Luiz Eduardo Tassi

Desempenho de ratos em jogo estratégico e modulação
dopaminérgica

São Paulo

2011

Luiz Eduardo Tassi

Desempenho de ratos em jogo estratégico e modulação
dopaminérgica deste desempenho

*Performance of rats in a strategic game and
dopaminergic modulation of their choice policy*

Tese apresentada ao Instituto de
Biociências da Universidade de São Paulo,
para a obtenção de Título de Doutor em
Ciências, na Área de Fisiologia Geral.

Orientador(a): Gilberto Fernando Xavier

São Paulo

2011

Tassi, Luiz Eduardo

O desempenho de ratos em jogo estratégico e os efeitos da modulação dopaminérgica / Luiz Eduardo Tassi ; orientador Gilberto Fernando Xavier. –São Paulo, 2011.

141 f.

Tese (Doutorado) – Universidade de São Paulo, Instituto de Biociências da Universidade de São Paulo. Departamento de Fisiologia.

1. Tomada de decisão. 2. Neuromodulação. I. Universidade de São Paulo, Instituto de Biociências, Departamento de Fisiologia. II. Título.

Comissão Julgadora:

Prof(a). Dr(a).

Prof(a). Dr(a).

Prof(a). Dr(a).

Prof(a). Dr(a).

Prof. Dr. Gilberto Fernando Xavier

Orientador

“Venho de um país vertiginoso onde a loteria é base da realidade.”

Jorge Luis Borges

Agradecimentos

Ao Prof. Gilberto Xavier pela orientação e apoio que deu a este trabalho.

Ao Prof. Paulo Justiniano Ribeiro (UFP) pelas valiosas análises estatísticas.

À Evalnete Rodrigues por todo o companheirismo durante este longo percurso.

Resumo

A interação entre agentes inteligentes na disputa por recursos necessários à sobrevivência é um fato onipresente na luta pela vida. Este tipo de interação é estudado e formalizado matematicamente pela teoria dos jogos. Na literatura experimental encontramos vários estudos envolvendo primatas humanos e não humanos em tarefas de jogos estratégicos, mas, até o momento, não foi desenvolvido nenhum modelo deste comportamento com roedores.

Estudos do comportamento animal mostram que estes aprendem e aprimoram este tipo de estratégias através de aprendizagem por reforço. O elemento central dos modelos computacionais de aprendizado por reforço é o sinal de violação de expectativa, que sinaliza o quanto um resultado difere, para mais ou para menos, do esperado. Este sinal é utilizado pelo agente para atualização dos valores e, assim, da probabilidade de escolha das opções. A pesquisa neurofisiológica tem consistentemente demonstrado que o sinal fásico emitido pelo sistema dopaminérgico conforma-se muito de perto às características do sinal descrito pela teoria computacional.

Dessa maneira, os objetivos do presente estudo são pesquisar (1) se roedores são capazes de jogar um jogo estratégico simples e se a evolução do seu desempenho é consistente com o aprendizado por reforço e (2) se os efeitos de agonistas e antagonistas dopaminérgicos na estratégia de jogo são consistentes com a teoria segundo a qual o sinal dopaminérgico fásico desempenha função central na atualização constante da estratégia de jogo. Nossos resultados demonstram que, neste jogo estratégico, roedores efetivamente são capazes de um desempenho muito próximo do normativo, que seu desempenho é consistente com o aprendizado por reforço e, finalmente, que o sistema dopaminérgico está envolvido no processo de atualização da estratégia.

Abstract

Intelligent agents competing for the resources necessary for survival is a universal factor in the struggle for life. This type of interaction has been studied and mathematically formalized by game theory. In scientific literature we have come across several studies involving human and non-human primates carrying out strategic game tasks; however, until now, no model for such behavior has been developed for rodents.

Animal behavior studies have shown that animals learn and develop strategies through reinforcement learning. A central element of computational models of reinforcement learning is the reward-prediction error signal, which indicates how much a result differs, either positively or negatively, from the expected result. This signal is used by the agent to update the values of its options, and so their probability of being chosen. Neurophysiologic research has consistently shown that the phasic signal emitted by the dopamine system conforms very closely to the characteristics of the signal described by computational theory.

The purposes of this study are: (1) to discover whether rodents are capable of playing a simple strategic game and whether the evolution of their performance is consistent with reinforcement learning; and (2) whether the effects of dopamine agonists and antagonists on game strategy are consistent with the theory that phasic dopamine signals have a primary role in the constant update of game strategy. Our results prove that, in this strategic game, rodents are effectively capable of finding a strategy that is very close to the normative one, that their performance is consistent with reinforcement learning and, finally, that the dopamine system is involved in the process of strategic updating.

Índice

1 INTRODUÇÃO	10
1.1 Sobre pesquisas neuroeconômicas	12
1.2 Participação da Dopamina	14
1.3 Neurofisiologia da escolha	15
1.4 Teoria dos jogos	17
1.4.1 Formalização da teoria dos jogos	23
1.5 Abordagem neuroeconômica dos transtornos mentais	26
1.6 Aprendizagem por reforço	30
1.6.1 Modelagem matemática da aprendizagem por reforço	33
1.6.2 Aprendizagem por reforço e modelos do ambiente	39
1.6.3 Modelos do circuito córtico-estriato-tálamo-cortical e a modulação dopaminérgica	46
1.7 “Matching pennies”	59
1.8 Breve síntese	66
2. OBJETIVO	70
3. EXPERIMENTOS	71
3.1 Experimento I. Desempenho de ratos no jogo estratégico de “matching pennies”	71
3.1.1 Material e Métodos	71
3.1.2 Resultados (do Experimento I)	81
3.1.3 Discussão	87
3.2 Experimento II. Os efeitos de agonistas e antagonistas dopaminérgicos na estratégia de jogo dos animais	92
3.2.1 Material e Método	94
3.2.2 Resultados (do Experimento II)	97
3.2.2.1 Agonistas Dopaminérgicos Específicos	98
3.2.2.1.1 SKF 81297	98
3.2.2.1.2 Bromocriptina	101
3.2.2.2 Antagonistas Dopaminérgicos Específicos	103
3.2.2.2.1 SCH 23390	103
3.2.2.2.2 Eticloprida	104
3.2.2.3 Agonista Dopaminérgico Inespecífico	104
3.2.3 Síntese dos Resultados	108
3.2.3.1 Discussão	109
3.3 Discussão Geral	124

3.3.1 Discussão	124
4 CONCLUSÃO	129
REFERÊNCIAS	130

1 INTRODUÇÃO

A perspectiva da neuroeconomia é a da integração de ideias oriundas da psicologia, economia e neurociência para a produção de modelos acurados dos processos de tomada de decisão, em outras palavras, o estudo neurocientífico do comportamento econômico, no sentido amplo de decidir qual é a escolha mais adaptativa. Tem como meta a aplicação da topologia de David Marr no estudo da tomada de decisão, isto é, a análise do processo do ponto de vista computacional, algorítmico e de implementação neural. Dentro desta perspectiva, entende-se por computacional a definição do problema a ser resolvido (por exemplo, a determinação da melhor escolha quando diante de duas opções que diferem em magnitude, prazo e probabilidade). O nível algorítmico descreve o modo como esta computação é feita (descreve o *input*, o *output* e o algoritmo para transformação). Por fim, o nível da implementação descreve como esta transformação de *input* em *output* é feita na estrutura neural (para revisão ver GLIMCHER, 2003a; NIV, 2009).

