 0. No DP, o comportamento de IP se estabiliza, em média, a partir da 4692º rodada de simulação, quando P age aleatoriamente. Nos jogos BS e CG são necessárias 5362 e 6113 rodadas, respectivamente, para obter o mesmo resultado considerando o comportamento aleatório de P.

As Figuras 8, 9 e 10 apresentam os resultados provenientes da variação de ϵ empregando o algoritmo RE para os jogos DP, BS e CG, respectivamente. Seguindo o mesmo padrão de disposição dos gráficos apresentados anteriormente, nessas Figuras, cada linha apresenta os resultados do comportamento de IP em resposta a um tipo de comportamento de P.

O parâmetro de experimentação foi introduzido ao algoritmo RE, e mantido em suas modificações, com o objetivo de fazer com que o agente não se fixasse em uma primeira estratégia, podendo assim experimentar as demais estratégias disponíveis e a partir disso decidir a melhor opção. Quando ϵ foi variado no RE, os resultados esperados para o DP continuaram sendo alcançados. Como consta na Figura 8, nos cenários em que P age de forma aleatória ou opta por Não Delatar, o comportamento de IP se estabiliza mais rapidamente, quando $\epsilon = 0,02$, por exemplo, isso ocorre a partir das 4080º e 3297º rodadas, respectivamente. Por outro lado, quando P adota a estratégia Delatar, o comportamento de IP estabiliza a partir da 4871º rodada de simulação. Esse comportamento pode ser justificado pelos baixos valores de recompensa atribuídos ao cenário em que ambos os jogadores delatam, fazendo com que a atualização das probabilidades de escolha ocorra mais lentamente.

A Figura 9 apresenta os resultados obtidos para o jogo Batalha dos Sexos, ao contrário do DP este jogo não apresenta uma estratégia que seja dominante em qualquer que seja o cenário. Portanto, a depender do comportamento de resposta determinado para P, os resultados mudam em termos de estratégia preferível. Nessa situação de conflito, uma estratégia é preferível a cada um dos agentes. Enquanto o agente P foi modelado para preferir a estratégia C, o agente IP apresenta preferência pela estratégia F. Dessa forma, mesmo que a estratégia de preferência do agente IP seja F, à medida que o modelo constata as repetições de P em somente adotar C, o

Figura 8 – Resultados da simulação do Dilema dos Prisioneiros, variando ϵ e mantendo ϕ igual a 0 para as três possibilidades de comportamento do agente P - Agente IP aprende a partir do RE



Fonte: A autora (2019)

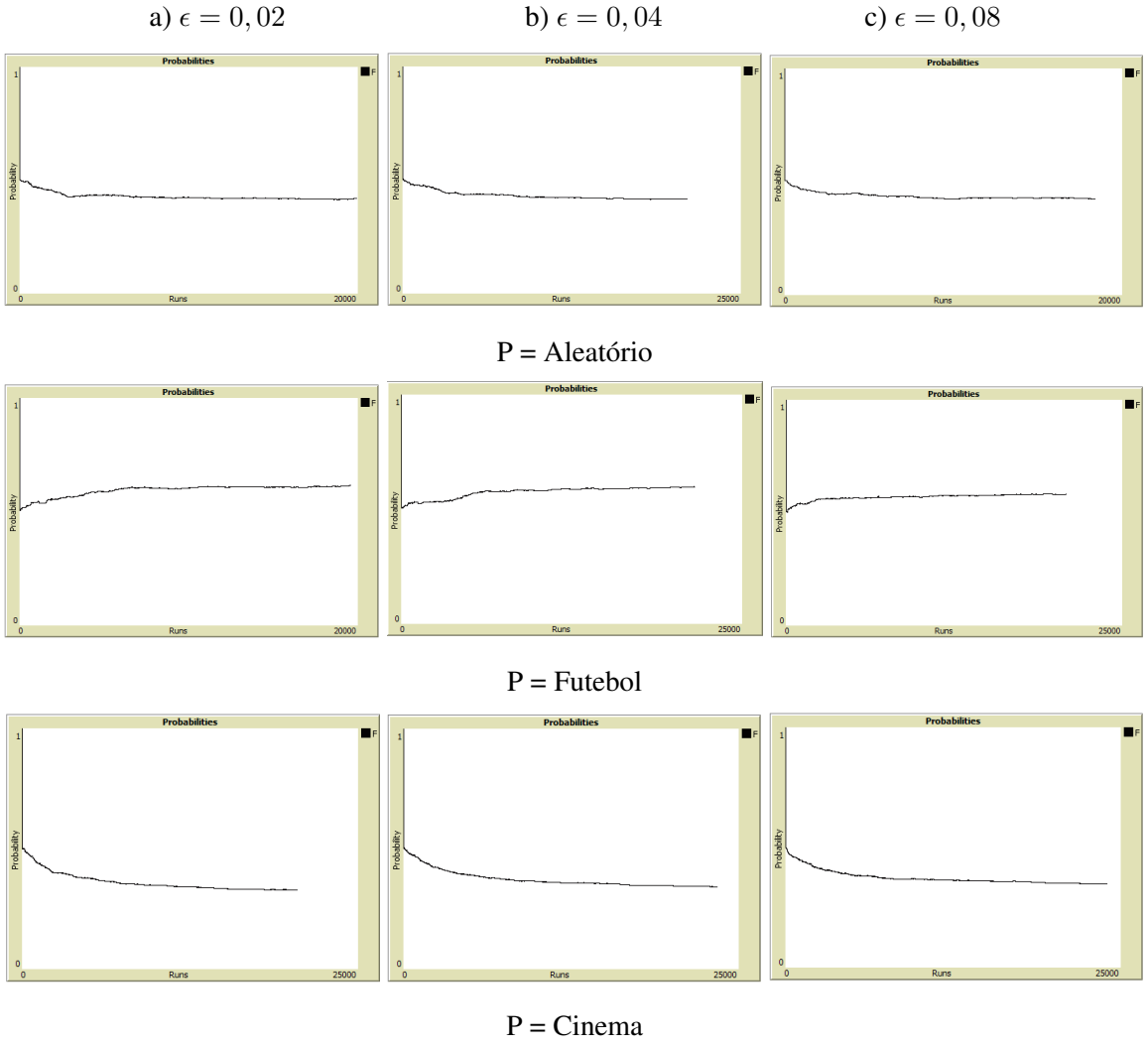
algoritmo altera o comportamento de resposta de IP, pois ao adotar a mesma estratégia do outro jogador, o agente IP recebe 2, enquanto que ao discordarem ambos recebem 0.

Os resultados permanecem confirmando os equilíbrios teóricos, no entanto à medida que o valor de ϵ aumenta, o comportamento de IP se estabiliza mais rapidamente. No cenário em que o agente P age aleatoriamente, quando os parâmetros são mantidos iguais a 0, são necessárias, em média, 5362 rodadas de simulação para que seja possível determinar o comportamento de IP como estável. Em contrapartida, quando $\epsilon = 0,08$, esse número é reduzido para 3459 rodadas de simulação.

Foi possível constatar, por meio das simulações repetidas, que quando o agente P adota somente a estratégia C, estratégia de sua preferência, o agente IP necessita de mais rodadas de

simulação para apresentar um comportamento estável, isso ocorre em torno da 7231º rodada, ao considerar $\epsilon = 0,02$.

Figura 9 – Resultados da simulação do Batalha dos Sexos, variando ϵ e mantendo ϕ igual a 0 para as três possibilidades de comportamento do agente P - Agente IP aprende a partir do RE



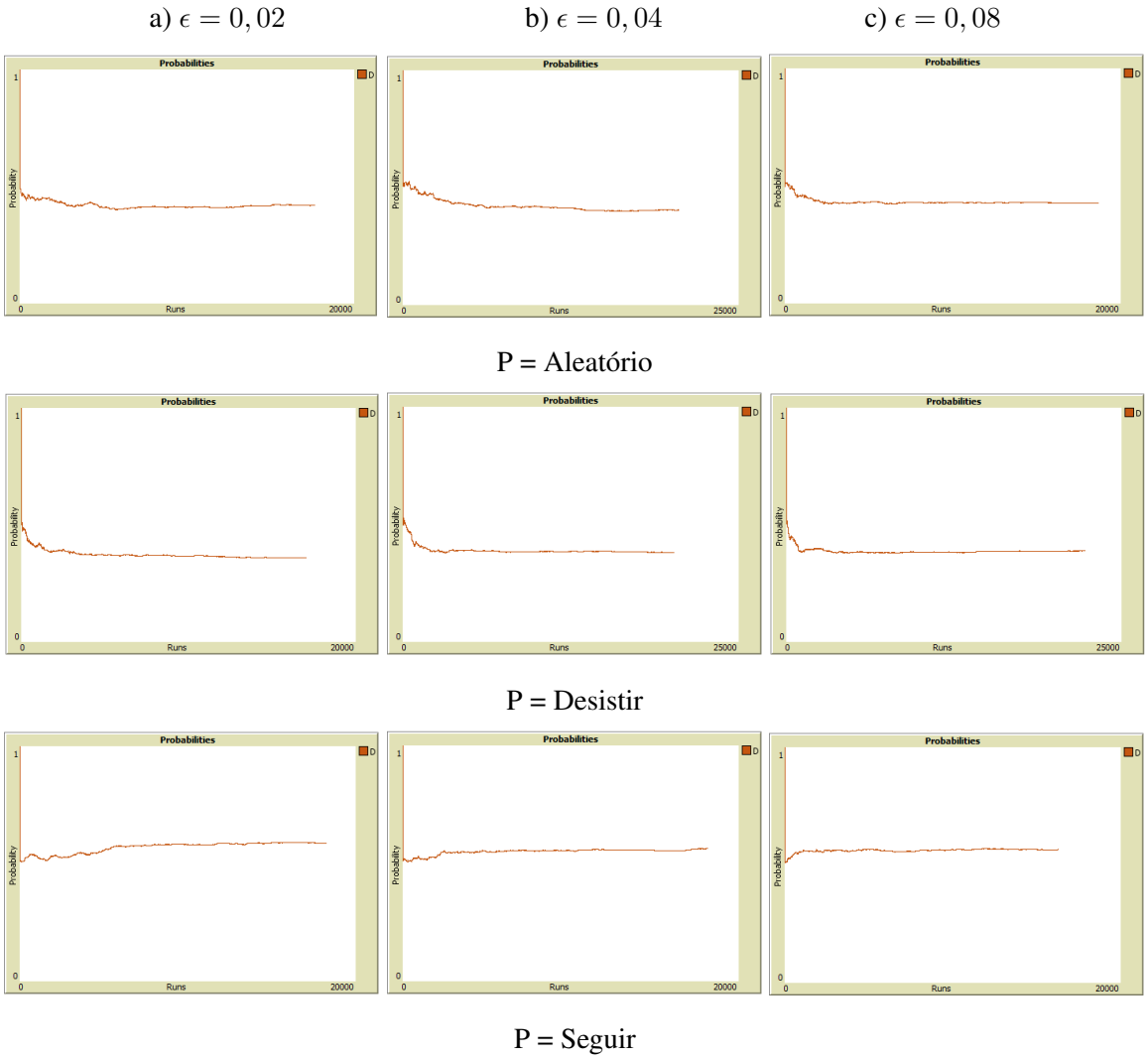
Fonte: A autora (2019)

A Figura 10 contém os resultados obtidos para o CG, em que, assim como no BS, o comportamento de IP se modifica de acordo com a estratégia preferível a cada cenário, tendo em vista que esse jogador tem capacidade de mapear o comportamento do seu oponente.

Assim como para os demais jogos, os equilíbrios teóricos do jogo *Chicken Game* foram confirmados ao simular o comportamento do agente que aprende por meio do algoritmo *Roth-Erev*, enquanto o parâmetro de experimentação sofre variação. De uma maneira geral, ao utilizar o RE para simular a inteligência de IP, nos três jogos, juntamente ao viés comportamental introduzido ao algoritmo por meio do parâmetro de experimentação, o número de rodadas

necessárias para que o agente apresentasse comportamento estável foi menor. É válido ressaltar que quanto maior o valor de ϵ, mais rápido o comportamento de IP se estabiliza.

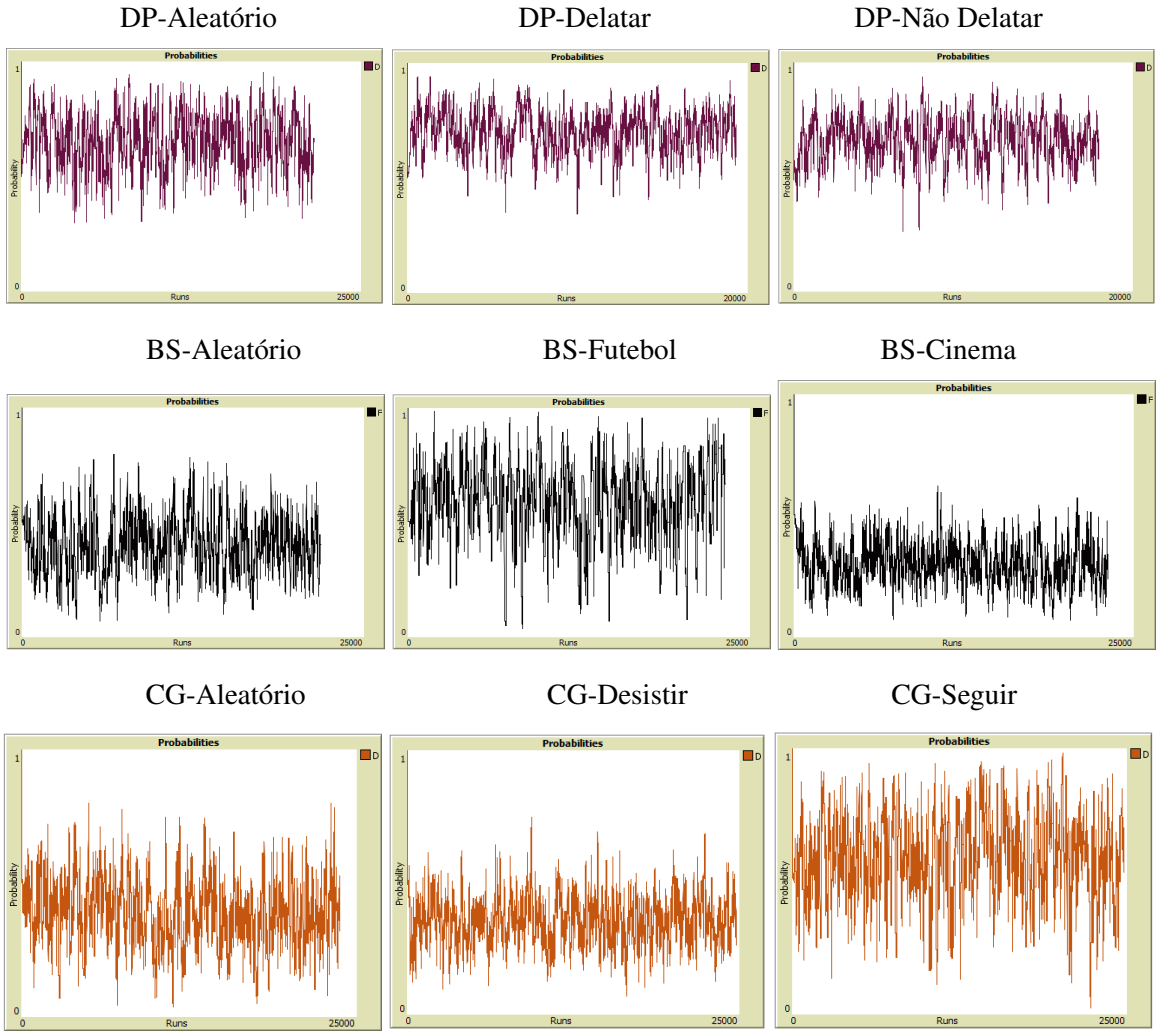
Figura 10 – Resultados da simulação do *Chicken Game*, variando ϵ e mantendo ϕ igual a 0 para as três possibilidades de comportamento do agente P - Agente IP aprende a partir do RE



Fonte: A autora (2019)

Posteriormente às análises das variações feitas em ϵ, os cenários com ϕ variando foram analisados. Os resultados das simulações constam na Figura 11. Comportamentos instáveis foram observados para as curvas de probabilidade correspondentes aos três jogos testados. O parâmetro de esquecimento faz com que o agente inteligente perca sua capacidade de posicionamento estratégico diante das situações de conflito. Percebe-se, em alguns casos, uma alternância contínua entre as estratégias disponíveis como, por exemplo, quando no jogo BS o agente P adota Futebol, em algumas rodadas a probabilidade de F apresenta alternância entre 0,1 e 0,9,

Figura 11 – Resultados da simulação dos três jogos, com $\phi = 0,02$ e $\epsilon = 0$, considerando as três possibilidades de comportamento do agente P - Agente IP aprende a partir do RE



Fonte: A autora (2019)

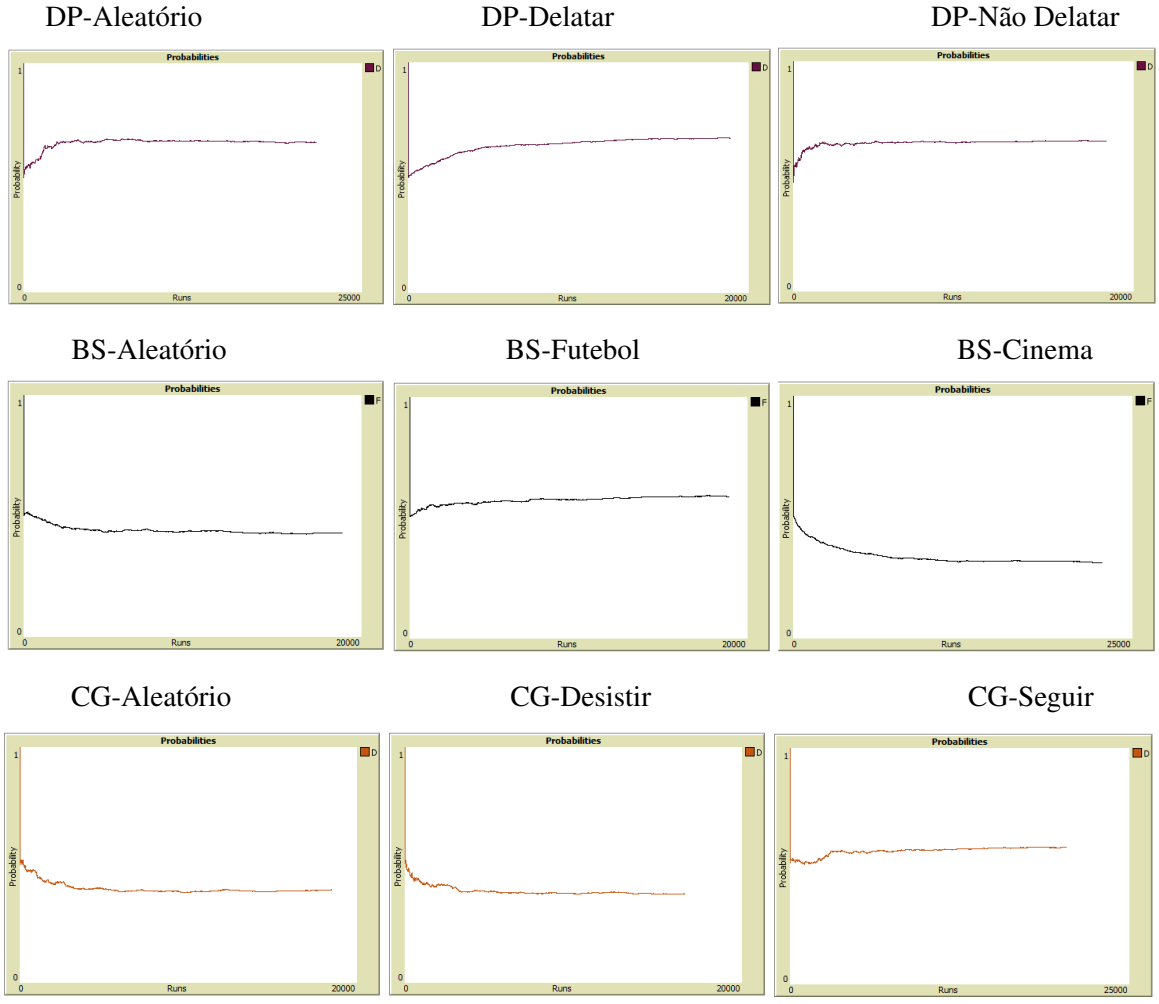
tendo em vista que valores de probabilidade superiores a 0,5 indicam que a estratégia Futebol está sendo adotada e valores inferiores indicam o contrário.

No entanto, em alguns cenários, apesar de o comportamento de IP ter apresentado instabilidade, os valores de probabilidade permaneceram indicando que uma única estratégia foi adotada na maior parte das rodadas de simulação. Esses resultados confirmam os equilíbrios teóricos e correspondem aos cenários: DP-Delatar, DP-Não Delatar, BS-Cinema e CG-Desistir. Os resultados correspondentes às demais variações feitas em ϕ, apresentaram comportamento similar e estão dispostos no Apêndice B.

5.2 **Modified Roth-Erev RL**

Ao submeter o algoritmo MRE aos mesmos testes realizados para o RE, foi possível constatar que para valores nulos de ϕ e ϵ, as curvas de probabilidade obtidas para os dois algoritmos, considerando os três jogos, se comportam de maneira semelhante. Os gráficos que demonstram o comportamento de resposta de IP podem ser observados na Figura 12.

Figura 12 – Resultados das simulações dos três jogos, considerando ϕ e ϵ iguais a 0, para as três possibilidades de resposta do agente P - Agente IP aprende por meio do MRE



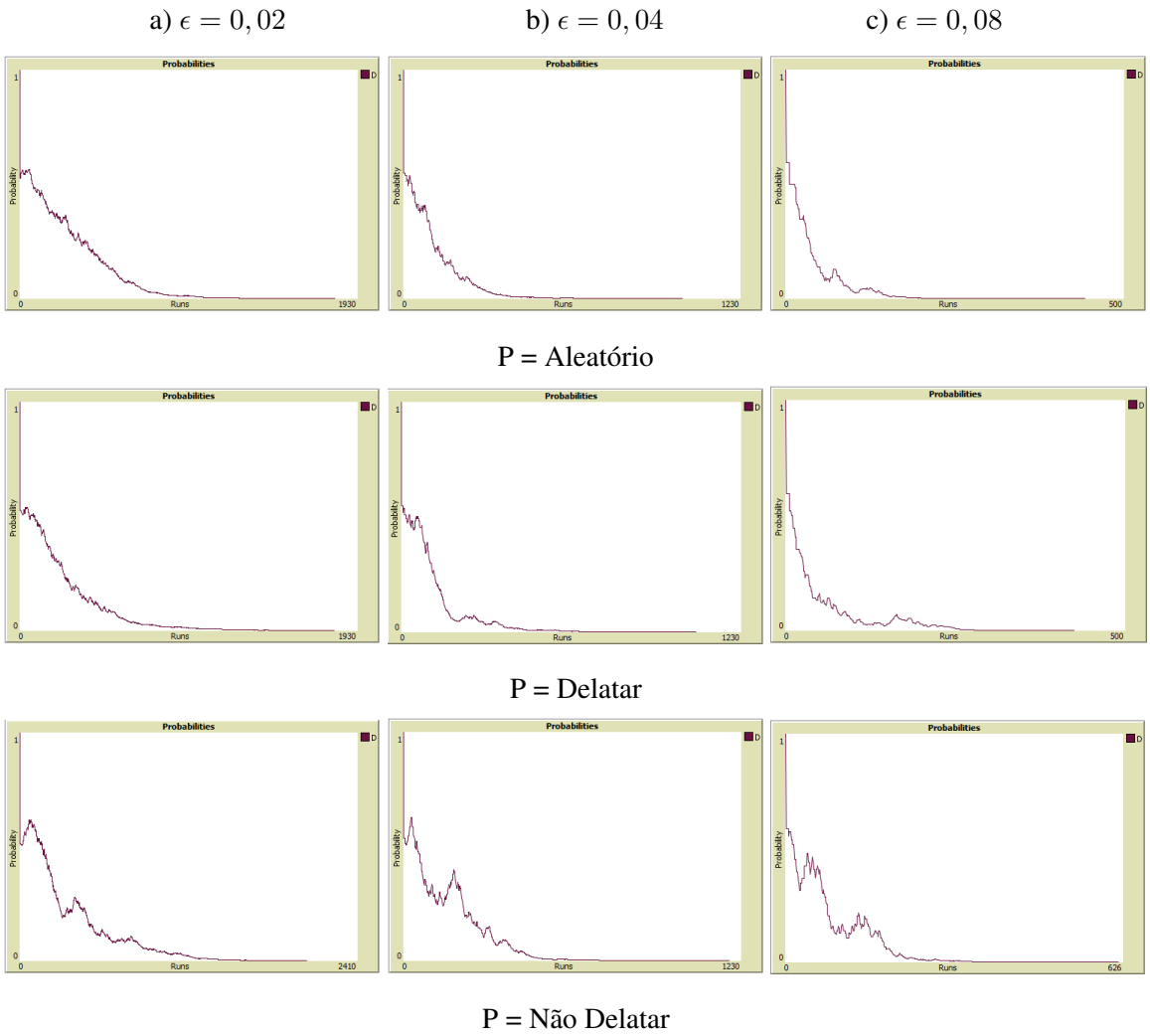
Fonte: A autora (2019)

No jogo DP, o agente inteligente optou por Delatar para qualquer que tenha sido o comportamento de seu oponente. Para os demais jogos, BS e CG, o agente IP modificava suas respostas à medida que mapeava o comportamento de P. Dessa forma, para essas condições de simulação, o algoritmo MRE foi capaz de atribuir ao agente inteligente a capacidade de se posicionar estrategicamente dentro das três situações de conflito. Quanto ao número de rodadas

necessárias para que o comportamento de IP fosse dito como estável, os três jogos apresentaram valores muito próximos. Tomando como exemplo o cenário em que P age aleatoriamente, IP apresenta comportamento estável a partir das 5824º, 6049º e 6371º rodadas de simulação, para os jogos DP, BS e CG, respectivamente.

Em contrapartida às afirmações feitas anteriormente, ao incorporar a variação de ϵ, no algoritmo de aprendizagem MRE, o comportamento das curvas de probabilidade difere completamente do comportamento previsto para o primeiro algoritmo, assim como mostram as Figuras 13, 14 e 15.

Figura 13 – Resultados da simulação do Dilema dos Prisioneiros, variando ϵ e mantendo ϕ igual a 0 para as três possibilidades de comportamento do agente P - Agente IP aprende a partir do MRE



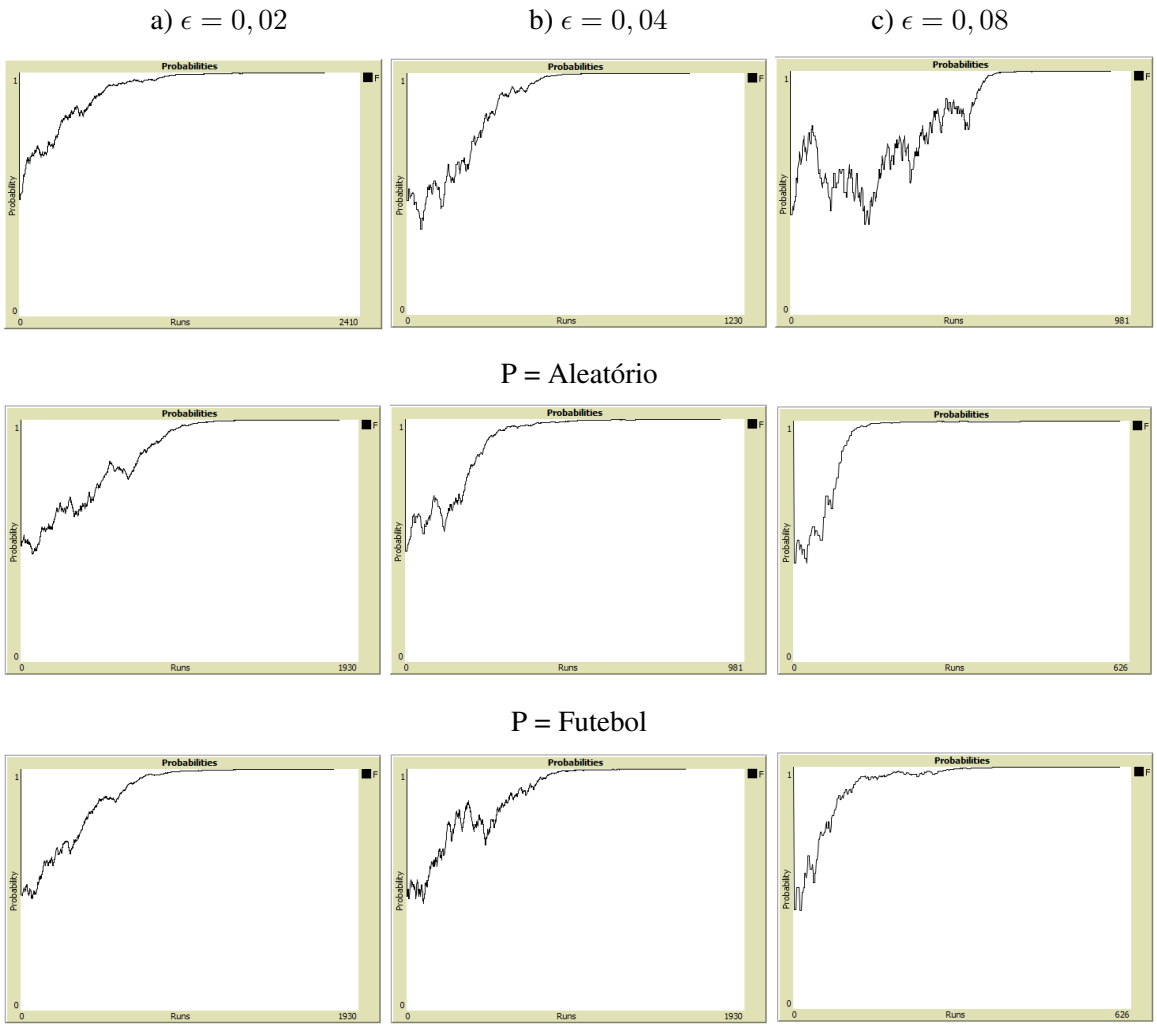
Fonte: A autora (2019)

Conforme demonstrado na Figura 13, duas principais alterações foram percebidas no comportamento de IP, quando ϵ é considerado. Primeiro, os valores de probabilidade, que

anteriormente se estabilizavam entre 0 e 1, passaram a atingir os valores 0 e 1. A segunda alteração está no comportamento de resposta de IP, que passou a adotar a estratégia ND em resposta a todo e qualquer comportamento de P, o levando a obter *payoffs* mínimos em comparação ao seu oponente. Esse comportamento fez com que o equilíbrio teórico do jogo não fosse alcançado durante as simulações. Isso se deu por meio da influência do parâmetro de experimentação no algoritmo de aprendizagem.

O mesmo comportamento foi observado para os demais jogos. Na Figura 14, se encontram os resultados obtidos para o jogo Batalha dos Sexos.