As origens da neuroeconomia (para uma revisão GLIMCHER *et al.*, 2009) estão no nascimento da economia neoclássica nos anos 30 e da neurociência cognitiva nos anos 90 do século passado. No início dos anos 30, um grupo de economistas começou a estudar a estrutura matemática do comportamento de escolhas que resultaria do uso de ideias básicas mais primitivas ou axiomas. A partir destes axiomas pode-se deduzir uma série de variáveis intermediárias da teoria econômica, tais como a de utilidade, que posteriormente serviu de base para o desenvolvimento de teorias econômicas gerais. A extensão destes estudos levou ao tratamento da escolha probabilística e da escolha em que os resultados são obtidos após um prazo. Posteriormente, vários questionamentos vieram a ser feitos a respeito da validade dos resultados do método axiomático por economistas e psicólogos que passaram a denominar-se economistas comportamentais. Basicamente, o argumento deste grupo é que ideias provenientes da psicologia poderiam melhorar os modelos de comportamento humano gerados pela

economia neoclássica; propõe-se modelos de limites na capacidade de cálculo racional, força de vontade e interesse pessoal, e procura-se codificar estes limites formalmente e explorar suas implicações empíricas. Por exemplo, ao contrário da visão da economia neoclássica, propõe-se que os agentes econômicos valorizam diferentemente ganhos e perdas da mesma magnitude, tendo a perda maior valor subjetivo que o ganho. Esta evolução no pensamento econômico foi muito importante para o nascimento da neuroeconomia, porque os economistas comportamentais também frequentemente propunham teorias que podiam ser traduzidas como algoritmos versando sobre como a informação é processada e as escolhas como resultando deste processamento de informação.

A outra vertente na origem da neuroeconomia situa-se na evolução da neurociência cognitiva. Tal como na economia, também aí encontramos a interação de duas correntes: no caso, as correntes neurológica e neurofisiológica. Na neurologia os estudos centravam-se em efeitos facilmente observáveis no desempenho de humanos ou animais de laboratório portadores de lesões principalmente de sistemas sensoriais ou motores. Já estudos com lesões que visavam o estudo de operações cognitivas ou estados emocionais eram mais raros em função da dificuldade de estudá-los precisamente. Mas, são precisamente estes estudos que permitem a compreensão das bases neurais da tomada de decisão e processos afins estudados pela neuroeconomia. A abordagem fisiológica envolve a correlação de medidas diretas do funcionamento neural, tais como potenciais de ação, mudanças de fluxo sanguíneo, liberação de neurotransmissores, com eventos do mundo externo. Inicialmente estes métodos foram muito eficientes para o estudo das características básicas do funcionamento nervoso, mas extremamente limitados, como no caso dos estudos neurológicos, na sua aplicabilidade ao estudo de fenômenos cognitivos e emocionais.

No contexto da evolução das técnicas de estudo do sistema nervoso, menção especial deve ser feita aos progressos da neuroimagem que permitiram, de modo não invasivo, uma melhor compreensão da relação entre função neural e atividade cognitiva. Diante da grande quantidade de

dados gerados pelos progressos das novas técnicas de estudo da função neural o desafio passou a ser a criação do instrumental teórico para organizar estes dados.

Neste contexto, economistas comportamentais e cientistas cognitivos encontraram na neuroimagem funcional um instrumento para testar e desenvolver teorias alternativas às teorias neoclássicas. Ao mesmo tempo neurofisiologistas e neurocientistas cognitivos procuraram a teoria econômica como uma ferramenta para testar e desenvolver modelos algorítmicos dos sistemas neurais responsáveis pela tomada de decisão. Estas duas correntes, de certa forma, antagônicas, persistem até hoje dentro da pesquisa neuroeconômica.

1.1 Sobre pesquisas neuroeconômicas

A seguir apresentamos uma síntese dos resultados dos estudos neuroeconômicos realizados em primatas humanos e não humanos.

Um modelo de tomada de decisão deve responder duas questões fundamentais. Primeiramente, como os valores subjetivos das várias opções sob consideração são apreendidos, armazenados e representados. Em segundo lugar, como uma ação de alto valor é escolhida entre as demais opções para ser executada pelo sistema motor. De forma sumária, a literatura sugere que o processo de computar o valor das opções envolve setores ventromediais do córtex pré-frontal e as áreas associadas do estriado, enquanto que o processo de escolha está ligado às regiões corticais pré-frontais laterais e parietais (KABLE; GLIMCHER, 2009).

Uma característica central de todas as teorias sobre tomada de decisão é que as várias dimensões de uma opção devem ser integradas numa medida única de valor subjetivo. Comparações entre diferentes espécies de opção são baseadas nesta medida abstrata de valor subjetivo, algo como

uma “moeda comum” para permitir a escolha. Embora esta “moeda comum” seja postulada pela economia neoclássica para explicar a consistência das opções, entre os economistas a existência desta moeda é apenas teórica, não havendo para eles o menor interesse na questão de sua implementação psíquica ou neural nos tomadores de decisão. No entanto, em estudos neurofisiológicos são crescentes as evidências de que as representações de valor subjetivo de fato têm uma função ao nível dos algoritmos neurais e que estas representações são codificadas primariamente no córtex pré-frontal ventro-medial e no estriado (KABLE; GLIMCHER, 2009).

Padua-Schioppa e Assad (2006) registraram a atividade de neurônios do córtex órbito-frontal (OFC) enquanto macacos faziam escolhas entre pares de sucos que lhes eram oferecidos. Através das sequências de escolha comportamentais o valor subjetivo de cada oferta (tipo de suco e quantidade) foi calculado. A atividade de neurônios no córtex órbito-frontal foi registrada na busca de evidências deste valor subjetivo na hipotética escala (ou moeda) comum. Em 80% dos neurônios pesquisados, três padrões diferentes de resposta foram encontrados: neurônios cuja taxa de disparos era linearmente correlacionada com o valor subjetivo dos sucos ofertados no par. Este sinal poderia servir como um valor subjetivo, uma “moeda comum”, para comparação entre as opções oferecidas. Também havia neurônios cuja atividade era relacionada ao valor subjetivo do suco que havia sido escolhido e outros neurônios com uma resposta que codificava os sabores ofertados. Num estudo posterior, os mesmos pesquisadores mostraram que a atividade linearmente correspondente ao valor de cada oferta era invariável em relação à oferta concorrente, isto é, a atividade relativa à determinada quantidade de unidades de um determinado suco era a mesma diante de qualquer oferta concorrente (PADUA-SCHIOPPA; ASSAD, 2008).

Outros estudos documentaram respostas semelhantes no estriado. Lau e Glimcher (2008) registraram a atividade de neurônios do núcleo caudado enquanto macacos executavam uma tarefa de escolha óculo-motora. Encontraram neurônios cuja atividade representava o valor das ações, independentemente destas serem ou não escolhidas. É interessante observar o paralelismo destes

resultados com os de Padua-Schioppa e Assad (2006). Samejima e colaboradores (2005) também encontraram no estriado representações de valor de ações numa escala comum. Resumidamente, as respostas encontradas no caudado e putamen nestes dois estudos são paralelas às encontradas no OFC, exceto que representando o valor de ações em lugar do valor dos reforços.

Estudos em humanos envolvendo imagens funcionais por ressonância magnética (fMRI) levam a conclusões semelhantes. Kable e Glimcher (2007), estudando a escolha entre reforços monetários imediatos e postergados, encontraram que o sinal BOLD no córtex pré-frontal medial e no estriado ventral se correlacionava com o valor subjetivo de reforço postergado nos diversos blocos do experimento. Também a atividade neural variou nos diversos indivíduos de acordo com os índices individuais de desvalorização aplicados ao reforço em função do prazo de entrega.

1.2 Participação da Dopamina

Uma vez que as evidências sugerem a existência de representações neurais de valor subjetivo de reforços e ações, devemos nos perguntar como são geradas estas representações. Uma série de evidências, que veremos mais adiante, aponta para uma importante função do sistema dopaminérgico na geração e atualização destes sinais. O sistema dopaminérgico está na origem de três tipos de sinais que têm como alvo diferentes redes cerebrais. O mais conhecido e que ocupará nosso interesse neste trabalho é o sinal fásico dopaminérgico, que codifica o valor motivacional de reforços e estímulos condicionados relacionados a eles. Este sinal está relacionado a sistemas cerebrais ligados à avaliação, aprendizado e procura de reforços e fuga de estímulos aversivos. Um segundo sinal codifica a saliência motivacional dos estímulos, sejam eles reforços ou punições e tem como alvo redes cerebrais relacionadas à orientação, cognição e motivação geral. Finalmente um

terceiro sinal dopaminérgico tem uma função de desencadear estados de alerta (para uma revisão BROMBERG-MARTIN *et al.*, 2010)).