Figura 14 – Resultados da simulação do Batalha dos Sexos variando ϵ e mantendo ϕ igual a 0 para as três possibilidades de comportamento do agente P - Agente IP aprende a partir do MRE

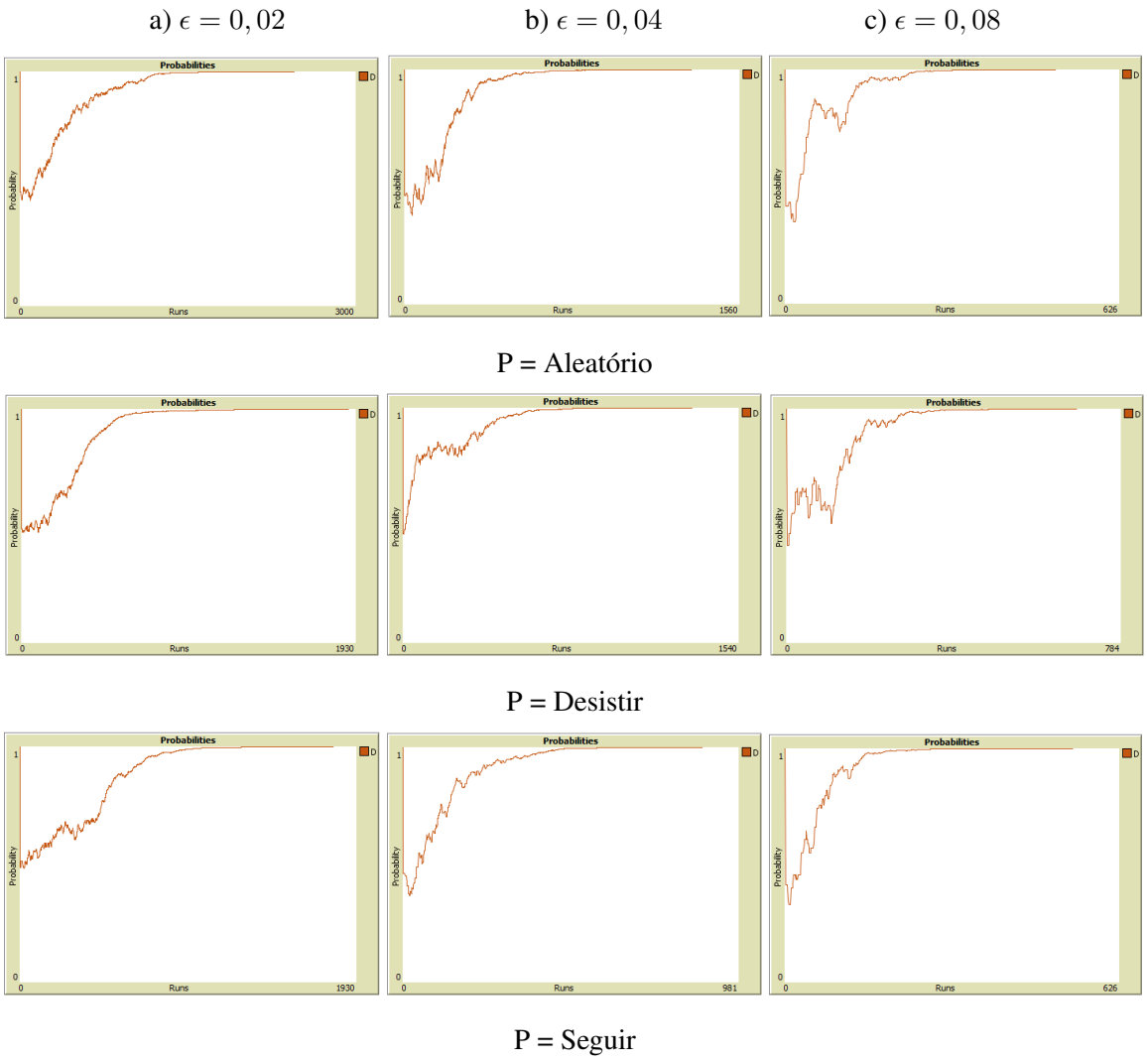


Fonte: A autora (2019)

Ao empregar valores para ϵ durante a simulação do BS, os resultados obtidos confirmam

apenas um equilíbrio de Nash, correspondente a combinação de estratégias F_1F_2. Para os demais cenários os resultados foram contraditórios ao resultado canônico do jogo, o mesmo ocorreu durante a simulação do jogo CG, Figura 15. Quanto maior o valor assumido por ϵ, menor o número de rodadas necessárias para que a probabilidade de escolha de F seja igual a 1.

Figura 15 – Resultados da simulação do *Chicken Game*, variando ϵ e mantendo ϕ igual a 0 para as três possibilidades de comportamento do agente P - Agente IP aprende a partir do MRE



Fonte: A autora (2019)

Da mesma forma como ocorreu para os demais jogos, no CG, o agente IP passou a adotar uma única estratégia em resposta a todo e qualquer comportamento de P. Ao somente Desistir, IP deixa de receber *payoffs* máximos, como no cenário em que P também desiste.

De uma maneira geral, quando incorporado ao MRE, o parâmetro de experimentação, atribui falha ao processo de aprendizagem, tornando o agente incapaz de mapear o compor-

tamento do seu oponente, bem como de se posicionar estrategicamente diante dos diferentes cenários simulados. As justificativas para esses resultados podem estar relacionada às modificações feitas por Nicolaisen, Petrov e Tesfatsion (2001), no algoritmo original e até mesmo às condições que delimitam um jogo de matriz 2 x 2. Uma vez que, o número de ações/estratégias, M, disponíveis aos agentes é igual a 2, levando o denominador da Equação (3), $M - 1$, a ser igual a 1. Isso faz com que a propensão de escolha da estratégia que não foi adotada seja sempre maior, tendo em vista que $j = k$ implica em $R(j)(1 - \epsilon)$, enquanto $j \neq k$ implica em $q_{nj}(t)\epsilon$, e mesmo que $(1 - \epsilon) > \epsilon$, o valor da propensão do período corrente, $q_{nj}(t)$, será sempre maior que a recompensa $R(j)$, resultando em um processo de aprendizagem falho.

O parâmetro de esquecimento também foi variado, com o objetivo de identificar sua influência no processo de aprendizagem, como um todo. Os resultados das simulações mostraram que o algoritmo MRE é sensível, também, às variações de ϕ. A Figura 16 apresenta os resultados obtidos para os três jogos, empregando $\phi = 0,02$.

O parâmetro de esquecimento exerce a função de fazer com que as ações tomadas em um passado recente tenham maior peso e, portanto, maiores chances de serem repetidas, por se tratar de um processo probabilístico. Assim como mostra a Figura 16, o algoritmo MRE atribui um comportamento instável ao agente inteligente, quando ϕ recebe valores maiores que 0. Apesar de, em cenários como o representado em CG-Seguir, o agente variar continuamente o seu posicionamento diante do conflito, ϕ, ao contrário de ϵ, não faz com que o agente perca totalmente a sua capacidade de se posicionar estrategicamente.

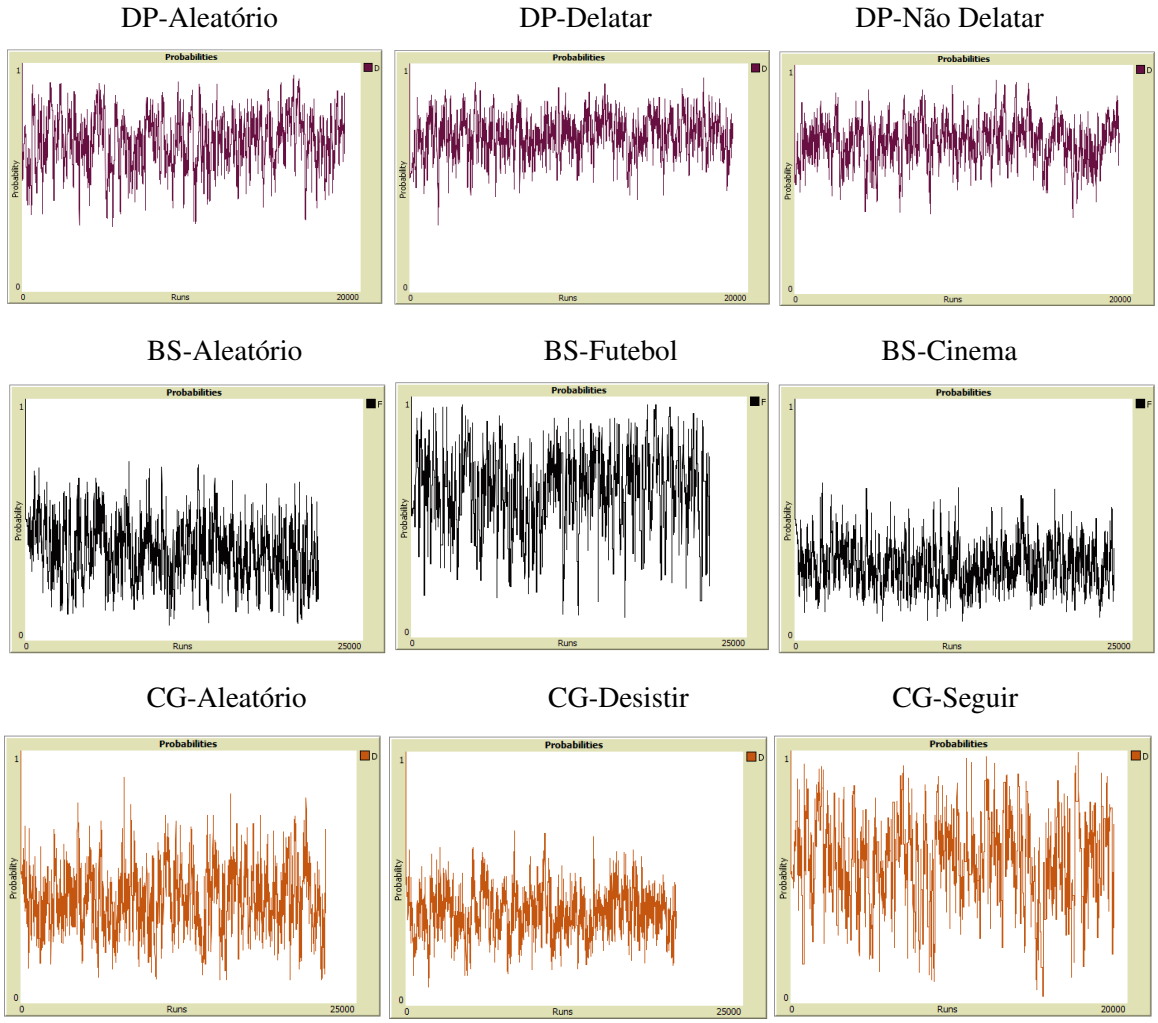
Em cenários como os representados em: DP-Delatar, DP-Não Delatar, BS-Cinema e CG-Desistir, os valores de probabilidade, associados às escolhas inerentes a cada conflito, variam, no entanto permanecem indicando o posicionamento adequado na maioria das rodadas de simulação, confirmando os equilíbrios teóricos dos três jogos.

Durante as simulações em que ambos os parâmetros sofreram variação, foi identificada uma alternativa de correção às falhas decorrentes das variações feitas somente em ϵ. O parâmetro de esquecimento, quando variado juntamente ao parâmetros de experimentação, atribui determinado equilíbrio às reações causadas por ϵ, permitindo que o agente volte a se posicionar de maneira estratégia e passe a receber recompensas satisfatórias.

A Figura 17, apresenta os resultados obtidos por meio da simulação em que $\phi = 0,03$ e $\epsilon = 0,02$. O comportamento das curvas de probabilidade, confirmam os equilíbrios teóricos dos três jogos.

Foi identificado, durante as repetidas simulações, que, para que esse equilíbrio seja

Figura 16 – Resultados das simulações dos três jogos clássicos, variando ϕ e mantendo ϵ igual a 0 para as três possibilidades de comportamento do agente P - Agente IP aprende a partir do MRE



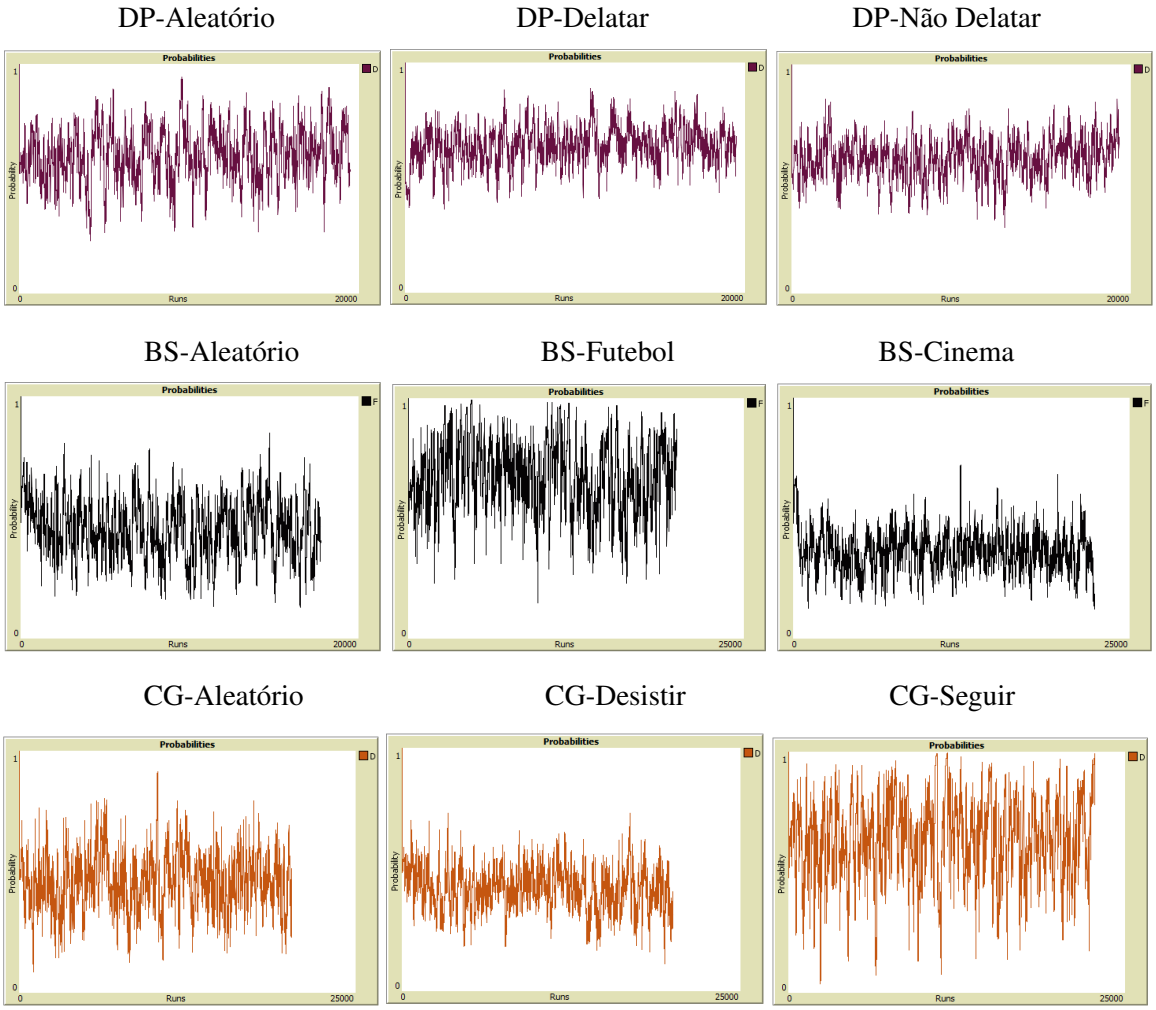
Fonte: A autora (2019)

alcançado, os valores atribuídos a ϕ devem ser maiores ou iguais aos valores atribuídos a ϵ. Ao assumirem valores iguais, o aprendizado ocorre mais lentamente.

Após submeter o algoritmo MRE às variações de ϕ e ϵ, separadamente e de forma conjunta, simulando todos os cenários possíveis dentro dos três jogos clássicos, foi possível concluir que esse algoritmo de aprendizagem apresenta melhor desempenho quando os vieses comportamentais, introduzidos por meio dos parâmetros de esquecimento e experimentação, não são considerados durante as simulações. Dessa forma agente inteligente pode se posicionar estrategicamente, a partir do mapeamento de ações do seu oponente, e obter melhores recompensas levando em consideração o total de recompensas acumuladas durante as rodadas de simulação.