O aprendizado do valor de reforços e ações escolhidas se dá fundamentalmente pela experiência passada. Em experimentos com animais, onde não é possível a comunicação linguística, a experiência é o único modo de descoberta de valor de situações e ações. Entre os sinais provenientes do sistema dopaminérgico descritos acima, o sinal que codifica o valor motivacional é o que nos interessa no aprendizado. Ele corresponde a um breve pulso de atividade dos neurônios dopaminérgicos do tronco encefálico que ocorre todas as vezes que o agente se defronta tanto com um reforço inesperado, como com um sinal preditor deste reforço. Também, em situações em que um reforço previsto não ocorre temos uma modificação do nível de descarga dos neurônios dopaminérgicos; porém, neste caso, há uma diminuição no momento em que o reforço previsto não ocorre (SCHULTZ *et al.*, 1997). Este esquema geral foi observado tanto em animais de laboratório (por exemplo, BAYER *et al.*, 2007) como em estudos de neuroimagem com humanos (D'ARDENNE *et al.*, 2008). O tema da função dopaminérgica, por ser central neste trabalho, será desenvolvido mais extensivamente adiante.

1.3 Neurofisiologia da escolha

Através do processo de escolha a alternativa com maior valor subjetivo é identificada e os dados relativos a esta alternativa, mas não os relativos àquela de menor valor, são passados ao sistema motor para a execução da ação correspondente. Teoricamente o processo equivale a um mecanismo de comparação do valor das alternativas e à identificação da alternativa de maior valor. Os dados disponíveis atualmente implicam principalmente os córtices pré-frontal lateral e parietal no processo de seleção e implementação da alternativa escolhida. Sabe-se que este processo envolve

uma estocasticidade intrínseca, no sentido de que quando duas opções têm valores subjetivos próximos, aquela de menor valor pode ser também escolhida com uma menor frequência.

Teoricamente, podemos conceber o processo de escolha como envolvendo um mecanismo que compara duas ou mais opções e identifica a de maior valor. Qualquer sistema que faça esta comparação deve poder ser capaz de representar os valores de cada opção antes de fazer a escolha e, após esta, enviar a informação da opção selecionada para circuitos subordinados.

Os processos de decisão foram, por razões de ordem técnica, intensamente estudados no sistema visual sacádico dos primatas (GLIMCHER, 2003b). Este sistema é composto pelos campos oculares frontais, colículos superiores e área lateral intraparietal (LIP), que formam uma rede intensamente conectada, onde existem sucessivos mapas das diversas sacadas possíveis. Os dados disponíveis sugerem uma rede análoga composta pelo córtex pré-motor, área motora suplementar e área LIP no córtex parietal adjacente como desempenhando o mesmo tipo de função na tomada de decisão não sacádica dos primatas. Nos colículos superiores, onde existem mapas relativos à direção e amplitude das sacadas, no período que antecede a identificação da sacada de maior valor, as representações neurais das diversas sacadas contempladas têm sua atividade proporcional ao valor da sacada representada. Com a evolução da tomada de decisão, quando a sacada de maior valor é identificada, a representação desta tem sua atividade intensamente aumentada e a atividade dos neurônios representando a opção alternativa é diminuída, num mecanismo computacionalmente conhecido como “winner-take-all” (BASSO; WURTZ, 1998). Assim caracteriza-se o mecanismo acima descrito de representação de valor das opções e execução da escolha. Na área LIP, que envia informações aos colículos superiores, Platt e Glimcher (1999) mostraram que a atividade neuronal que ocorre antes do início da atividade intensa nos colículos superiores também está relacionada ao valor subjetivo das sacadas, sendo quase uma função linear da magnitude e probabilidade do reforço associado àquela sacada. Desta maneira, o mecanismo teórico para a escolha de uma ação estaria implementado, ao nível neural, nestas estruturas nervosas.

1.4 Teoria dos jogos

Na discussão acima tratamos da tomada de decisão de um agente solitário que, portanto, não tem os resultados de suas escolhas afetados pelas escolhas de outros. No entanto, para animais vivendo em grupo ou isoladamente em um ecossistema, não é esta a situação usual, pois o resultado de uma decisão tomada por um agente depende das decisões tomadas por outros. Basta para isto examinarmos situações muito frequentes da vida social, como por exemplo, um convite ou compra de ações em bolsa de valores, ou, num ecossistema, a caça ou disputa por um recurso com outros animais. Este tipo de tomada de decisão é estudado formalmente na economia pela teoria dos jogos. O estudo do comportamento humano ou de outros animais durante jogos nos dá a possibilidade de comparar o desempenho efetivo dos agentes com as previsões teóricas e, utilizando vários paradigmas experimentais da neurociência, examinar as estruturas neurais ativas durante os diferentes jogos, os efeitos de lesões ou de manipulações farmacológicas. Os benefícios deste tipo de abordagem são duplos. De um lado a neurociência pode revelar limites biológicos importantes nos processos neurocognitivos envolvidos que sejam importantes para a compreensão dos fatos empíricos. De fato, a pesquisa tem revelado que muitos dos processos envolvidos em tomada de decisões complexas envolvem os mesmos mecanismos usados em tomadas de decisão mais simples e fundamentais e, portanto, apresentarão características computacionais em parte similares. Outro ponto bastante importante que a pesquisa empírica tem revelado é que o comportamento de tomada de decisão dos animais raramente está em conformidade com as previsões teóricas. Consequentemente, uma caracterização mais precisa do comportamento é um passo importante no aperfeiçoamento destes modelos teóricos.

O que a teoria dos jogos fornece é um grupo de modelos bem especificados para o estudo da interação social. Em sua essência, a teoria dos jogos é uma coleção de modelos rigorosos para o entendimento e explicação de situações em que dois ou mais agentes devem decidir numa interação recíproca. Estes modelos fornecem, além de especificações sobre o número de jogadores, regras sobre a decisão que cada jogador pode tomar a cada momento e em posse de qual informação a respeito das escolhas dos outros. Além disto, especifica o retorno de cada decisão em função das escolhas de todos os outros.

Um conceito central na teoria dos jogos é o equilíbrio de Nash (NASH, 1950). Segundo este conceito, um grupo de jogadores, racionais e interessados em maximizar seus ganhos, vai convergir para um conjunto de estratégias, das quais nenhum jogador poderá aumentar seu ganho afastando-se unilateralmente. Em outras palavras, quando em equilíbrio de Nash os jogadores terão atingido a estratégia de retorno máximo. Vejamos alguns exemplos.

No Jogo do Ultimato o comportamento dos jogadores é diferente do previsto pela teoria dos jogos, o que nos permite avançar no estudo das motivações envolvidas. Neste jogo, o proponente deve dividir uma quantidade fixa de dinheiro, digamos \$10. Uma vez feita a proposta, o receptor pode aceitar ou rejeitar. Em caso de aceitação da proposta cada um recebe o valor proposto. Por exemplo, o proponente pode oferecer \$1 e conservar os outros \$9 para si. Se o receptor não aceitar a proposta, ambos não receberão nada. Se para cada jogador a motivação da ação é a maximização dos ganhos, podemos esperar que o proponente ofereça a menor fração possível da moeda e que o receptor aceite esta oferta, baseado na ideia de que qualquer ganho é melhor que nada. Prevendo isto, o proponente tenderá a fazer esta oferta mínima. No entanto, este não é o comportamento observado experimentalmente. Nos experimentos feitos em países industrializados, ofertas menores que 20% do total são rejeitadas 50% das vezes. Vemos aqui que o comportamento não corresponde ao previsto no modelo baseado puramente no interesse de maximização de ganhos. Uma hipótese plausível para explicar esse resultado seria a evolução de instintos competitivos, que levariam o receptor a rejeitar

ofertas muito díspares em favor de ofertas mais igualitárias, que levariam ao desenvolvimento da reciprocidade (NOWAK *et al.*, 2000). Este comportamento sofre alterações importantes se o jogador souber que está interagindo com um computador em vez de outro ser humano. Se souber que está interagindo com um computador, a aceitação das ofertas inferiores a 20% da soma sobe para 80%. Investigações sobre as bases neurais deste comportamento mostraram que a ativação da ínsula anterior aumenta à medida que a iniquidade (desigualdade) das ofertas aumenta. Quando o proponente é um humano este aumento da ativação é maior do que quando o proponente é um computador. No que diz respeito à recusa de uma oferta, a intensidade desta ativação tem valor preditivo (SANFEY *et al.*, 2003). A ínsula anterior também é responsiva a estímulos dolorosos e desagradáveis. Esta região, juntamente com outras envolvidas no processamento de emoções, pode contribuir para registrar certas interações sociais como aversivas e assim desencorajar a confiança em certos parceiros no futuro. Tais reações emocionais poderiam ter evoluído precisamente para desenvolver a reciprocidade mútua, tornar a reputação importante, e encorajar a punição daqueles que têm comportamentos que trazem desvantagens para outros membros do grupo (SANFEY, 2007).

Este exemplo mostra que jogos oferecem vantagens em relação a tarefas de tomada de decisão que não envolvem interação entre os agentes, pois entre outras coisas, permitem avaliar não apenas as bases neurofisiológicas desse comportamento, como processos complexos como reputação, confiança e cooperação.

A teoria dos jogos começou como teoria matemática e continua nesta tradição, estudando a solução ideal que jogadores ideais usariam para maximizar seus ganhos num jogo. Este modo de tratar o problema ignora o fato de que para os jogadores possam existir outros interesses mais complexos além da maximização quantitativa dos retornos, como no exemplo acima citado.

Estudos sobre as bases neurais da tomada de decisão podem lançar uma luz sobre os desvios do comportamento observados em relação às previsões da teoria dos jogos. Sabe-se que

neurônios do estriado têm resposta que é proporcional à magnitude do reforço ou punição (CROMWELL; SCHULTZ, 2003) e estas respostas, em humanos, não se limitam a reforços concretos como dinheiro. O estriado humano está centralmente envolvido em decisões sociais para além de qualquer ganho financeiro, uma vez que a reciprocidade na cooperação com outro humano no jogo do dilema do prisioneiro, descrito mais adiante, leva a uma ativação estriatal, o que não ocorre quando a mesma interação envolve um computador e mesma compensação monetária está em jogo (RILLING *et al.*, 2002). A reciprocidade de outro ser humano numa interação social tem, no estriado, manifestações semelhantes a de outros reforços concretos, inclusive sua evolução temporal como será descrito adiante.

No jogo da confiança, um jogador, o investidor, decide o quanto, de uma quantia que recebeu, investirá junto ao outro jogador, o curador. A parte que é transferida é multiplicada por um fator, geralmente três ou quatro na maioria dos experimentos. O curador deve devolver ao investidor parte ou toda a quantia que lhe foi confiada, mas pode também, se assim decidir, conservar para si tudo o que recebeu. Se o curador honrar a confiança e devolver pelo menos uma parte do que recebeu ao investidor, ambos terão, ao final, mais do que inicialmente; caso contrário, *i.e.*, se não devolver nada, o investidor terá uma perda. Quando os jogadores interagem só uma vez, a teoria dos jogos prevê que um curador racional, tentando maximizar seus ganhos, jamais honrará a confiança nele depositada pelo investidor e este, sabendo disto, não confiará nada ao curador. Apesar destas previsões, novamente observa-se que os investidores de fato confiam quantias aos curadores e estes devolvem parte do que foi confiado.

Quando o jogo da confiança é repetido diversas vezes envolvendo os mesmos jogadores, a ativação no núcleo caudado do curador relaciona-se à confiança demonstrada pelo investidor nas jogadas anteriores e corresponde a uma intenção de retribuição por parte deste. Também a evolução temporal deste sinal é semelhante ao sinal de erro de previsão dos modelos de aprendizagem por reforço. Inicialmente ocorre depois que o investidor faz a escolha, sendo que posteriormente o sinal

ocorre muito antes que o investidor revele sua escolha (KING-CASAS *et al.*, 2005). Também contrariando as previsões do equilíbrio de Nash quando os investidores tinham, em outra versão deste jogo, a possibilidade de punir os curadores que não mostravam reciprocidade na confiança depositada, não só o faziam, mas esta punição também estava associada a uma maior ativação no núcleo caudado. É importante ressaltar que esta punição tinha um custo para os investidores. Também neste caso havia ativação associada do núcleo caudado (QUERVAIN *et al.*, 2004).

Além dos fatores envolvendo preferências sociais, devemos considerar questões de capacidade computacional, uma vez que esta pode diferir da pressuposta nos jogadores ideais da teoria dos jogos. Esta limitação poderia levar ao uso de algoritmos de cálculo e decisão menos exigentes computacionalmente e conseqüentemente menos eficientes que os propostos. Outro ponto importante a considerar é que a teoria dos jogos apenas descreve a estratégia ideal, mas não como o agente chega a ela; portanto, ignora os efeitos da aprendizagem (CAMERER, 2003, 2009). O ponto central da teoria matemática dos jogos é o que um jogador pensa que o outro vai fazer e, em geral, este pensamento é iterado. Por exemplo, no jogo do dilema do prisioneiro analisado na próxima sessão, o jogador 1 (P1) imagina que o jogador 2 (P2) não vai colaborar, porque imagina que P2 imagina que P1 não vai colaborar. Este processo dá origem a respostas mutuamente consistentes, neste caso, nem P1 nem P2 colaboram. Em alguns jogos este processo pode ser bastante complexo, limitando o número de iterações principalmente quando o jogo é novo para os participantes. Tomemos como exemplo o jogo chamado concurso de beleza-p (“p-beauty-contest”), excelente para o estudo do pensamento estratégico. Neste jogo, os jogadores escolhem simultaneamente um número entre zero e 100. O jogador que tiver escolhido o número que é “p vezes” a média dos números escolhidos pelo grupo ganha um prêmio fixo. O valor de p é sempre menor que um, em geral 2/3. Este jogo modela situações de mercado como leilões de serviços públicos, em que vence quem oferecer o menor preço. Nesta situação, o ideal é diminuir os preços para vencer a concorrência, mas não demais para não comprometer a lucratividade. O equilíbrio de Nash para este jogo é a escolha do número zero, que é o

único número que é a melhor escolha para todos. No entanto, não é o que encontramos experimentalmente. Nos experimentos encontramos que alguns jogadores escolhem números aleatoriamente entre zero e 100, outros escolhem p vezes 50 (a média de escolhas aleatórias no intervalo proposto). Outros jogadores, mais sofisticados, com maior capacidade analítica e capazes de antever as estratégias dos outros, escolhem p vezes o resultado anterior, isto é, p^2 vezes 50. Estes resultados, repetidos em várias pesquisas, ilustram não só as dificuldades do pensamento estratégico, como diferentes capacidades de pensamento estratégico entre os diferentes jogadores, fatores que explicam os desvios do equilíbrio previsto teoricamente. Por outro lado, quando este jogo é repetido diversas vezes com o mesmo grupo e os resultados são conhecidos, as escolhas tendem gradualmente a zero, *i.e.*, a estratégia de equilíbrio.

Quando um jogo é repetido várias vezes, os jogadores podem aproximar as estratégias de equilíbrio, como nos mostra o exemplo discutido acima. Este progresso está relacionado ao aprendizado. Existem vários modelos de aprendizado nos jogos. Um modelo é o aprendizado por reforço em que os jogadores repetem as estratégias que produziram bons resultados. Outro modelo é o jogo fictício (“belief-learning”), no qual os jogadores usam a história das jogadas dos outros para formar crenças sobre o que os outros farão e respondem de acordo com estas crenças. Jogo fictício é matematicamente equivalente ao modelo de aprendizagem por reforço generalizado, em que as estratégias não escolhidas têm seu valor atualizado segundo o resultado que teriam fornecido caso tivessem sido escolhidas. Um terceiro modelo é de atração ponderada pela experiência (“experience-weighted-attraction”, ou EWA) no qual os jogadores dão um peso aos ganhos contrafactuais de estratégias não escolhidas. O jogo fictício e o aprendizado por reforço modelam melhor a realidade que as teorias do equilíbrio, pois estas não prevêm qualquer aprendizado, o que as torna menos realistas do ponto de vista cognitivo. EWA é a mais robusta das três, pois suas previsões são tão boas quanto as dos outros dois modelos, mas ela é mais exata em relação ao curso temporal do aprendizado

humano (CAMERER, 2003). Já experimentos de jogos com primatas não humanos podem ser adequadamente modelados com aprendizado por reforço (LEE *et al.*, 2004).