As simulações empregando valores maiores para ambos os parâmetros, resultaram em curvas de probabilidade que são semelhantes às observadas na Figura 17. Os resultados das

Figura 17 – Resultados das simulações dos três jogos clássicos, com $\epsilon = 0,02$ e $\phi = 0,03$, para as três possibilidades de comportamento do agente P - Agente IP aprende a partir do MRE



Fonte: A autora (2019)

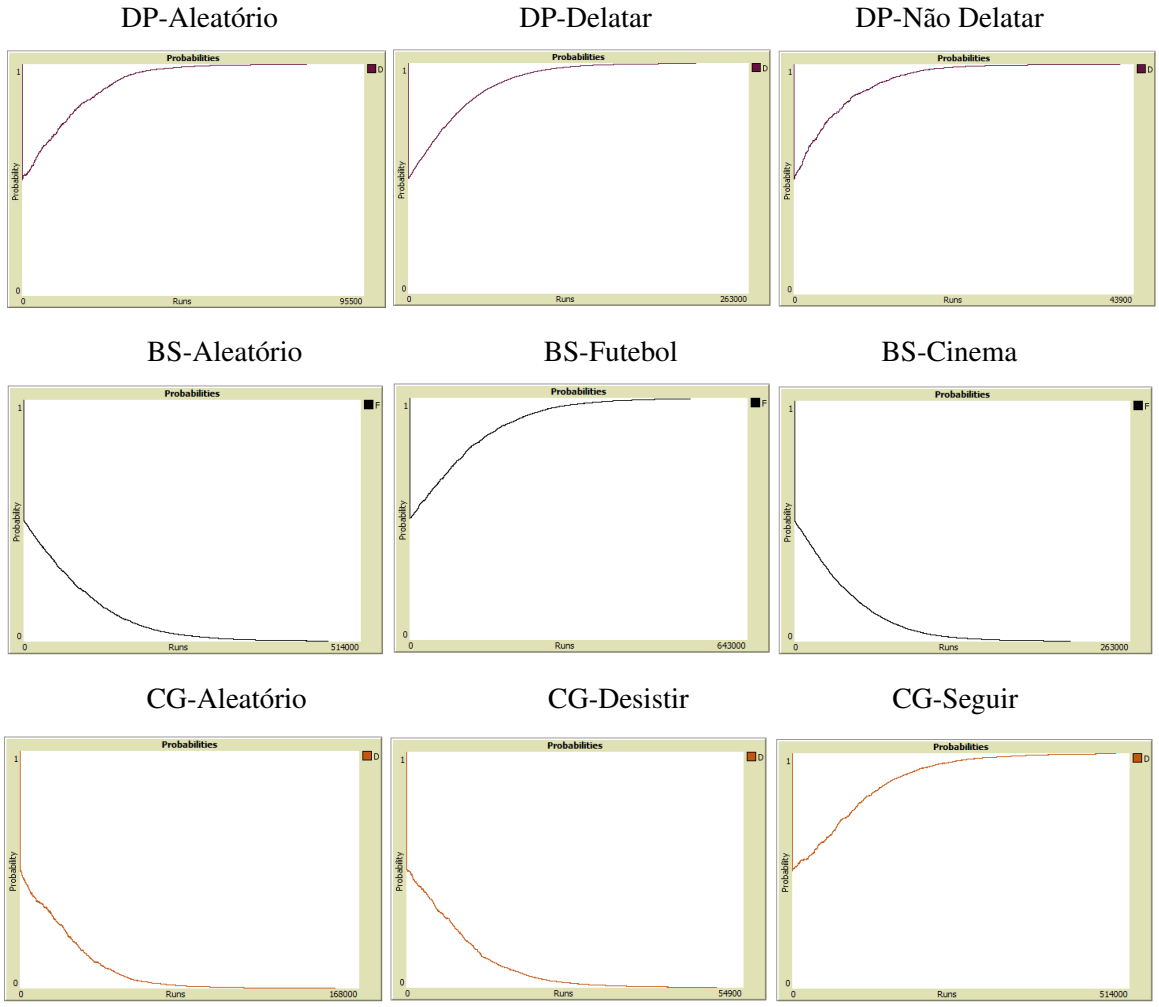
demais simulações podem ser encontrados no Apêndice B.

5.3 **Variant Roth-Erev RL**

O algoritmo VRE, mantém as alterações feitas por Nicolaisen, Petrov e Tesfatsion (2001), na versão original, e acrescenta uma nova forma de calcular as probabilidades de escolha. Devido a alteração proposta por Sun e Tesfatsion (2007), novos padrões de comportamento puderam ser observados durante as simulações, esse algoritmo faz com que as probabilidades sempre apresentem convergência para um dos limites. A Figura 18, apresenta os resultados obtidos durante a simulação dos três jogos, enquanto os parâmetros de experimentação e esquecimento foram considerados iguais a 0.

O comportamento das curvas de probabilidade, obtido para os respectivos cenários

Figura 18 – Resultados das simulações dos três jogos clássicos, com $\phi = \epsilon = 0$, considerando as três possibilidades de comportamento do agente P - Agente IP aprende a partir do VRE



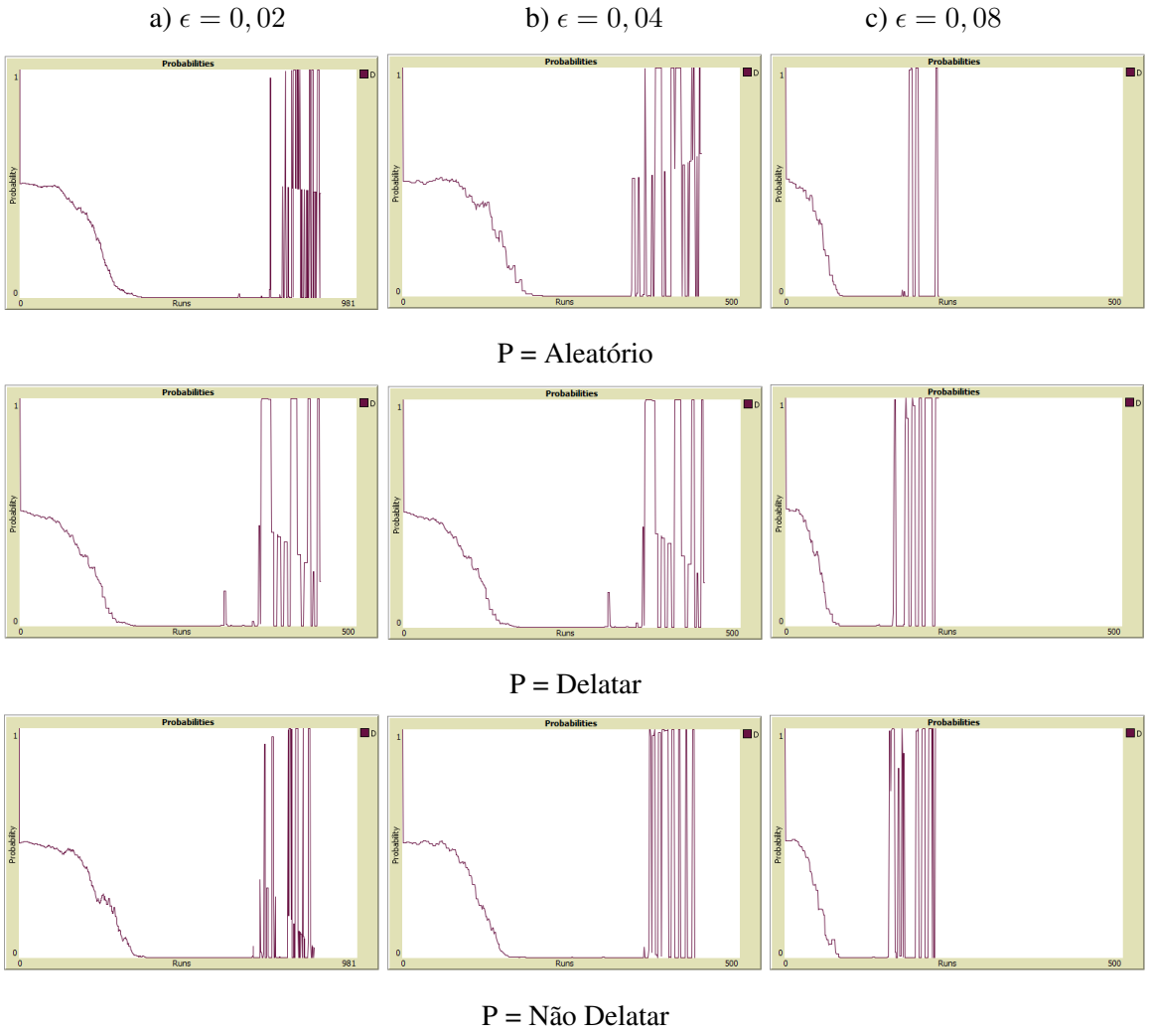
Fonte: A autora (2019)

pertencentes a cada um dos três jogos clássicos, confirmam os equilíbrios teóricos. Para que houvesse convergência para os valores 0 e 1, foram necessárias, na maioria dos cenários, um valor superior a cem mil rodadas de simulação, os valores específicos podem ser consultados nas Tabelas 5, 6 e 7. Apesar do elevado número de rodadas necessárias para alcançar tal resultado, logo nas primeiras 50 rodadas o agente apresentou capacidade de posicionamento estratégico, adotando as estratégias preferíveis de acordo com o posicionamento do seu oponente.

Uma vez identificado o padrão de comportamento das curvas de probabilidade sob as condições iniciais de simulação, ou seja, $\epsilon = \phi = 0$, o parâmetro de experimentação foi incorporado ao algoritmo com o objetivo de identificar as implicações deste no processo de aprendizagem, os resultados dessa simulação se encontram na Figura 19.

Ao empregar o algoritmo de aprendizagem VRE, na simulação do DP, com $\epsilon > 0$,

Figura 19 – Resultados da simulação do Dilema dos Prisioneiros, variando ϵ e mantendo ϕ igual a 0 para as três possibilidades de comportamento do agente P - Agente IP aprende a partir do VRE



Fonte: A autora (2019)

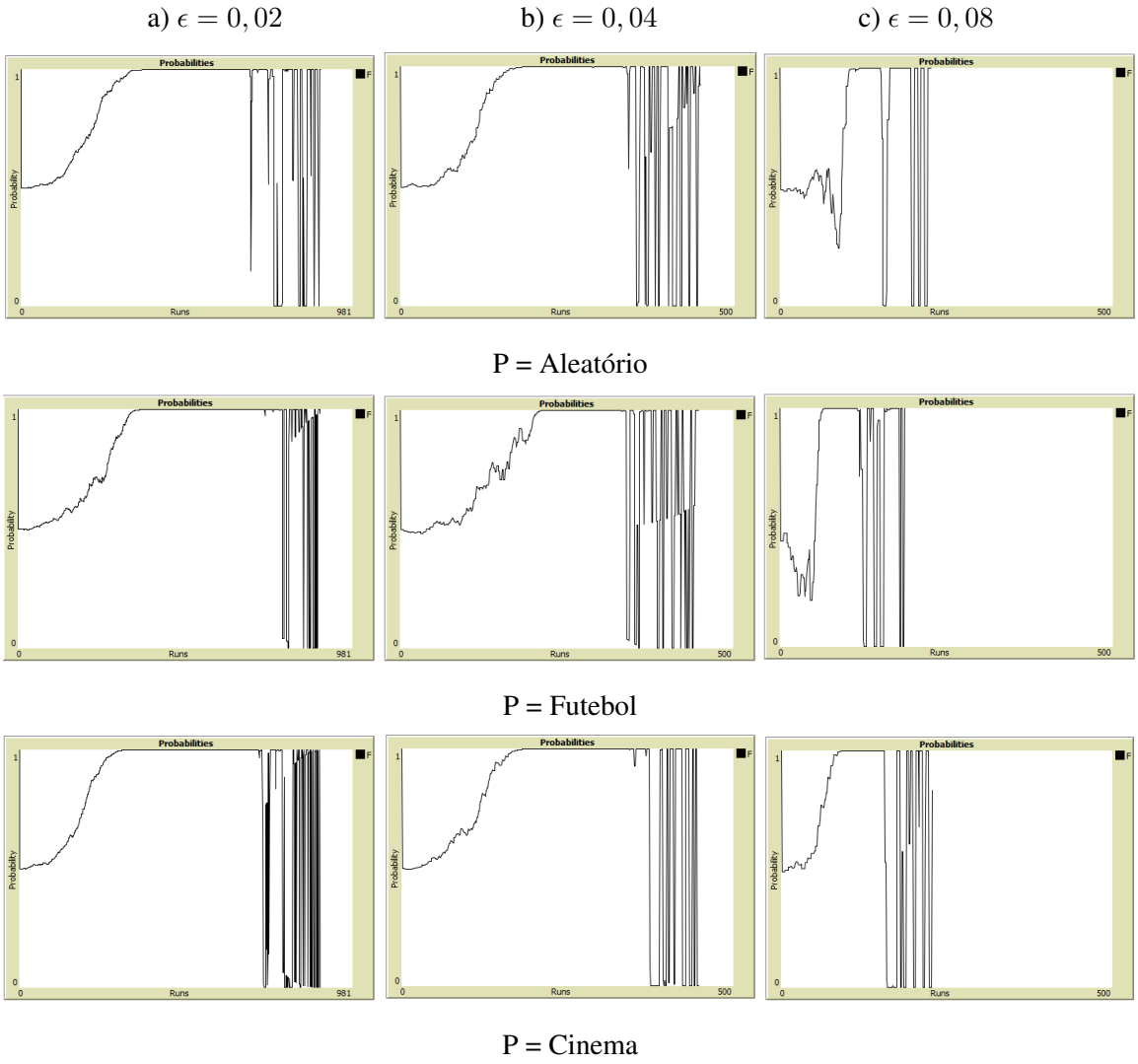
falhas ocorridas na aprendizagem puderam ser observadas. O parâmetro de experimentação acelerou a convergência da curva de probabilidade para os limites, no entanto fez com que o agente inteligente perdesse a capacidade de posicionamento estratégico, adotando estratégias que não refletem os *payoffs* máximos que podem ser adquiridos a cada cenário.

O número de rodadas necessárias para que as curvas de probabilidade apresentassem convergência, sofreu uma significativa redução de mais de cem mil para menos de mil rodadas, aproximadamente.

O mesmo comportamento falho, foi observado durante as simulações dos jogos BS e CG, como mostram as Figuras 20 e 21, respectivamente.