Em síntese, na teoria dos jogos o conceito de equilíbrio de Nash (NASH, 1950) nos fornece a estratégia para a qual convergem jogadores ideais levando em conta apenas seu interesse de maximizar ganhos em cada rodada do jogo. Também assume que todos tenham a mesma capacidade cognitiva e que esta seja suficiente para compreender perfeitamente a complexidade do jogo desde o primeiro contato e a realizar as computações necessárias para jogar tendo em vista a maximização dos ganhos. O desvio constante das previsões de equilíbrio que encontramos na pesquisa experimental ocorre por violação das duas classes de pressupostos acima: os jogadores humanos não são guiados apenas pelas perspectivas de ganho em cada rodada de um jogo, pois agem também motivados por interações sociais incluindo aversão à desigualdade e sentimento de reciprocidade, entre outros. Outro tipo de desvio em relação às estratégias de equilíbrio é que estas exigem uma capacidade computacional que nos jogadores reais, humanos ou não, apresenta limitações que não são homogeneamente distribuídas nas populações; ademais, estas estratégias requerem um período de aprendizagem para seu desenvolvimento até o ponto máximo possível para cada jogador.

1.4.1 Formalização da teoria dos jogos

Para introduzirmos a formalização matemática utilizada na teoria dos jogos (KRUEGER *et al.*, 2008) começaremos com o jogo do dilema do prisioneiro. Neste jogo os dois jogadores devem decidir simultânea e isoladamente cooperar ou não cooperar. A melhor opção para ambos é a mútua cooperação com um resultado de, digamos, 2 unidades para cada um (2, 2), comparada com o retorno (1, 1) se ambos não cooperam. Mas caso os jogadores tenham escolhas diferentes, aquele que escolhe não cooperar tem um ganho bem maior que o que escolheu cooperar (3, 0). Neste caso, a

escolha ótima, o equilíbrio de Nash, é a não cooperação para ambos, pois as possibilidades são receber 3 caso o outro jogador escolha cooperar, ou receber 1 caso o outro jogador escolha também não cooperar, sem, no entanto, correr o risco de nada receber.

		P2	
		C	D
P1	C	2 2	3 0
	D	0 3	1 1

FIGURA 1 – MATRIZ DO JOGO DO DILEMA DO PRISIONEIRO (PD). P1 E P2: JOGADORES, C (COOPERAR) E D (NÃO COOPERAR).

Nash (1950) propôs uma teoria para prever o comportamento dos jogadores. Este comportamento é o Equilíbrio de Nash (NE). Num jogo de estratégia pura com dois jogadores, no qual um dos jogadores escolhe uma única estratégia s_1^i , do conjunto $S_1 = \{s_1^1, s_1^2, s_1^3, \dots, s_1^n\}$ e o outro uma pura estratégia s_2^j do conjunto $S_2 = \{s_2^1, s_2^2, s_2^3, \dots, s_2^n\}$, o NE é um conjunto de estratégias (s_1^*, s_2^*) que satisfaz as seguintes condições:

$$U_1(s_1^*, s_2^*) \geq U_1(s_1, s_2^*) \quad \text{sendo } s_1 \in S_1$$

$$U_2(s_1^*, s_2^*) \geq U_2(s_1^*, s_2) \quad \text{sendo } s_2 \in S_2$$

Obviamente no caso do jogo do dilema do prisioneiro a não colaboração mútua é o único par de estratégias que satisfaz as duas condições acima. Tanto P1 quanto P2 têm como estratégias apenas a colaboração ou não colaboração, isto é, o mesmo par de estratégias (C, D). Assim, vemos na matriz acima que o resultado para P1 quando ele colabora e P2 não colabora é zero, e para P2 é 3.

$$U_1(C, D) = 0 \text{ e}$$

$$U_2(C, D) = 3$$

O único par de estratégias que satisfaz as condições do equilíbrio de Nash é (D, D), pois:

$$U_1(D, D) = 1 \text{ e}$$

$$U_2(D, D) = 1$$

uma vez que, por exemplo, se P1 escolhe D, a escolha de P2 que maximiza seu ganho é também D, caso contrário receberá zero e P1 receberá três. Pelo mesmo motivo P1 escolherá D. A escolha de colaborar, antevista pelo oponente que procura a maximização de seus resultados, levá-lo-á a D.

Existem jogos para os quais não existe um par de estratégias que satisfaçam as duas condições acima como, por exemplo, o jogo de *matching pennies*. Neste jogo, dois jogadores escolhem qual face de uma moeda colocar sobre a mesa. Se ambos colocarem a mesma face um dos jogadores ganha (*matcher*), caso contrário o outro será vitorioso (*non-matcher*). Vale notar que se trata de uma variante cultural do jogo do par-ou-ímpar e que tem exatamente a mesma estrutura matemática.

		P2	
		H	T
P1	H	-1 +1	+1 -1
	T	+1 -1	-1 +1

FIGURA 2 – MATRIZ DE RESULTADOS DO JOGO DE *MATCHING PENNIES* (MP), P1 E P2 REPRESENTAM OS JOGADORES, H = CARA, T = COROA (EXTRAÍDO DE KRUEGER *et al*, 2008)

Vê-se nesta matriz que não existe um par de estratégias que satisfaça as inequações acima. Não existe um par de estratégias que seja a melhor escolha para os dois jogadores. Nash ampliou seu conceito de equilíbrio para estratégias mistas, isto é, distribuições de probabilidade sobre S_1 e S_2 .

Agora o equilíbrio ocorrerá não mais pela escolha de um par de estratégias puras como no jogo do dilema do prisioneiro (no caso (D, D)), mas pela probabilidade com que são escolhidas as estratégias. No caso do MP, se a estratégia de escolher H for 50% para um dos jogadores, o outro não poderá aumentar seu ganho escolhendo com outra probabilidade. De fato se o fizer, digamos optar por H 70% das vezes, o outro jogador usará esta informação em seu benefício. Isto nos leva a outra característica dos equilíbrios mistos: a necessidade da randomização da sequência de escolhas, pois no caso de escolhas previsíveis, dentro de distribuição de probabilidades, o outro jogador usará esta informação em seu benefício. Esta mesma questão da imprevisibilidade está presente também nos jogos de estratégia pura como o dilema do prisioneiro. Formalmente, o equilíbrio de Nash dos jogos de estratégia mista com duas escolhas e dois jogadores, P1 e P2, como o MP, é o par de estratégias (p, q) que deve satisfazer as seguintes condições:

$$EU_1(p^*, q^*) \geq EU_1(p, q^*) \quad e$$

$$EU_2(p^*, q^*) \geq EU_2(p^*, q)$$

onde p e q são as probabilidades de escolha de uma estratégia (H, por exemplo) e EU_1 e EU_2 as expectativas matemáticas de retorno (utilidade) para P1 e P2, respectivamente. Neste caso prova-se que na situação de equilíbrio a expectativa matemática de ganho para cada jogador é a mesma para cada estratégia pura, isto é, em equilíbrio a expectativa de ganho para cada jogador é a mesma tanto para escolha de cara ou coroa.

1.5 Abordagem neuroeconômica dos transtornos mentais

A disponibilidade de dados envolvendo desempenho em tarefas de tomada de decisão econômica e atividade neural concomitante permite pesquisar as relações entre ambos em estados patológicos do sistema nervoso e compará-las com dados de indivíduos sadios. Isto abre a

possibilidade de utilizar tarefas de tomada de decisão econômica como instrumento de pesquisa em doenças neuropsiquiátricas e eventualmente como instrumento auxiliar no diagnóstico.

Por exemplo, Bickel *et al.* (2008) pesquisaram a desvalorização temporal do reforço em humanos dependentes de nicotina e em humanos sadios. Os resultados mostraram que uma função hiperbólica de desvalorização descreve e diferencia os dois grupos. O comportamento de escolha dos não fumantes mostrou uma desvalorização hiperbólica, como o dos fumantes. No entanto, os fumantes desvalorizaram perdas e ganhos hipotéticos passadas e futuras significativamente mais que os controle não fumantes. Em outro estudo (PAULUS *et al.*, 2008), dependentes de metanfetamina (e seu respectivo grupo-controle) tiveram seu desempenho e ativação neural medidos por fMRI comparados numa tarefa que avaliava a mudança de estratégia após erros. Estes erros ocorriam em três frequências distintas em diferentes blocos do experimento. Comparando os dependentes de metanfetamina aos controles, evidenciou-se que os usuários da droga apresentaram um menor ajuste da estratégia após erros. Eles mostraram uma menor probabilidade de manter respostas corretas e também de mudar seu comportamento após respostas que levavam a perdas. Também foram constatadas diferenças nos padrões de ativação neural nos córtices insular esquerdo e pré-frontal dorsolateral bilateral em função da média de erros de cada etapa da tarefa. O grau de ajuste de estratégia em função da média de erros do bloco estava correlacionado com a diferença de ativação no córtex insular nas fases de alta e baixa média de erros. Estes resultados mostram que o processo de tomada de decisão de usuários de psicoestimulantes é menos sensível à frequência dos erros cometidos e que menor processamento neural é utilizado para antecipar os resultados da tomada de decisão. Estes achados poderiam também indicar predisposição ao uso de drogas; a avaliação do significado de sua relevância clínica, portanto, requer mais pesquisas.

Pesquisas com jogos de reciprocidade (“fairness games”), utilizando os mesmos parceiros interagindo diversas vezes seguidas, permitem estudar os mecanismos computacionais envolvidos na interação social. Correlações dos resultados comportamentais com dados de atividade neural abrem

perspectivas no estudo das computações envolvidas em intercâmbios sociais, tanto em indivíduos sadios como em pacientes portadores de transtornos neuropsiquiátricos.

Coletivamente, os jogos de reciprocidade envolvem três fatos fundamentais que estão na base das interações sociais: (1) a existência de normas sociais sobre o que é esperado em interações recíprocas entre dois indivíduos; (2) a existência de um mecanismo de detecção de violação de normas que gera um sinal que leve a uma mudança de comportamento do jogador que detectou a violação, visando alterar o comportamento do outro; e (3) uma modelagem dos processos mentais do parceiro, que permite antever sua resposta a esta mudança de comportamento.

Mesmo nas suas formas mais simples, o intercâmbio social requer algumas capacidades computacionais nos agentes envolvidos (KISHIDA *et al.*, 2010). Num simples jogo de reciprocidade com outro ser humano, cada um deve ser capaz de (1) computar normas sociais sobre o que é esperado em interações recíprocas entre dois indivíduos; (2) detectar desvio destas normas; e (3) selecionar respostas apropriadas a estes desvios.

Para estudar estes mecanismos computacionais e sua possível alterações em transtornos de personalidade, King-Casas e colaboradores (2008) estudaram a evolução da interação no Jogo da Confiança e a resposta hemodinâmica nervosa com fMRI, comparando a evolução da cooperação em díades compostas apenas por voluntários sadios e díades compostas por voluntários sadios e pacientes com transtorno *borderline* de personalidade. Portadores deste transtorno exibem um padrão de relações interpessoais intensas e muito instáveis, grandes esforços para evitar abandonos, irritabilidade e instabilidade afetiva. Como exposto anteriormente, no jogo da confiança um dos jogadores, o investidor, recebe uma quantia da qual pode enviar uma parte ou tudo ao outro jogador, o curador. A quantia enviada é automaticamente multiplicada por três ao ser passada ao curador. Este, então, dividirá com o investidor a quantia recebida. Importante ressaltar que tanto o investidor quanto o curador são livres para investir ou devolver qualquer porção das quantidades em jogo, podendo

inclusive nada investir ou nada devolver. Nesta situação, a quantidade que um jogador manda para outro sem qualquer obrigação externa quantifica a confiança; a confiança do investidor de que o curador retornará mais do que foi originalmente investido e a confiança do curador de que nas próximas jogadas o investidor continuará investindo quantidades grandes. Esta situação também pode refletir a norma presente em ambos os jogadores sobre a forma de dividir um recurso. Foi comparada a evolução da interação de díades formadas por dois indivíduos sadios com a evolução de díades formadas por um portador de personalidade *borderline* como curador e um indivíduo sadio como investidor. A diferença entre os grupos manifestou-se quando os investidores começaram a investir pequenas somas. Nesta situação os receptores sadios apresentaram maior ativação hemodinâmica do córtex insular anterior e responderam retornando maiores somas ao investidor. Esta estratégia de “sedução” sinalizava o quanto eram confiáveis e tipicamente levou a maiores investimentos por parte do parceiro. Diante das pequenas quantias investidas, as alterações hemodinâmicas dos pacientes *borderline* não foram significativas nem os mesmos adotaram a mesma estratégia dos controles sadios. Ao contrário, com maior frequência, respondiam a um investimento pequeno com um retorno também pequeno. Vários estudos implicam o córtex insular anterior na representação e atualização de normas (e.g., MONTAGUE; LOHRENZ, 2007) e os dados deste estudo sugerem que a diferença da ativação entre os grupos refletem a sensibilidade destes a violações de normas comportamentais neste jogo. As diferentes estratégias podem ser explicadas como resultantes desta sensibilidade diferente.

Os exemplos resumidos acima mostram o quanto a abordagem neuroeconômica pode contribuir para a pesquisa das bases neurais do comportamento social e dos transtornos mentais em que a interação social é comprometida e tem se mostrado resistente a abordagens quantitativas.

1.6 Aprendizagem por reforço

Uma questão fundamental no estudo do comportamento animal é a do processo de tomada de decisão envolvendo selecionar opções em face de recompensas e punições. Este tópico vem sendo investigado na psicologia em associação com os conceitos de aprendizado por reforço, seja o condicionamento Pavloviano ou o condicionamento instrumental. No primeiro o agente aprende uma relação entre um estímulo ambiental inicialmente neutro do ponto vista motivacional e um estímulo de valor motivacional, chamado de estímulo incondicionado; por exemplo, um animal aprende que um som indica ou antecede a liberação de alimento ou de uma punição como um choque elétrico. Com o aprendizado o estímulo inicialmente neutro passa ter a capacidade de desencadear reações fisiológicas típicas do estímulo incondicionado, tornando-se neste ponto um estímulo condicionado. No condicionamento instrumental o agente aprende a relação entre uma ação sua e a liberação de estímulo incondicionado. Uma diferença fundamental entre os dois tipos de aprendizagem é que no condicionamento Pavloviano (ou clássico) o agente apenas aprende relações entre estímulos no meio ambiente, devendo sofrê-las passivamente, enquanto que no condicionamento instrumental o agente tem controle sobre os estímulos futuros através de suas ações; por exemplo, um animal que determina a liberação de alimento pressionando uma alavanca. Para a neurociência existe ainda a questão da implementação neural deste aprendizado, *i.e.*, em que estruturas neurais e de que forma as informações sobre o meio ambiente são armazenadas e como, a partir destas informações, decisões adaptadas são geradas. Por outro lado, os estudos do comportamento animal forneceram aos estudiosos da inteligência artificial e robótica uma inspiração para procurar algoritmos computacionais que permitem a uma máquina adquirir uma variedade de funções a partir de um sinal de reforço. Os produtos destas pesquisas computacionais são coletivamente chamados de algoritmos de aprendizagem por reforço e têm sido utilizados numa grande gama de problemas de controle e otimização.

O campo de aprendizado por reforço computacional fornece um quadro normativo, dentro do qual o comportamento condicionado pode ser analisado e compreendido. Nele, a seleção de ações otimizadas é baseada nas previsões das futuras consequências das escolhas. Estas previsões são baseadas na experiência anterior e, à medida que a experiência se repete num determinado ambiente, estas previsões ou modelos do mundo são corrigidos ou atualizados, com base na discrepância entre expectativa e resultado efetivo da ação. Esta diferença entre expectativa e resultado da ação corresponde ao conceito de erro de previsão (“prediction error”), neste caso, erro de previsão do reforço (“reward prediction error”). Este conceito se tornou a ponte entre a teoria computacional e a pesquisa neurocientífica, quando, nos anos 1990, o grupo de Wolfram Schultz (SCHULTZ *et al.*, 1997) encontrou no sinal codificado pelos neurônios dopaminérgicos do tronco encefálico o equivalente do sinal de previsão de erro dos modelos computacionais. O uso da atualização de representações de valor, ou outras, através de sinais de erro de previsão é consistente com a pressão evolucionária de um processamento de informações econômico do ponto de vista energético. A solução encontrada pelo esquema em questão consiste basicamente no armazenamento de predições em níveis cerebrais mais elevadas e calcular em níveis cerebrais mais periféricos a diferença entre as previsões e a nova informação ambiental. Esta diferença será utilizada para a atualização da representação dos centros superiores com um gasto energético muito menor que o que seria necessário caso uma nova representação nos centros superiores tivesse que ser feita a cada mudança detectada no ambiente (SCHULTZ, 2010). Este esquema é utilizado pelo encéfalo não somente na atualização de valor de estados e opções, mas também na atualização de representações perceptivas sem relação com reforço (den OUDEN *et al.*, 2010). Para Friston (2009), a troca de mensagens de erro de previsão entre diferentes níveis hierárquicos do SNC constitui-se no princípio fundamental de funcionamento encefálico e de organização do comportamento.