Sob o efeito do parâmetro de experimentação, o agente inteligente passou a adotar

Figura 20 – Resultados da simulação do Batalha dos Sexos, variando ϵ e mantendo ϕ igual a 0 para as três possibilidades de comportamento do agente P - Agente IP aprende a partir do VRE



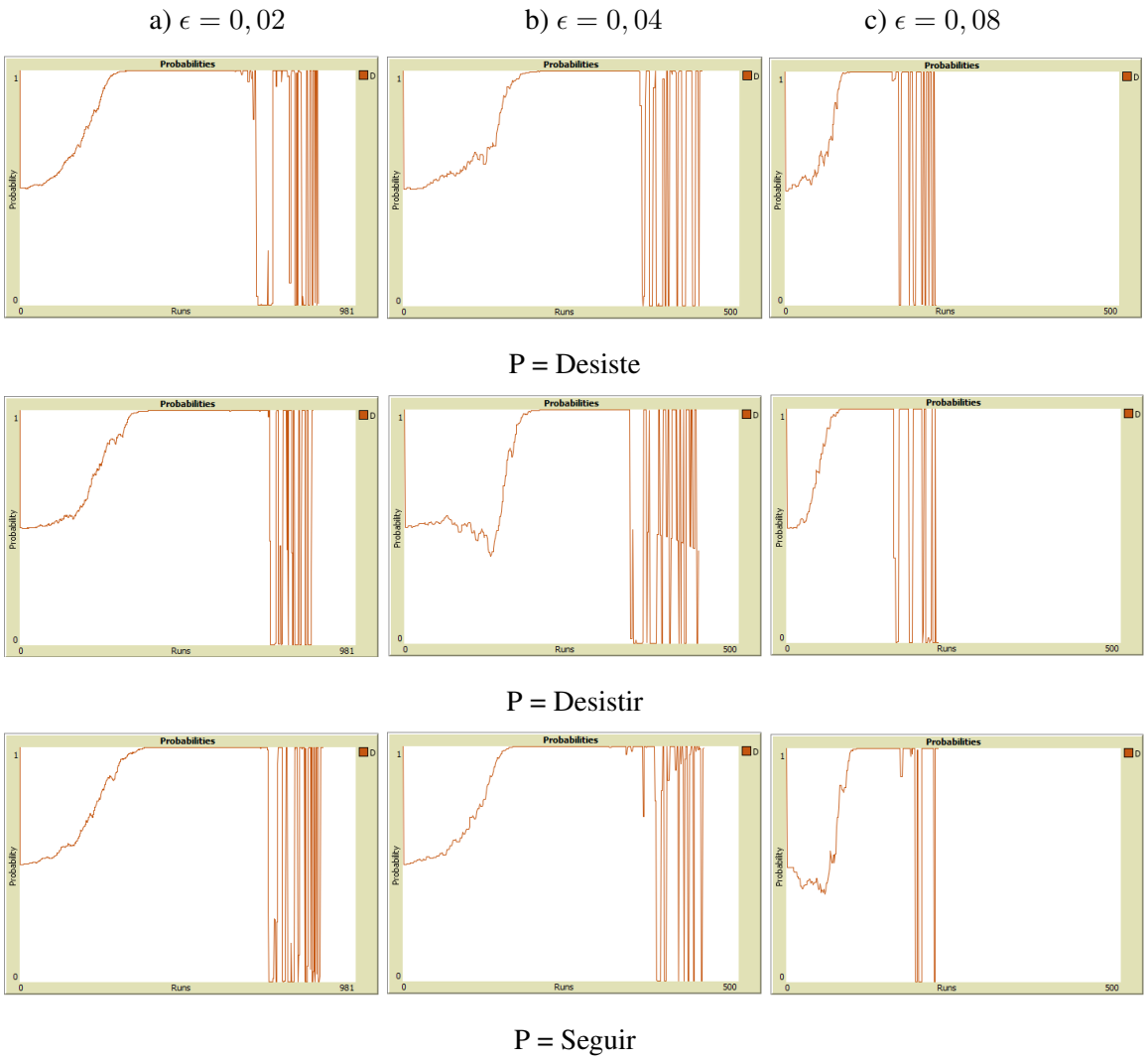
Fonte: A autora (2019)

somente a estratégia Futebol, tendo em vista que o jogo foi desenhado para que essa fosse a sua estratégia de preferência no decorrer do conflito. À medida que ϵ aumenta, mais rápido a probabilidade de escolha associada a F se torna 1.

A probabilidade de F se estabiliza em 1 durante algumas rodadas e, posteriormente, inicia uma série de picos que fazem com que a probabilidade alterne entre os limites 0 e 1. Esse comportamento foi identificado para o DP e, também, para o CG, como mostra a Figura 21.

Como mostra a Figura 21, quanto maior o valor assumido por ϵ, mais fortemente o agente IP adota uma estratégia que não lhe proporciona o ganho máximo. No CG, o agente inteligente passou a adotar somente a estratégia Desistir, independentemente do posicionamento

Figura 21 – Resultados da simulação do Chicken Game, variando ϵ e mantendo ϕ igual a 0 para as três possibilidades de comportamento do agente P - Agente IP aprende a partir do VRE



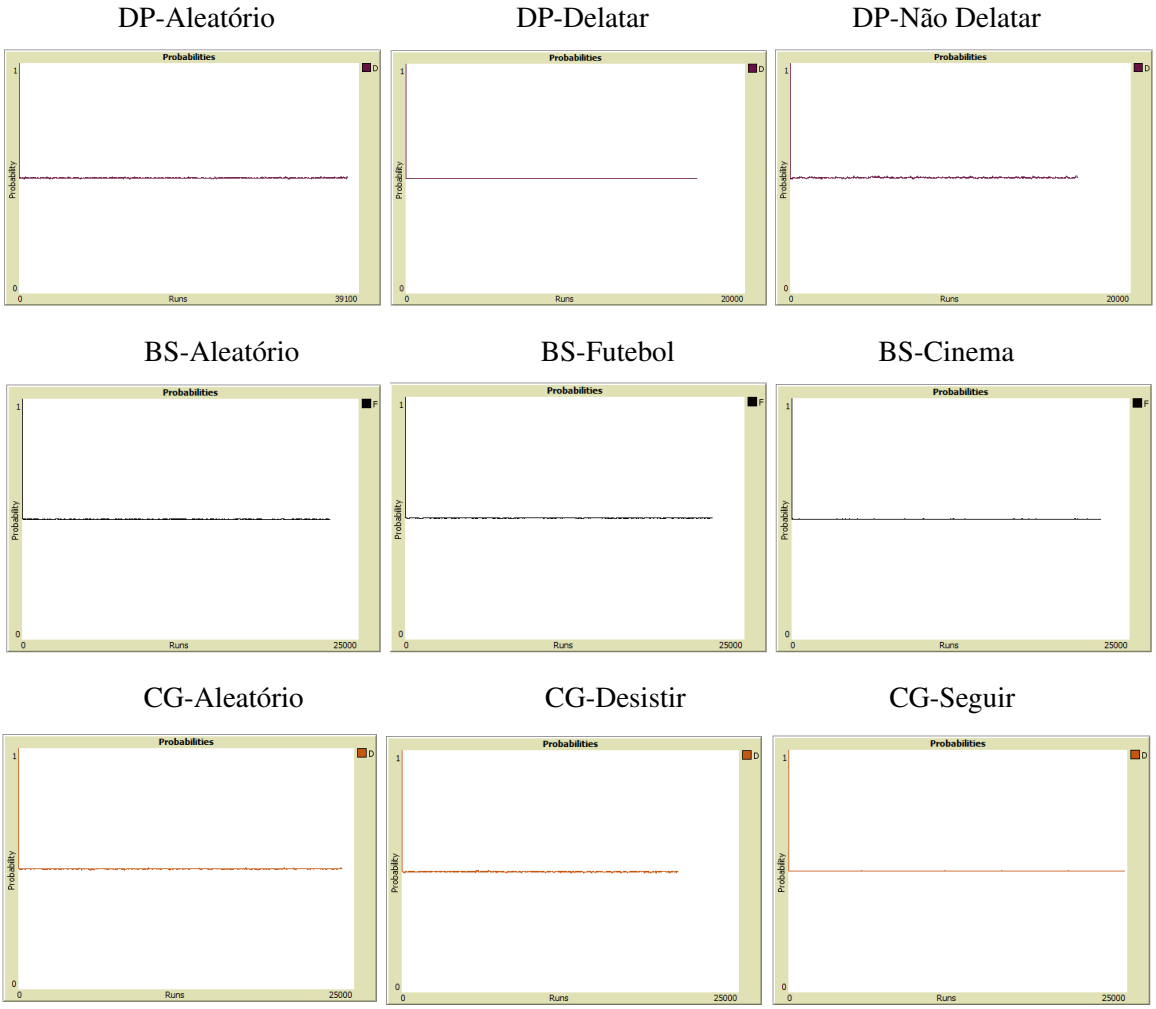
Fonte: A autora (2019)

de P.

As simulações realizadas empregando a variação de ϕ, mostraram que esse parâmetro faz com que não haja posicionamento por parte de IP, tendo em vista que as probabilidades de escolha se estabilizam exatamente em 0,5 durante infinitas rodadas de simulação. Os resultados das simulações que incorporam ϕ ao desempenho do algoritmo VRE, para os três jogos, podem ser observados na Figura 22.

As simulações em que ϕ e ϵ sofreram variação ao mesmo tempo, não indicaram um potencial de correção para a falha identificada na aprendizagem, por meio do VRE. Nos cenários em que $\phi \geq \epsilon$, os resultados confirmaram o padrão de comportamento exibido na Figura 22, já

Figura 22 – Resultados das simulações dos três jogos, com $\phi = 0,02$ e $\epsilon = 0$, considerando as três possibilidades de comportamento do agente P - Agente IP aprende a partir do VRE



Fonte: A autora (2019)

nos cenários em que $\phi < \epsilon$, os resultados confirmaram o padrão de comportamento exibido nas Figuras 19, 20 e 21, ou seja, a combinação de ambos os parâmetros não torna o agente inteligente capaz de se posicionar estrategicamente.

5.4 **Considerações finais do capítulo**

Neste estudo, foi possível constatar que logo nas primeiras 50 rodadas, os três algoritmos atribuem maiores valores de probabilidade às estratégias preferíveis, dado o cenário simulado. No entanto, o agente não apresenta um comportamento estável de forma tão rápida. As informações presentes nas Tabelas 5, 6 e 7, demonstram que são necessárias mais de quatro mil rodadas de simulação, no mínimo, para que o comportamento do agente apresente um percentual de variação menor ou igual a 5%.

Tabela 5 – Valor médio de rodadas necessárias para que IP tenha comportamento estável - Dilema dos Prisioneiros

Parâmetros	RE	MRE	VRE	RE	MRE	VRE	RE	MRE	VRE
	Aleatório			Delatar			Não Delatar		
$\epsilon = \phi = 0$	4692	5824	48500	4933	6905	126600	3414	5149	30430
$\epsilon = 0,02; \phi = 0$	4080	1436	307	4871	1353	375	3297	1475	368
$\epsilon = 0,04; \phi = 0$	3453	778	234	4365	677	171	3055	734	183
$\epsilon = 0,08; \phi = 0$	3435	384	85	4165	348	123	2448	436	150

Fonte: A autora (2019)

Somente foi possível obter resultados para os cenários em que ambos os parâmetros são nulos e empregando variações em ϵ, isso se deve ao fato de que o parâmetro de esquecimento introduz instabilidade no comportamento de IP, quando os algoritmos RE e MRE estão em uso. No caso particular do VRE, o parâmetro de esquecimento resulta em valores de probabilidade que não se atualizam e não refletem posicionamento algum por parte de IP.

Tabela 6 – Valor médio de rodadas necessárias para que IP tenha comportamento estável - Batalha dos Sexos

Parâmetros	RE	MRE	VRE	RE	MRE	VRE	RE	MRE	VRE
	Aleatório			Futebol			Cinema		
$\epsilon = \phi = 0$	5362	6049	355000	6108	6490	470000	7792	6806	161300
$\epsilon = 0,02; \phi = 0$	4276	1307	363	5825	1326	278	7231	1401	310
$\epsilon = 0,04; \phi = 0$	3967	719	186	5172	730	218	7245	707	185
$\epsilon = 0,08; \phi = 0$	3459	394	121	4794	397	112	7415	417	93

Fonte: A autora (2019)

De maneira geral, o algoritmo RE produz um comportamento estável mais rapidamente, tendo em vista que o número de rodadas necessárias para isso, diminuem ao passo em que ϵ aumenta. O mesmo efeito foi observado para os demais algoritmos.

Tabela 7 – Valor médio de rodadas necessárias para que IP tenha comportamento estável - *Chicken-Game*

Parâmetros	RE	MRE	VRE	RE	MRE	VRE	RE	MRE	VRE
	Aleatório			Seguir			Desistir		
$\epsilon = \phi = 0$	6113	6371	110600	5978	6317	391000	5702	5839	36300
$\epsilon = 0,02; \phi = 0$	5898	1394	311	5445	1358	368	5079	1389	377
$\epsilon = 0,04; \phi = 0$	5606	751	204	5294	742	166	4835	774	206
$\epsilon = 0,08; \phi = 0$	4896	418	98	4697	391	108	4646	372	85

Fonte: A autora (2019)

O algoritmo VRE, resulta em valores de probabilidade que convergem para 0 ou 1, a depender de cada cenário. Além de ser um processo demasiado lento, tendo em vista os resultados das Tabelas 5, 6 e 7, esse não é um comportamento dito como estratégico, uma vez que ao atingirem 0 ou 1, os valores de probabilidade não se alteram. Nos cenários em que P apresenta

comportamento de resposta fixo, o posicionamento de IP dado o algoritmo VRE, resulta em ganhos acumulados mais satisfatórios em relação aos demais algoritmos. No entanto, levando em consideração os resultados que podem ser alcançados quando P age aleatoriamente, não é vantajoso para IP, apresentar comportamento fixo e sem atualização.

As probabilidades médias permitem avaliar a percepção do agente em relação as estratégias, ou seja, a sua capacidade de identificar a melhor estratégia de acordo com a situação de conflito e o posicionamento do seu oponente. Quanto mais próxima de 1, maior é a certeza, do agente, de que aquela determinada estratégia é preferível. As Tabelas 8, 9 e 10, apresentam os valores médios de probabilidade que indicam o comportamento estável de IP, em cada cenário, de acordo com o jogo simulado.

Tabela 8 – Probabilidade média atingida durante o comportamento estável de IP - Dilema dos Prisioneiros

Parâmetros	RE	MRE	VRE	RE	MRE	VRE	RE	MRE	VRE
	Aleatório			Delatar			Não Delatar		
$\epsilon = \phi = 0$	0,645	0,649	1	0,642	0,653	1	0,653	0,646	1
$\epsilon = 0,02; \phi = 0$	0,642	0	0	0,639	0	0	0,639	0	0
$\epsilon = 0,04; \phi = 0$	0,633	0	0	0,625	0	0	0,636	0	0
$\epsilon = 0,08; \phi = 0$	0,625	0	0	0,612	0	0	0,622	0	0

Fonte: A autora (2019)

O comportamento de IP, quando orientado por RE, também, sofre influência do parâmetro de experimentação. No entanto, isso ocorre de maneira sutil e quase imperceptível. À medida que ϵ aumenta, as probabilidades de escolha, no RE, apresentam tendência ao centro do gráfico, fazendo com que a certeza sobre aquela estratégia diminua. É possível observar um padrão de comportamento em todos os resultados obtidos para o algoritmo RE que estão exibidos nas Tabelas 8, 9 e 10. Esse resultado pode ser explicado, dada a função do parâmetro de experimentação, no processo de aprendizagem. Os autores do algoritmo original, Erev e Roth (1998), incorporaram ϵ como um viés comportamental que se refere à experimentação das estratégias disponíveis, ou seja, o agente adota as estratégias disponíveis até ter certeza sobre qual é a preferível a cada cenário e isso permite que ele explore todas as opções e evita que ele adote uma estratégia menos preferível e mantenha seu comportamento fixo.

Quanto maior a frequência com que determinada estratégia é adotada, maiores valores de probabilidade serão atribuídos a ela, pois as recompensas serão incorporadas durante a atualização da escolha. Um comportamento com viés de experimentação, por sua vez, reflete probabilidades menores, tendo em vista a baixa repetição inicial.

Já quando variado dentro dos algoritmos MRE e VRE, o parâmetro de experimentação

Tabela 9 – Probabilidade média atingida durante o comportamento estável de IP - Batalha dos Sexos

Parâmetros	RE	MRE	VRE	RE	MRE	VRE	RE	MRE	VRE
	Aleatório			Futebol			Cinema		
$\epsilon = \phi = 0$	0,433	0,428	0	0,587	0,582	1	0,347	0,355	0
$\epsilon = 0,02; \phi = 0$	0,434	1	1	0,579	1	1	0,356	1	1
$\epsilon = 0,04; \phi = 0$	0,440	1	1	0,576	1	1	0,362	1	1
$\epsilon = 0,08; \phi = 0$	0,448	1	1	0,564	1	1	0,375	1	1

Fonte: A autora (2019)

faz com que o agente não possua capacidade de se posicionar estrategicamente, inviabilizando os seus usos, com essas configurações, para este tipo de estudo. Buscou-se uma explicação acerca da justificativa para tal comportamento, mas nada parecido foi encontrado na literatura. A inversão no comportamento do agente, no algoritmo MRE, foi corrigida quando os parâmetros ϕ e ϵ foram empregados ao mesmo tempo, uma vez que ϕ equilibra ϵ. Nesse sentido, nos cenários em que ϕ é mantido igual ou superior a ϵ o comportamento do agente que aprende volta a confirmar os equilíbrios teóricos dos jogos testados, mesmo que a partir de um comportamento instável.