Assim, antes de estudarmos a neurobiologia da tomada de decisão, passaremos a um breve resumo da teoria computacional do aprendizado por reforço.

Do ponto de vista computacional, o aprendizado da relação entre eventos no meio ambiente, o condicionamento Pavloviano, é considerado o protótipo do aprendizado de previsão. No condicionamento instrumental aparece um elemento novo que são as ações do agente. O agente aprende a relação entre suas ações e a probabilidade de reforços e/ou eventos aversivos. Do ponto de vista computacional, esta tomada de decisão visa otimizar as consequências a longo prazo das ações. Deste modo, o estudo do condicionamento instrumental pode ser considerado como uma investigação daquilo que pode ser, talvez, a mais básica forma de tomada de decisão racional. Em função disto convém realçar que o fato de ratos, macacos, pombos e outros animais fazerem esta tomada de decisão com facilidade num mundo computacionalmente complexo e em constante mutação é um fato nada trivial.

Como ressaltamos, os modelos computacionais têm sido utilizados para o estudo destes comportamentos. Todos os modelos têm em comum o uso de um sinal escalar de reforço para direcionar o aprendizado. Como vimos acima, estes modelos de aprendizado por reforço fornecem um quadro normativo dentro do qual se pode analisar e interpretar o condicionamento animal. Isto é, estes modelos fornecem previsões moleculares e molares da forma ótima de comportamento; sugerem meios pelos quais previsões e seleção de ações a nível ótimo podem ser alcançadas, e expõem as computações que devem ser realizadas a serviço destas metas. Deve-se ressaltar que estes modelos não são descritivos, mas normativos. Como modelos normativos descrevem os comportamentos do ponto de vista de sua função hipotética, isto é, como deve ser para realizar suas metas em nível ótimo. Dois fatores fazem os modelos normativos importantes. Primeiramente, porque, uma vez que na evolução animal é selecionado o mais apto, é razoável a suposição de se desenvolverem comportamentos de escolha otimizados. Em segundo lugar, as discrepâncias entre os modelos normativos e os dados observados são frequentemente muito informativas, pois podem apontar limitações no processamento de informação dos agentes ou indicar que o que está sendo otimizado não é o que assume o modelo. Nos exemplos em que o jogador punia um oponente mesmo incorrendo

em um custo para tanto, temos um exemplo de um agente estar otimizando coisa diversa da prevista pelo modelo, que seria a maximização dos ganhos financeiros na jogada.

1.6.1 Modelagem matemática da aprendizagem por reforço

Passamos agora a um estudo da modelagem matemática dos fenômenos de aprendizagem, iniciando pelo condicionamento simples ou Pavloviano. Um dos primeiros modelos matemáticos foi proposto por Rescorla e Wagner em 1972 (para revisões NIV, 2009; NIV; MONTAGUE 2009). Este modelo, que gozou de muita influência nos meios científicos, capaz de explicar uma série de fenômenos encontrados no condicionamento, tem como ideia fundamental que o aprendizado ocorre quando os eventos violam as expectativas. Matematicamente, o valor de um estímulo condicionado será modificado segundo a equação:

$$V_{\text{novo}} = V_{\text{antigo}} + \eta(\text{resultado} - \text{expectativa})$$

ou

$$V_{\text{novo}} = V_{\text{antigo}} + \eta(R - V_{\text{antigo}}) \quad \text{eq. (1)}$$

Isto significa que numa tentativa o poder preditivo de estímulo (V_{novo}) mudará, para mais ou para menos, somente se o valor do resultado, isto é, do reforço obtido (R), for maior ou menor que o esperado ou previsto (V_{antigo}). Caso seja igual, nenhuma mudança ocorrerá no valor. A intensidade de variação será dada pelo coeficiente de aprendizagem η ($0 < \eta \leq 1$). Quanto maior o valor do coeficiente de aprendizagem η , maior será a importância do resultado da última tentativa na determinação do valor do estímulo. O conceito de valor é o de expectativa de reforço e determina a intensidade da resposta do animal. Alguns pontos nessa regra de Rescorla-Wagner merecem ser ressaltados. Quando existe mais de um estímulo condicionado sinalizando o próximo resultado, o valor da variável *valor antigo* na fórmula é a soma do valor de ambos. Em outras palavras, em cada tentativa

não é considerada a relação temporal entre os estímulos, nem a relação temporal entre estímulos condicionados de um lado e estímulo incondicionado de outro. A regra trata a tentativa como um objeto temporal discreto, isto é, sem dimensão temporal interna; considera apenas a relação preditiva entre antecedentes e o estímulo condicionado. Outro ponto importante é que a regra de Rescorla-Wagner trata o estímulo condicionado e incondicionado como irredutivelmente diferentes, isto é, a regra não admite a possibilidade de que um estímulo condicionado possa condicionar outro, o chamado condicionamento secundário.

Sutton e Burto (1998) (para revisões NIV, 2009; NIV; MONTAGUE, 2009) propuseram um novo modelo matemático para o condicionamento Pavloviano. Trata-se da regra de diferença temporal (DT), que é uma extensão do modelo de Rescorla-Wagner. O modelo de DT, ao contrário do anterior, não trata o tempo como objeto matematicamente discreto, mas leva em consideração as relações temporais entre os eventos em cada tentativa. Também no modelo de DT a meta de aprendizado do agente é diferente. Este busca fazer uma estimativa dos valores dos diferentes estados ou situações, em termos de reforços ou punições futuras totais que se pode prever em cada estado. O modelo anterior, como vimos, se ocupa apenas do aprendizado em uma tentativa. Já o modelo DT tem como meta, no caso de um animal experimental percorrendo uma arena, a previsão de todos os reforços futuros que se pode obter a partir de cada ponto. Uma maneira de fazer isto seria estimar para cada localização a média do total de reforços que se obtém começando-se, diversas vezes, daquele ponto. Aqui se faz patente uma das diferenças importantes com a regra de Rescorla-Wagner: a questão da incorporação da expectativa dos reforços futuros no valor de um estado. Assim, -um animal que passa por três estados sucessivos, sendo que no primeiro um estímulo sinaliza um reforço no segundo estado, uma vez passado para o segundo estado, aí recebe novo reforço passando para um terceiro estado onde, novamente, recebe um reforço. O modelo de Rescorla-Wagner permite que no primeiro estado, com a repetição da experiência, seja aprendido o valor do estímulo como preditor do primeiro reforço, mas nada será aprendido do reforço do estado 3. Já no modelo de DT será possível aprender

que o estímulo condicionado sinaliza o próximo reforço e também a passagem (determinística neste exemplo) a um estado seguinte que comportará também um reforço; a atualização do valor de um estímulo (um estado) englobando os reforços em todos os estados futuros. Esta diferença entre os modelos se torna muito importante quando consideramos que o valor dos estados será utilizado em um modelo mais elaborado de DT para tomada de decisão. Na formalização matemática do aprendizado no modelo de DT o sinal de erro, que sinaliza surpresa ou violação de expectativa, como no modelo de Rescorla-Wagner, tem importância fundamental. Agora, no entanto, este sinal engloba também o valor dos estados futuros. Iniciamos a modelagem matemática do modelo de DT apresentando primeiramente a equação para cálculo do valor dos estados englobando o valor dos possíveis estados sucessivos aos que se fará a transição (probabilística). Assim o valor do estado “s” no momento “t” é:

$$V(s,t) = E[\gamma^0 r_t + \gamma^1 r_{t+1} + \gamma^2 r_{t+2} + \gamma^3 r_{t+3} + \dots + \gamma^n r_{t+n}] \quad \text{eq.(2)}$$

com $0 < \gamma \leq 1$. O operador $E[.]$ indica que o V_t é a expectativa matemática ou média dos valores entre colchetes referentes às diversas séries de tentativas anteriores (o modelo parte da pressuposição que agente percorreu diversas vezes a sequência de estados de $t=0$ até $t=n$). A eq. (2), apresentada na sua forma estendida, informa que o valor do estado “s” no momento “t” é igual ao valor do reforço imediatamente disponível, somado ao valor descontado por γ do reforço do estado seguinte, mais o valor do reforço do estado sucessivo descontado por γ^2 e assim sucessivamente, de maneira que a cada passo temporal o valor do reforço respectivo sofre uma desvalorização semelhante, fazendo com que os reforços mais distantes tenham valor menor que os mais recentes. É o que chamamos de desconto exponencial. Podemos reescrever a eq.(2) na forma abreviada abaixo. Com a finalidade de simplificar a notação passamos a utilizar V_t em lugar de $V(s,t)$.