Tabela 10 – Probabilidade média atingida durante o comportamento estável de IP - *Chicken-Game*

Parâmetros	RE	MRE	VRE	RE	MRE	VRE	RE	MRE	VRE
	Aleatório			Seguir			Desistir		
$\epsilon = \phi = 0$	0,415	0,414	0	0,575	0,570	1	0,370	0,372	0
$\epsilon = 0,02; \phi = 0$	0,417	1	1	0,563	1	1	0,378	1	1
$\epsilon = 0,04; \phi = 0$	0,422	1	1	0,567	1	1	0,385	1	1
$\epsilon = 0,08; \phi = 0$	0,424	1	1	0,562	1	1	0,395	1	1

Fonte: A autora (2019)

Portanto, dentre os três algoritmos que foram incorporados ao comportamento do agente IP, a versão original, *Roth-Erev*, demonstrou melhor desempenho diante dos estímulos realizados no decorrer do estudo, durante a simulação dos três jogos. Apesar de demonstrar sensibilidade às variações de ϕ, o resultado esperado permaneceu sendo observado, ou seja, o algoritmo foi capaz de proporcionar ao agente, a aprendizagem necessária para que o mesmo adotasse em cada situação a estratégia mais indicada.

6 CONCLUSÃO

O presente estudo emprega abordagens que vêm sendo utilizadas mais frequentemente, nos últimos anos, para estudar o comportamento individual e as maneiras com que a combinação de diferentes comportamentos impactam no resultado global. O emprego da modelagem baseada em agentes visa a construção dos agentes de maneira a representar o comportamento como ele é na realidade, atribuindo a cada agente as suas especificidades. É assumido, dessa forma, que os indivíduos são diferentes e que essas diferenças precisam ser consideradas, principalmente, quando mais de um indivíduo é abordado no estudo.

Situações em que se faz necessário o posicionamento estratégico, estão presentes nos mais variados campos de atuação. O processo decisório é inerente ao ser humano, visto que ele toma decisões a todo tempo, sendo elas de nível baixo a alto de complexidade. Já o pensamento estratégico precisa ser trabalhado.

A partir dos jogos clássicos que foram testados, foi possível observar que diferentes cenários podem ser obtidos a depender do comportamento estratégico dos indivíduos que interagem entre si. Enquanto um dos indivíduos, tomava decisão de forma fixa ou aleatória, sem desempenhar o pensamento estratégico, o outro recorria a um histórico de informações e de posicionamentos anteriores, para respaldá-lo acerca do próximo passo.

Os algoritmos de aprendizagem que foram incorporados ao comportamento do agente que aprende, deram-lhe a possibilidade de questionar as estratégias que foram adotadas a cada rodada de simulação. E mesmo dotado de um processo de aprendizagem que lhe permitia estar em situação de vantagem, em relação ao outro indivíduo, em muitas situações, o agente adotou estratégias que o levaram a cenários que não representavam o melhor resultado que poderia ser alcançado. Esse tipo de comportamento pode ser explicado pela racionalidade limitada, tendo em vista que mesmo que o indivíduo apresente comportamento racional, a racionalidade não garante que o mesmo tomará as melhores decisões em 100% das situações. O caso particular do jogo Dilema dos Prisioneiros reflete uma afirmação de Steingraber e Fernandez (2013), a respeito da racionalidade limitada, em que ele relata que em muitos casos o agente maximizador não adotará a estratégia que lhe proporcione a melhor recompensa, apenas uma recompensa satisfatória. Nesse sentido, durante a busca pelo resultado máximo, individualmente, ambos os jogadores acabam obtendo um resultado que não representa o melhor cenário de acordo com a matriz de incentivos.

A aprendizagem incorporada ao comportamento do agente, no modelo de simulação, faz com que o comportamento dele se aproxime do comportamento previsto em situações reais,

em que indivíduos diferentes precisam entrar em acordo sobre determinadas questões. Os três jogos que foram testados no estudo são, na verdade, metáforas de situações de conflito que se repetem diariamente.

A partir das repetidas simulações levando em consideração os três diferentes comportamentos de resposta do agente que aprende e empregando diferentes combinações dos parâmetros de esquecimento e experimentação, conclui-se que o algoritmo *Roth-Erev RL* apresenta resultados mais robustos, de uma maneira geral, em comparação aos demais algoritmos estudados. Como apresentado nos resultados, o algoritmo RE foi capaz de confirmar os equilíbrios teóricos sob todas as interferências feitas no decorrer das simulações como, por exemplo, a variação dos parâmetros de experimentação e esquecimento, tanto quando foram variados individualmente quanto de forma conjunta. Resultado este que não foi obtido com o mesmo desempenho pelos demais algoritmos.

Diferentemente dos algoritmos *Modified Roth-Erev RL* e *Variant Roth-Erev RL*, o RE continuou confirmando os resultados canônicos para cada um dos jogos testados mesmo após os parâmetros serem submetidos a variações. Para este tipo de estudo e empregando os valores que compõem a matriz de incentivos, como foram empregados, os algoritmos MRE e VRE demonstraram sensibilidade às variações do parâmetro de experimentação, levando o agente a tomar decisões que não representavam a melhor situação da matriz de incentivos. Esse comportamento se confirmou para os três jogos.

O parâmetro de experimentação provocou variações significativas no comportamento do agente IP, que passou a adotar estratégias de não-equilíbrio com maior frequência. Porém, esse comportamento pôde ser corrigido ao incorporar valores para ϕ, tendo em vista que o parâmetro de esquecimento neutralizou os efeitos do parâmetro de experimentação, proporcionando resultados mais coerentes a medida em que ϕ se demonstrava igual ou superior a ϵ. As simulações que empregaram o algoritmo *Modified Roth-Erev*, mantendo os parâmetros ϕ e ϵ iguais a 0, apresentaram resultados que convergem para os equilíbrios teóricos dos três jogos analisados.

A etapa de interpretação dos resultados apresentou limitações relacionadas a escassa literatura acerca do funcionamento dos algoritmos RE, MRE e VRE com foco no processo decisório utilizando teoria dos jogos e, principalmente, considerando a premissa de racionalidade limitada. A ausência de trabalhos que possam ser usados como base durante a determinação dos valores dos parâmetros de esquecimento e experimentação, também foi um fator limitante do trabalho. Dessa forma, propõe-se o desenvolvimento de trabalhos que explorem a influência desses vieses comportamentais no processo de aprendizagem, tendo em vista que tanto os valores

quanto os limites de variação estabelecidos para ϕ e ϵ foram determinados de acordo com o trabalho de Pentapalli (2008), não havendo justificativas fortes acerca desses valores.

Uma sugestão de trabalho futuro é empregar os algoritmos de aprendizagem no comportamento de todos os agentes envolvidos na situação de conflito que está sendo simulada, e a partir disso determinar se diferentes resultados podem ser alcançados quando ambos possuem a capacidade de mapear o comportamento do outro, podendo, dessa forma, proporcionar resultados mais representativos com os obtidos em situações reais. Até mesmo utilizar tanto a modelagem ABM quanto os algoritmos de aprendizagem para modelar outros jogos, não somente matriciais, que reflitam situações de conflitos reais.

Por fim, conclui-se que este tipo de estudo que simula o comportamento de indivíduos em situação de conflito incorporando as diferenças inerentes ao comportamento destes e incorporando aos mesmos a capacidade de aprendizagem, possibilitam que avaliações sejam feitas sob diferentes perspectivas, podendo assim explicar resultados que muitas vezes não são esperados.

REFERÊNCIAS

ABAR, S.; THEODOROPOULOS, G. K.; LEMARINIER, P.; M.P.O'HARE, G. Agent based modelling and simulation tools: A review of the state-of-art software. **Computer Science Review**, v. 24, n. 33, p. 13–33, maio 2017.

ALEDO, P. G. de; VLADIMIROV, A.; MANCA, M.; BAUGH, J.; ASAI, R.; KAISER, M.; BAUER, R. An optimization approach for agent-based computational models of biological development. **Advances in Engineering Software**, v. 121, p. 262–275, 2018.

ALEXANDRE, M.; LIMA, G. T. Combining monetary policy and prudential regulation: an agent-based modeling approach. **Springer**, 2017.

ALIABADI, D. E.; KAYA, M.; SAHIN, G. Competition, risk and learning in electricity markets: An agent-based simulation study. **Applied Energy**, n. 195, p. 1000–1011, abril 2017.

ARAUJO, F. C.; LEONETI, A. B. Game theory and 2x2 strategic games applied for modeling oil and gas industry decision-making problems. **Pesquisa Operacional**, SciELO Brasil, v. 38, n. 3, p. 479–497, 2018.

BELL, A. R. Informing decisions in agent-based models d a mobile update. **Environmental Modelling Software**, v. 93, p. 310e321, 2017.

BLOK, D. J.; LENTHE, F. J. van; VLAS, S. J. de. The impact of individual and environmental interventions on income inequalities in sports participation: explorations with an agent-based model. **International Journal of Behavioral Nutrition and Physical Activity**, v. 107, n. 15, 2018.

BOUKERCHE, A.; MACHADO, R. B.; JUCá, K. R.; SOBRAL, J. B. M.; NOTARE, M. S. An agent based and biological inspired real-time intrusion detection and security model for computer network operations. **Computer Communications**, v. 30, p. 2649–2660, 2007.

BUSCH, J.; ROELICH, K.; BALE, C. S. E.; KNOERI, C. Scaling up local energy infrastructure; an agent-based model of the emergence of district heating networks. **Energy Policy**, v. 100, p. 170–180, 2017.

CALABRIA, F. A.; SARAIVA, J. T.; ROCHA, A. P. Improving the brazilian electricity market: how to replace the centralized dispatch by decentralized market-based bidding. **Journal of Energy Markets**, v. 11, n. 2, p. 83–106, Junho 2018.

CLIFF, O. M.; HARDING, N.; PIRAVEENAN, M.; ERTEN1, E. Y.; GAMBHIR, M.; PROKOPENKO, M. Investigating spatiotemporal dynamics and synchrony of influenza epidemics in australia: An agent-based modelling approach. **Simulation Modelling Practice and Theory**, v. 87, p. 412–431, Setembro 2018.

CREPALDI, A. F.; FERREIRA, F. F.; RODRIGUES, J. de S. Jogo da minoria: um modelo baseado em agentes aplicado ao mercado financeiro. **Gestão Produção**, v. 19, n. 4, p. 793–809, 2012.

DILAVER, O.; JUMP, R. C.; LEVINE, P. Agent-based macroeconomics and dynamic stochastic general equilibrium models: Where do we go from here? **Journal of Economic Surveys**, p. 1–26, 2018.

EREV, I.; ROTH, A. E. Predicting how people play games: Reinforcement learning in experimental games with unique, mixed strategy equilibri. **The American Economic Review**, v. 88, n. 4, p. 848–881, 1998.

FARIAS, O. L. M. de; SANTOS, N. dos. Agent-based geographical information system. **IEEE**, 2005.

GAIVORONSKAIA, E.; TSYPLAKOV, A. Using a modified erev-roth algorithm in an agent-based electricity market model. **Journal of the New Economic Association**, v. 39, p. 55–83, 2018.

GIRARDI, R.; MARINHO, L. B.; OLIVEIRA, I. R. de. A system of agent-based software patterns for user modeling based on usage mining. **Interacting with Computers**, v. 17, p. 567–591, 2005.

GRIMM, V.; BERGER, U.; BASTIANSEN, F.; ELIASSEN, S.; GINOT, V.; GISKE, J.; GOSS-CUSTARD, J.; GRAND, T.; HEINZ, S. K.; HUSE, G.; HUTH, A.; JEPSEN, J. U.; JøRGENSEN, C.; MOOIJ, W. M.; MüLLER, B.; PE'ER, G.; PIOU, C.; RAILSBACK, S. F.; ROBBINS, A. M.; ROBBINS, M. M.; ROSSMANITH, E.; RüGER, N.; STRAND, E.; SOUISSI, S.; STILLMAN, R. A.; VABø, R.; VISSERAN, U.; DEANGELIS., D. L. A standard protocol for describing individual-based and agent-based models. **Ecological Modelling**, v. 198, p. 115–126, 2006.

GUERCI, E.; RASTEGAR, M. A.; CINCOTTI, S. Agent-based modeling and simulation of competitive wholesale electricity markets. **Energy Systems**, 2010.

HAFEZALKOTOB, A.; MAHMOUDI, R.; HAJISAMI, E.; WEE, H. M. Wholesale-retail pricing strategies under market risk and uncertain demand in supply chain using evolutionary game theory. **Kybernetes**, Emerald Publishing Limited, v. 47, n. 6, p. 1178–1201, 2018.

LI, M.; NGUYEN, B.; YU, X.; HAN, Y. Competition vs. collaboration: a four set game theory–innovation, collaboration, imitation, and'do nothing'. **International Journal of Technology Management**, Inderscience Publishers (IEL), v. 76, n. 3-4, p. 285–315, 2018.

MACAL, C.; NORTH, M. Tutorial on agent-based modelling and simulation. **Journal of Simulation**, v. 4, n. 3, p. 151–162, 2010.

MANTOVI, A.; SCHIANCHI, A. A game-theoretic traverse analysis: Price competition and strategic investment. **Structural Change and Economic Dynamics**, Elsevier, 2018.

MELO, T. M.; FUCIDJI, J. R. Racionalidade limitada e a tomada de decisão em sistemas complexos. **Revista de Economia Política**, Directory of Open Access Journals, v. 36, n. 3, p. 622–645, 2016.

MOYA, I.; CHICA, M.; SáEZ-LOZANO, J. L.; CORDóN Óscar. An agent-based model for understanding the influence of the 11-m terrorist attacks on the 2004 spanish elections. **Knowledge-Based Systems**, v. 123, p. 200–216, 2017.

MUREDDU, M.; MEYER-ORTMANNS, H. Extreme prices in electricity balancing markets from an approach of statistical physics. **Physica A**, v. 490, p. 1324–1334, 2018.

NICOLAISEN, J.; PETROV, V.; TESFATSION, L. Market power and efficiency in a computational electricity market with discriminatory double-auction pricing. **IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTATION**, v. 5, n. 5, p. 504–523, Outubro 2001.

PENTAPALLI, M. A comparative study of roth-erev and modied roth-erev reinforcement learning algorithms for uniform-price double auctions. **Iowa State University**, 2008.

RADHAKRISHNAN, B. M.; SRINIVASAN, D.; LAU, Y. F. A.; PARASUMANNA, B. G.; RATHORE, A. K.; PANDA, S. K.; KHAMBADKONE, A. A reinforcement learning algorithm for agent-based computational economics (ace) model of electricity markets. **IEEE Congress on Evolutionary Computation (CEC)**, Setembro 2015.

RINGLER, P.; KELES, D.; FICHTNER, W. Agent-based modelling and simulation of smartelectricity grids and markets – a literature review. **Renewable and Sustainable Energy Reviews**, v. 57, p. 205–215, 2016.

RODRIGUEZ-FERNANDEZ, J.; PINTO, T.; SILVA, F.; PRAçA, I.; VALE, Z.; CORCHADO, J. Context aware q-learning-based model for decision support in the negotiation of energy contracts. **Electrical Power and Energy Systems**, v. 104, p. 489–501, 2018.

SANCHEZ-CARTAS, J. M. Agent-based models and industrial organization theory. a price-competition algorithm for agent-based models based on game theory. **Complex Adaptive Systems Modeling**, SpringerOpen, v. 6, n. 1, p. 2, 2018.

SANCHEZ, S. M.; LUCAS, T. W. Exploring the world of agent-based simulations: Simple models. complex analyses. **Winter Simulation Conference**, 2002.

SENSFUB, F.; RAGWITZ, M.; GENOESE, M.; MöST, D. Agent-based simulation of electricity markets: a literature review. **ECONSTOR**, n. 5, p. 97–121, 2007.

SHIFLET, A. B.; SHIFLET, G. W. An introduction to agent-based modeling for undergraduates. **Procedia Computer Science**, v. 29, p. 1392–1402, 2014.

SIMON, H. A. A behavioral model of rational choice. **The quarterly journal of economics**, MIT Press, v. 69, n. 1, p. 99–118, 1955.

SMAJGL, A.; BROWN, D. G.; VALBUENA, D.; HUIGEN, M. G. Empirical characterisation of agent behaviours in socio-ecological systems. **Environmental Modelling Software**, v. 26, p. 837–844, Fevereiro 2011.

STEINGRABER, R.; FERNANDEZ, R. G. A racionalidade limitada de herbert simon na microeconomia. **Revista da Sociedade Brasileira de Economia Política**, v. 34, p. 123, 2013.

STREIT, R. E.; BORENSTEIN, D. An agent-based simulation model for analyzing the governance of the brazilian financial system. **Expert Systems with Applications**, v. 36, p. 11489–11501, 2009.

SUN, J.; TESFATSION, L. Dynamic testing of wholesale power market designs: An open-source agent-based framework. **Computational Economics**, v. 30, p. 291–327, Agosto 2007.

SUN, Z.; LORSCHEID, I.; MILLINGTON, J. D.; LAUF, S.; MAGLIOCCA, N. R.; GROENEVELD, J.; BALBI, S.; NOLZEN, H.; MüLLER, B.; SCHULZE, J.; BUCHMANN, C. M. Simple or complicated agent-based models? a complicated issue. **Environmental Modelling Software**, v. 86, p. 56–67, Setembro 2016.

THOBER, J.; SCHWARZ, N.; HERMANS, K. Agent-based modeling of environment-migration linkages: a review. **Ecology and Society**, v. 23, n. 2, 2018.

TISUE, S.; WILENSKY, U. Netlogo: Design and implementation of a multi-agent modeling environment. **International Conference on Complex Systems**, Maio 2004.

TRACY, M.; CERDA, M.; KEYES, K. M. Agent-based modeling in public health: Current applications and future directions. **Annual Review of Public Health**, n. 39, p. 77–94, 2018.

URBINA, D. A.; RUIZ-VILLAVERDE, A. A critical review of homo economicus from five approaches. **American Journal of Economics and Sociology**, Wiley Online Library, v. 78, n. 1, p. 63–93, 2019.

WEIDLICH, A.; VEIT, D. A critical survey of agent-based wholesale electricity market models. **Energy Economics**, n. 30, p. 1728–1759, janeiro 2008.

ZHOU, Z.; CHAN, W. K. V.; CHOW, J. H. Agent-based simulation of electricity markets: a survey of tools. **Energy Policy**, n. 28, p. 305–342, julho 2009.

ZSCHACHE, J. Melioration learning in two-person games. **PloS one**, Public Library of Science, v. 11, n. 11, 2016.

APÊNDICE A – ROTINA DE PROGRAMAÇÃO DO MODELO DE SIMULAÇÃO

```
breed [players player]
breed [IntelligentPlayers IntelligentPlayer]
breed [Conclusions Conclusion]
breed [Actions Action]
directed-link-breed [opinions opinion]
globals [Combinations
         strategy
         hidden-action
         NoCooperation
         Cooperation
         Combination
         Profit
         Rodadas
         R S T P C D]

players-own [payoff CurrentAction Cooperate?]
IntelligentPlayers-own [payoff CurrentAction]
Actions-own [payoff CurrentAction]
opinions-own [Propensity Probability]
conclusions-own [payoff CurrentAction]

to setup
  clear-all
  SetupActions
  SetupGames
  reset-ticks
  Rounds
end

to Rounds
set Rodadas 0
end

to SetupActions
  foreach ["C" "D"]
          [ create-ordered-Actions 1 [set color blue
           set size 3
           set shape "dot"
           setxy 15 3]]

ask Actions with [label = "C"] [rt 180 fd 0 set payoff 5]
ask Actions with [label = "D"] [rt 180 fd 1 set payoff 5]
end

to SetupGames
```

```
  set-default-shape players "person"
   create-players 1 [set heading 270
                     set color pink
                     set size 5
                     fd max-pxcor / 2
                     True
                    ]

  set hidden-action one-of ["act-randomly" "cooperate"
  "defend"]

 set-default-shape Conclusions "circle"
   ask patches with [pxcor = 0 and pycor = 0] [sprout-Conclusions 1]
   ask Conclusions [set color gray
                    set size 4]

  set-default-shape IntelligentPlayers "person"
  create-IntelligentPlayers 1 [set color blue
                               set heading 90
                               set size 5
                               fd abs min-pxcor / 2

  create-opinions-to actions [set label [who] of other-end

  set Propensity InitialPropensity
                              ]
    ask opinions [set Probability (exp(Propensity /
    CoolingParameter) / sum [exp(Propensity /
    CoolingParameter)] of [opinions] of myself)]

    set currentAction one-of actions with [who = 0]
                    set label [label] of currentAction
                    set payoff [payoff] of currentAction
  ]
end

to go
   go-a-round
   DetermineCombinations
   DetermineProfits
   LearningActivities
   ;tick
   DoPlot
   UpdateGames
end

to go-a-round
  ask players [ run strategy-player ]
```

```
end

to aleatorio
  set Cooperate? one-of [ true false ]
  if Cooperate? = true [set strategy 0]
  if Cooperate? = false [set strategy 1]
end

to NaoDelatar
  set Cooperate? true
  if Cooperate? = true [set strategy 0]
  if Cooperate? = false [set strategy 1]
end

to Delatar
  set Cooperate? false
  if Cooperate? = true [set strategy 0]
  if Cooperate? = false [set strategy 1]
end

to DetermineCombinations
  if strategy = 0 and CurrentAction = Action 0 [set Combination "R"]
  if strategy = 0 and CurrentAction = Action 1 [set Combination "S"]
  if strategy = 1 and CurrentAction = Action 0 [set Combination "T"]
  if strategy = 1 and CurrentAction = Action 1 [set Combination "P"]
end

to DetermineProfits
  ask IntelligentPlayers [if Combination = "R" [set Profit 5]]
  ask IntelligentPlayers [if Combination = "S" [set Profit 10]]
  ask IntelligentPlayers [if Combination = "T" [set Profit 0]]
  ask IntelligentPlayers [if Combination = "P" [set Profit 1]]
end

to LearningActivities
   if Algorithm = "Roth-Erev RL Algorithm" [UpdateOpinions1]
   if Algorithm = "Modified Roth-Erev" [UpdateOpinions2]
   if Algorithm = "Variant Roth-Erev RL Algorithm" [UpdateOpinions3]
   ChooseNewCurrentAction
end

to UpdateOpinions1
   ask IntelligentPlayers [if currentAction = Action 0
        [ask my-out-opinions with [label = 0]
        [set Propensity ((1 - RecencyParameter) * Propensity) +
        ((1 - ExperimentationParameter) * [profit] of myself)
       ]
        ask  my-out-opinions with [label != 0]
```

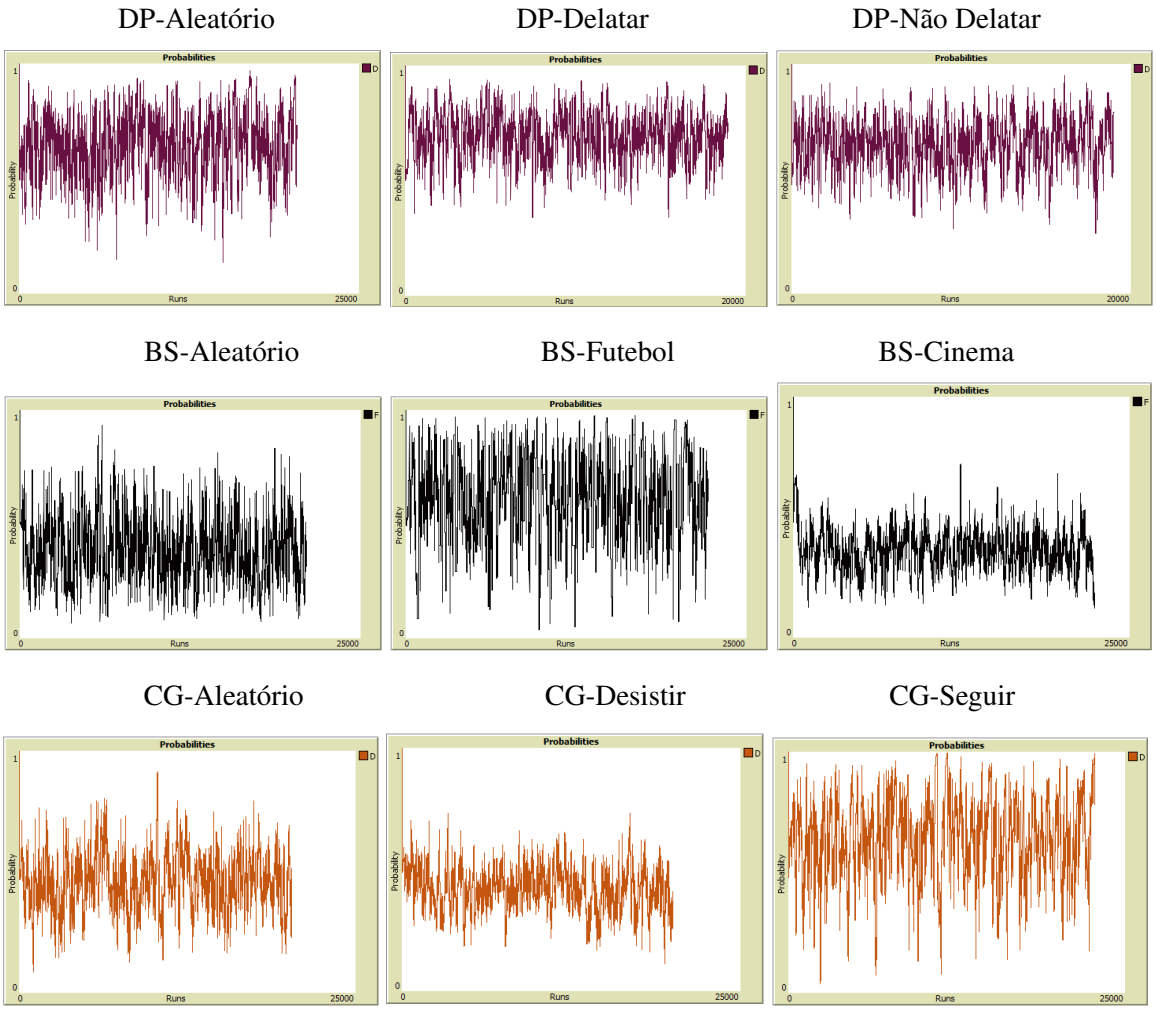
```
         [set Propensity ((1 - RecencyParameter) *
         Propensity) + ([profit] of myself *
         (ExperimentationParameter / 1))
        ]
         ask my-out-opinions [set Probability (Propensity) /
         (sum [Propensity] of [opinions] of myself)
                 if Probability > 1 [set Probability 1]
                 if Probability < 0 [set Probability 0]
                                                    ]
        ]

                              if currentAction = Action 1
          [ask my-out-opinions with [label = 1]
          [set Propensity ((1 - RecencyParameter) *
          Propensity) + ((1 - ExperimentationParameter) *
          [profit] of myself)
        ]
          ask  my-out-opinions with [label != 1]
          [set Propensity ((1 - RecencyParameter) * Propensity) +
          ([profit] of myself * (ExperimentationParameter / 1))
        ]
          ask my-out-opinions [set Probability (Propensity) /
          (sum [Propensity] of [opinions] of myself)
                  if Probability > 1 [set Probability 1]
                  if Probability < 0 [set Probability 0]
                                                    ]
        ]]
end

to UpdateOpinions2
ask IntelligentPlayers [if currentAction = Action 0
          [ask my-out-opinions with [label = 0]
          [set Propensity ((1 - RecencyParameter) *
          Propensity) + ((1 - ExperimentationParameter) *
          [profit] of myself)
         ]
           ask  my-out-opinions with [label != 0]
          [set Propensity ((1 - RecencyParameter) *
          Propensity) + (ExperimentationParameter *
          (Propensity / 1))
         ]
           ask my-out-opinions [set Probability Propensity /
           (sum [Propensity] of [opinions] of myself)
                   if Probability > 1 [set Probability 1]
                   if Probability < 0 [set Probability 0]
         ]
         ]
                          if currentAction = Action 1
```

```
          [ask my-out-opinions with [label = 1]
          [set Propensity ((1 - RecencyParameter) *
          Propensity) + ((1 - ExperimentationParameter) *
          [profit] of myself)
        ]
           ask  my-out-opinions with [label != 1]
          [set Propensity ((1 - RecencyParameter) *
          Propensity) + (ExperimentationParameter *
          (Propensity / 1))
        ]
           ask my-out-opinions [set Probability Propensity /
           (sum [Propensity] of [opinions] of myself)
                  if Probability > 1 [set Probability 1]
                  if Probability < 0 [set Probability 0]
        ]
        ]]
end

to UpdateOpinions3
ask IntelligentPlayers [if currentAction = Action 0
         [ask my-out-opinions with [label = 0]
         [set Propensity ((1 - RecencyParameter) *
         Propensity) + ((1 - ExperimentationParameter) *
         [profit] of myself)
        ]
          ask  my-out-opinions with [label != 0]
          [set Propensity ((1 - RecencyParameter) *
          Propensity) + (ExperimentationParameter *
          (Propensity / 1))
        ]
          ask my-out-opinions [set Probability
          (exp(Propensity / CoolingParameter)) /
          (sum [exp(Propensity /
          CoolingParameter)] of [opinions] of myself)]
        ]
                       if currentAction = Action 1
          [ask my-out-opinions with [label = 1]
          [set Propensity ((1 - RecencyParameter) *
          Propensity) + ((1 - ExperimentationParameter) *
          [profit] of myself)
        ]
          ask  my-out-opinions with [label != 1]
          [set Propensity ((1 - RecencyParameter) *
          Propensity) + (ExperimentationParameter *
          (Propensity / 1))
        ]
          ask my-out-opinions [set Probability
          (exp(Propensity / CoolingParameter)) /
```

```
          (sum [exp(Propensity /
          CoolingParameter)] of [opinions] of myself)]
        ]]
end

to ChooseNewCurrentAction
    if [label] of PrevailingOpinion = 0
    [ask IntelligentPlayers [set CurrentAction Action 0]]
    if [label] of PrevailingOpinion = 1
    [ask IntelligentPlayers [set CurrentAction Action 1]]
end

to-report PrevailingOpinion
    let pick random-float 1
    let winner nobody
    ask opinions
        [;; if there's no winner yet...
        while [winner = nobody and probability > 0.000001
        ]
        [ifelse probability > pick
                [set winner self ]
                [set pick pick - probability]
        ]
        ]
        report winner
end

to DoPlot
   PlotRodadas
   PlotProbabilities
end

to PlotRodadas
set Rodadas Rodadas + 1
end

to PlotProbabilities
  set-current-plot "Probabilities"
  set-current-plot-pen "D"
  ask opinions with [label = 1] [plot Probability]
end
```

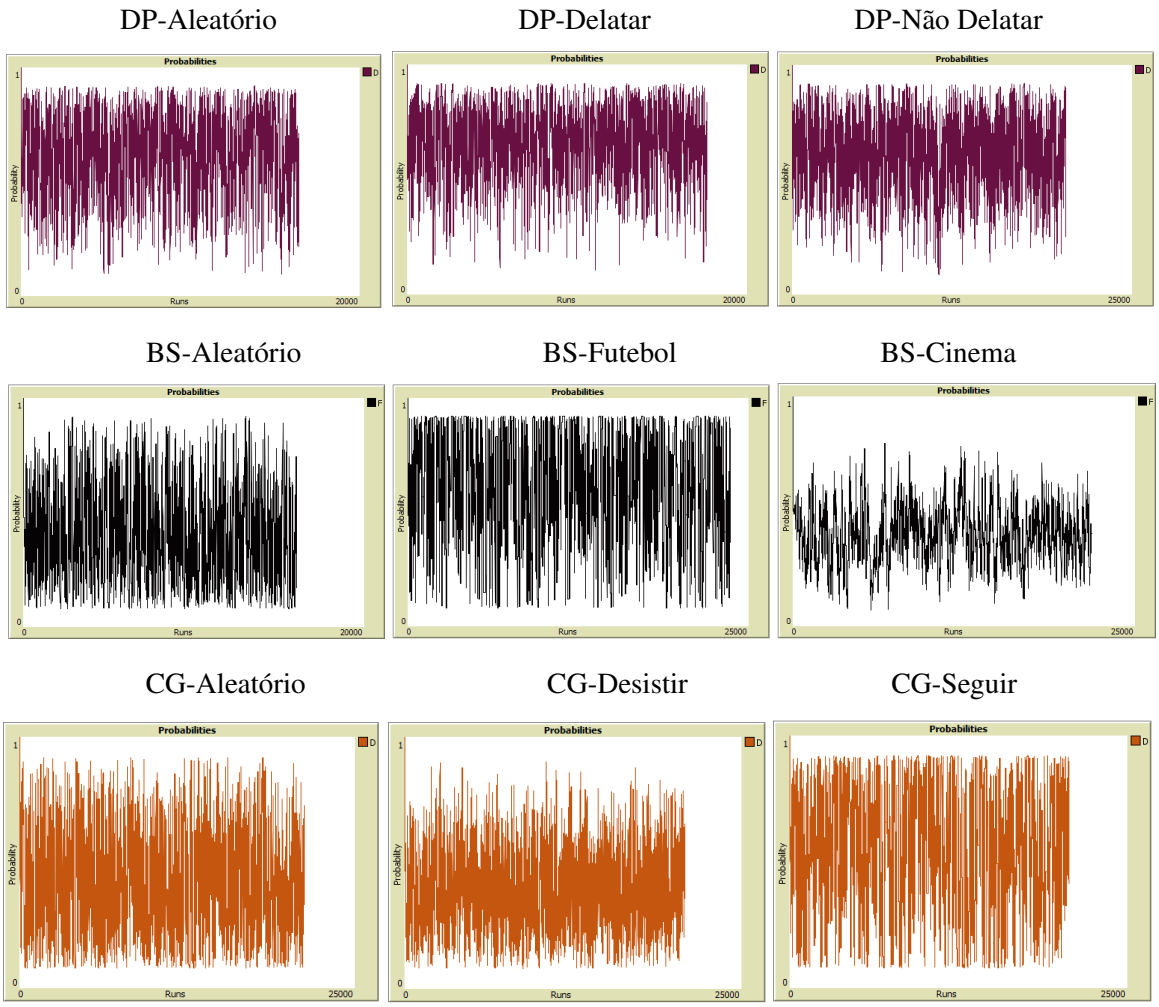
APÊNDICE B – GRÁFICOS COMPLEMENTARES

Figura 23 – Resultados das simulações dos três jogos clássicos, com = 0,02 eφ = 0,03, para as três possibilidades de comportamento do agente P - Agente IP aprende a partir do RE



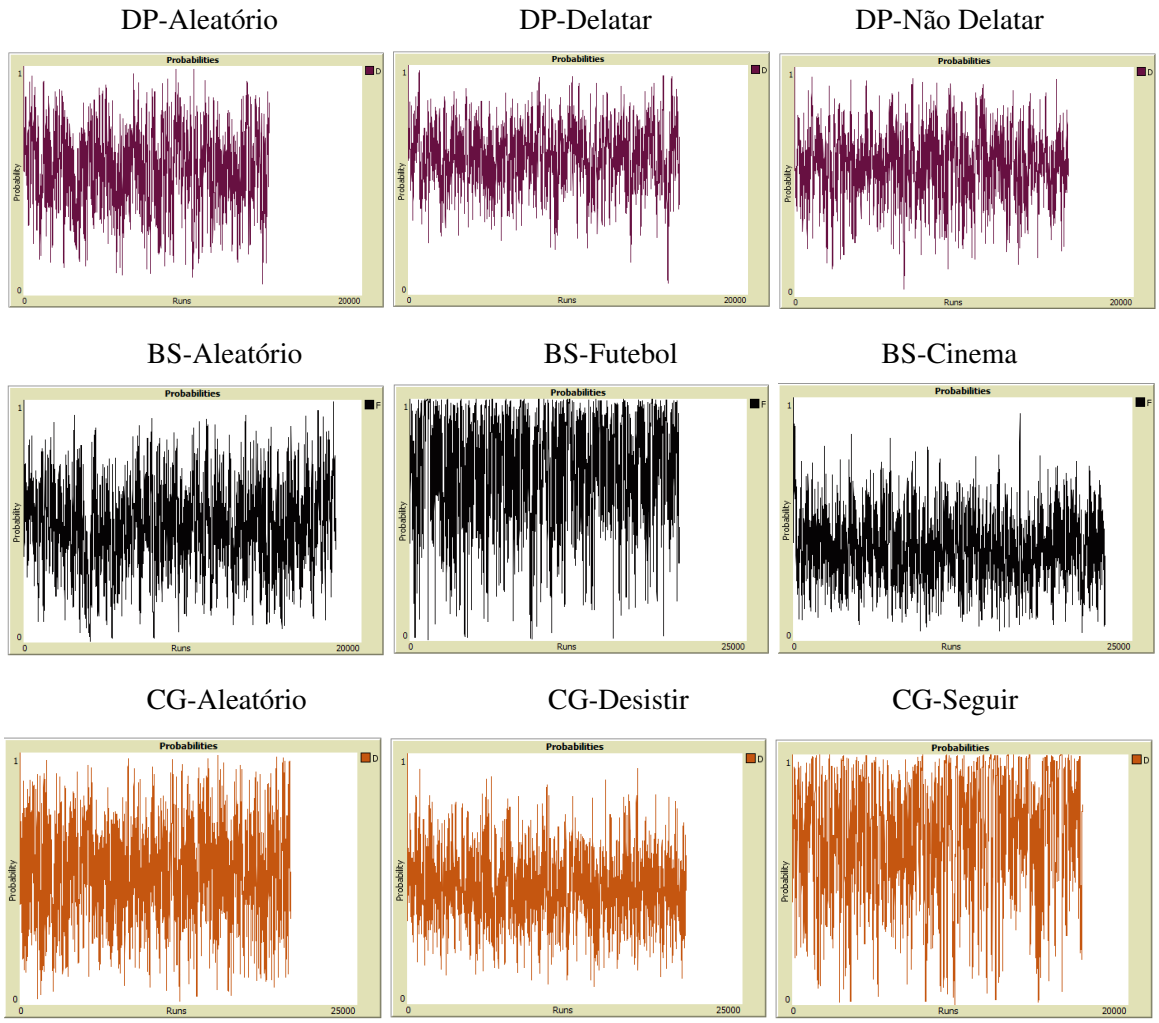
Fonte: A autora (2019)

Figura 24 – Resultados das simulações dos três jogos clássicos, com = 0,08 eφ = 0,09, para as três possibilidades de comportamento do agente P - Agente IP aprende a partir do RE



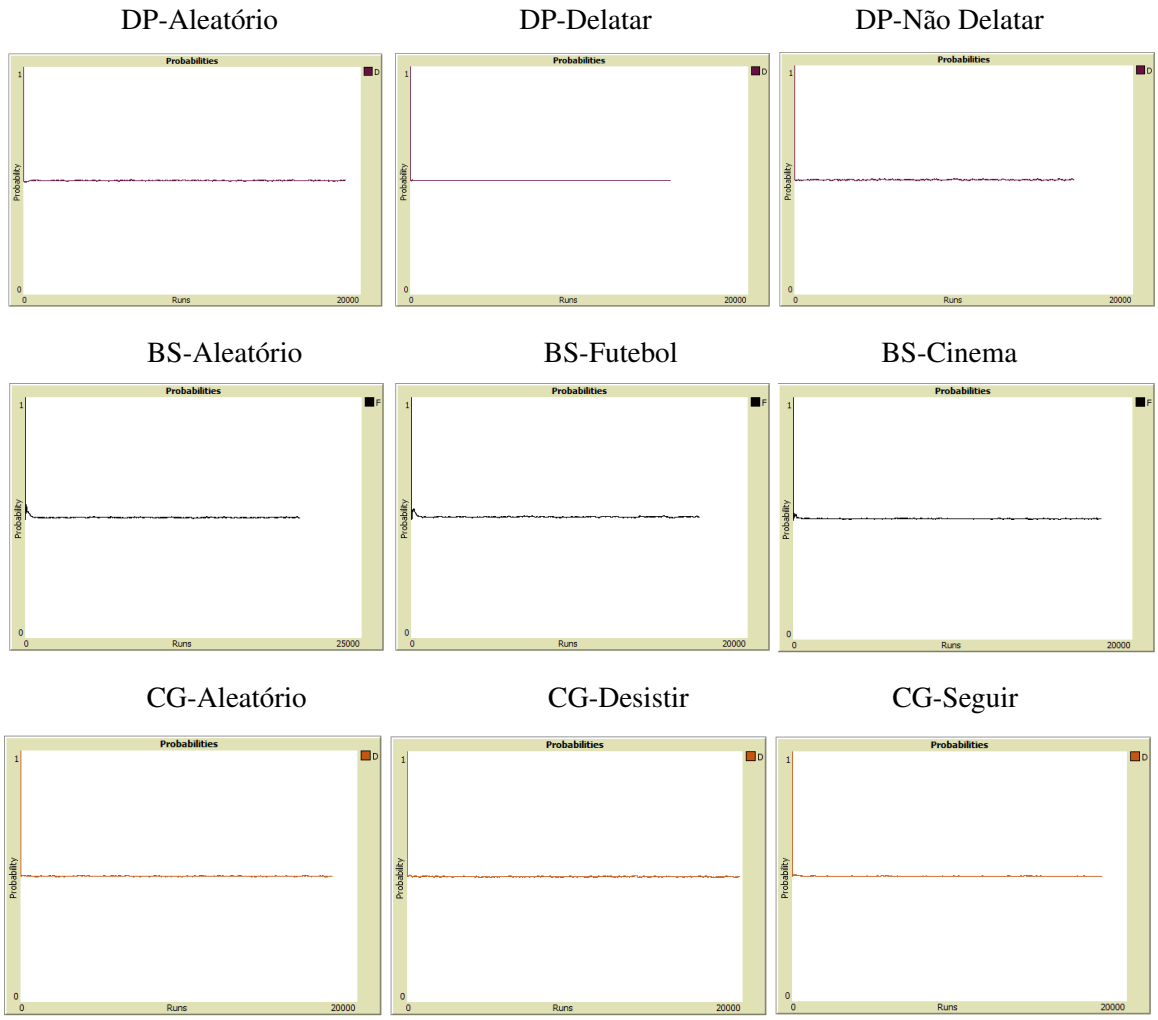
Fonte: A autora (2019)

Figura 25 – Resultados das simulações dos três jogos clássicos, com = 0,08 eφ = 0,09, para as três possibilidades de comportamento do agente P - Agente IP aprende a partir do MRE



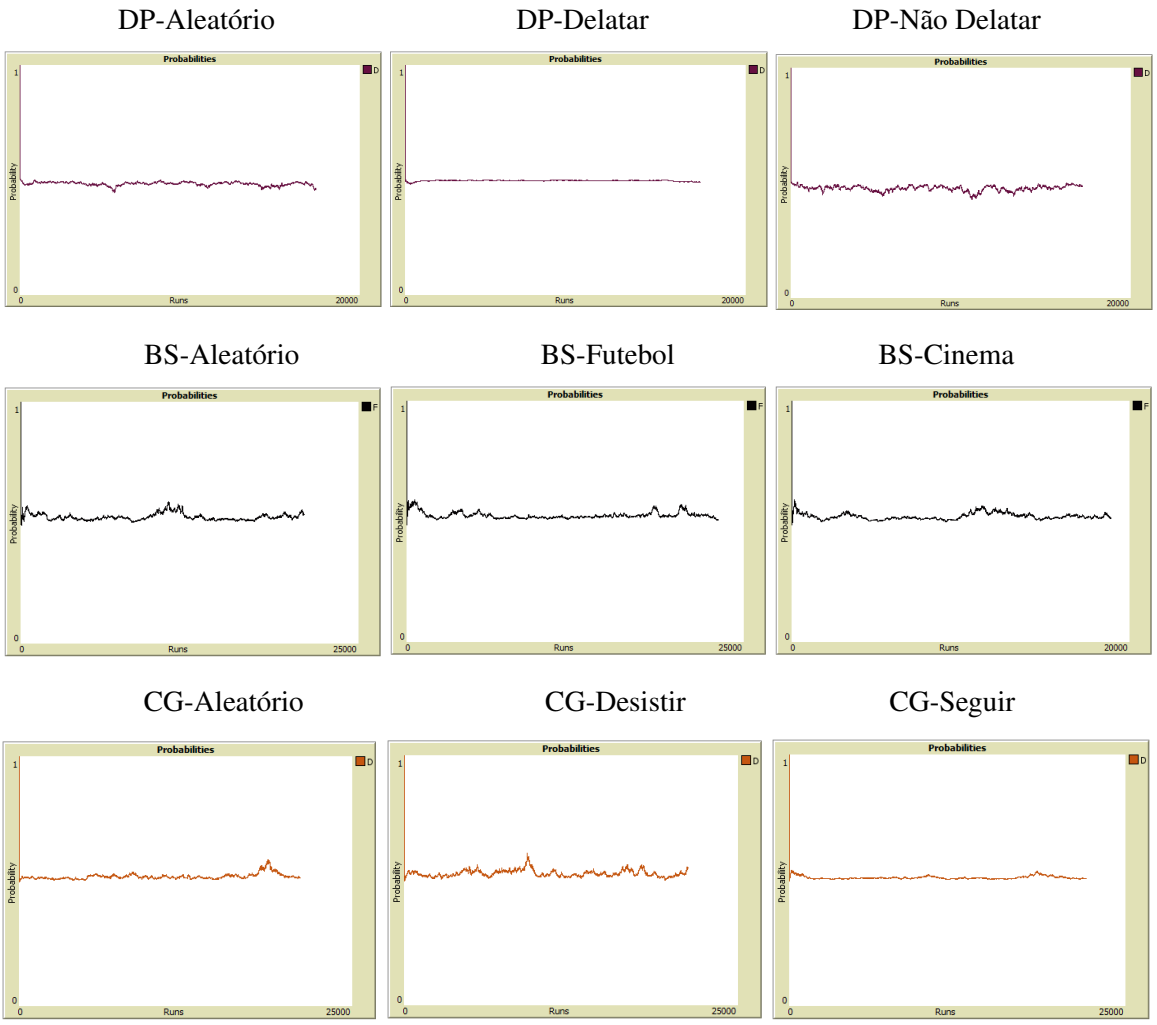
Fonte: A autora (2019)

Figura 26 – Resultados das simulações dos três jogos clássicos, com = 0,02 e φ = 0,02, para as três possibilidades de comportamento do agente P - Agente IP aprende a partir do VRE



Fonte: A autora (2019)

Figura 27 – Resultados das simulações dos três jogos clássicos, com = 0,03 e φ = 0,02, para as três possibilidades de comportamento do agente P - Agente IP aprende a partir do VRE



Fonte: A autora (2019)