$$V_t = r_t + \gamma V_{t+1} \quad \text{eq.(3)}.$$

Mas como se faz o aprendizado no DT? Como no modelo anterior o aprendizado se faz pela experiência de repetição frequente da sucessão de estados e pela correção repetida das estimativas. A cada estado corresponde uma expectativa, seu valor, calculada pela média dos reforços obtidos. Cada vez que o animal transita por este estado, esta expectativa se confirma ou é violada, tanto no que diz respeito ao reforço imediato como ao valor do estado que se segue probabilisticamente. Esta diferença entre o valor esperado e o encontrado é usada para atualizar o valor do estado.

Diferença = reforço estado atual + γ previsão próximo estado – previsão do estado atual:

$$\delta_t = r_t + \gamma V_{t+1} - V_t \quad \text{eq.(4)}$$

A equação (4) nos dá o erro “ δ_t ” que é utilizado para atualização do valor de “ V_t ”. Esta equação nada mais é que a diferença entre o esperado (V_t) e o obtido tanto como reforço imediato como transição a um novo estado ($r_t + \gamma V_{t+1}$). Todo este processo em que reiteradamente o animal passa pelos sucessivos estados deve ser compreendido como probabilístico de maneira que os valores dos estados devem ser entendidos como expectativas matemáticas ou médias. Assim, na próxima vez que o agente frequentar o estado “s” no momento “t”, o valor deste estado terá sido atualizado de acordo com as repetidas experiências anteriores. Note-se que “ δ_t ” poderá ser diferente de zero mesmo quando o estado pelo qual o animal passou não fornece nenhum reforço imediato, mas apresenta probabilidade de transição para estados promissores. Obviamente δ_t codifica a surpresa, erro de previsão ou violação de expectativa que, como no caso da regra de Rescorla-Wagner, será utilizada para a atualização do valor do estado após multiplicação por uma constante que determina, como no caso anterior, o quanto o valor será alterado em função deste sinal.

$$V_{\text{novo}} = V_{\text{antigo}} + \eta \delta_t \quad \text{eq.(5)}$$

Como no caso da regra de Rescorla-Wagner, a atualização do valor de um estado dependerá do coeficiente de aprendizagem η ($0 < \eta \leq 1$), que estabelece o quanto o resultado da experiência recente será determinante.

Até o momento estivemos trabalhando com a ideia de condicionamento Pavloviano. Porém, como nosso ponto é a tomada de decisão racional, passamos agora ao estudo do condicionamento operante. Como vimos, no condicionamento Pavloviano a transição entre os estados não depende das ações do agente. Já no condicionamento operante são as ações do agente que determinam as transições. Em função disto será necessário um modelo de como os agentes aprendem a fazer escolhas ou, em outras palavras, de como desenvolvem uma política de escolhas. Descreveremos brevemente dois modelos bastante usados em neurociência, o modelo de ator e crítico, e o modelo de aprendizagem-Q (para revisões DOYA, 2007; MAIA, 2009; NIV, 2009; NIV; MONTAGUE 2009).

No condicionamento operante a transição entre estados passa a depender das ações do agente e a meta deste passa a ser fazer escolhas que levem a estados associados à maior soma de reforços presentes e futuros. A grande dificuldade neste caso está em que cada ação determina não apenas a transição e o reforço imediato, mas a partir desta transição, também as possíveis transições subsequentes e consequentes reforços. Assim, quando o agente recebe um reforço após uma sequência de várias ações é necessário descobrir qual a ação anterior que aumentou a probabilidade de reforço. Este é problema de alocação temporal de crédito (“temporal credit assignment problem”). Formalmente, descrevemos a situação como sendo a transição probabilística de um estado a outro em função de uma ação (probabilidade de um novo estado dado o estado anterior e a ação tomada),

$$P(\text{novo estado} \mid \text{estado anterior, ação}),$$

e da probabilidade de reforço dada uma determinada ação em determinado estado,

$$P(\text{reforço} \mid \text{estado, ação}).$$

A meta do agente é melhorar sua política (“policy”) de escolha da ação em cada estado ou,

$$P(\text{ação} | \text{estado})$$

para maximizar a média dos reforços futuros. Como sempre nesta exposição, a maneira mais simples de entender um estado é compreendê-lo como um “input” sensorial (por exemplo, uma luz ou som sinalizando durante a execução de uma tarefa).

No modelo de ator e crítico, o crítico aprende, como no condicionamento Pavloviano, o valor de cada estado como descrito nas eq.(2) e eq.(3). Nesta equação obtemos a expectativa matemática (média) dos reforços futuros (desvalorizados pelo prazo) a serem obtidos deste estado seguindo a política atual (a distribuição probabilística das ações em cada estado subsequente). Como vimos na discussão anterior, a cada transição para um novo estado calculamos o δ , ou seja, o quanto este novo estado é melhor ou pior do que o previsto, em termos de reforços imediatos e futuros, e atualizamos o seu valor conforme a eq. (5). O valor de δ também será utilizado para atualizar a política, ou seja, aumentar a probabilidade de o agente escolher a ação que levou a um estado com maior valor do que anteriormente suposto, ou contrariamente diminuir a probabilidade da ação se esta levou a um estado de menor valor que o previsto. Formalmente:

$$p(\text{ação} | \text{estado})_{\text{novo}} = p(\text{ação} | \text{estado})_{\text{antigo}} + \eta \delta \quad \text{eq. (6)}$$

Outro algoritmo de aprendizagem por reforço para modelar o condicionamento operante é o chamado aprendizado-Q (“Q-learning”). Neste algoritmo, o agente aprende diretamente a encontrar a melhor política, ou seja, a melhor probabilidade de ação em cada estado sem necessidade de aprender o valor de cada estado como no modelo de ator e crítico.

$$Q(s,a) = E[r_t + \gamma^1 r_{t+1} + \gamma^2 r_{t+2} + \dots + \gamma^n r_{t+n} | \text{estado}(t)=s, \text{ação}(t)=a]$$

eq. (7)

Esta equação, onde $Q(s,a)$ é o valor médio da ação(a) no estado(s), nos diz o quanto o agente pode, em média, esperar se no estado s escolher a ação a e, a seguir, escolher sempre a ação que a experiência anterior ensinou ser a de maior retorno. Ao fazer esta escolha, o agente transita para o próximo estado e recebe ou não reforço conforme a probabilidade de transição e reforço vistas acima, bem como conhece o valor do estado subsequente. Neste ponto a diferença entre a previsão e o ocorrido será calculada, δ_t , e este será utilizado para atualizar o valor daquele par estado-ação,

$$Q(s,a): Q(s,a)_{\text{novo}} = Q(s,a)_{\text{antigo}} + \eta \delta_t \text{ eq.(8)}$$

enquanto os outros valores de Q para os outros pares de estado e ação mantêm-se inalterados. O processo de aprendizado-Q consiste na repetição de três passos: (1) fazer a previsão da expectativa de reforço das ações candidatas naquele estado; (2) selecionar a ação que tiver a maior expectativa de reforço; e (3) atualizar o valor de Q para aquele par estado-ação que foi escolhido, usando o erro ou discrepância entre valor obtido e esperado.

Deve-se ressaltar que embora o modelo de ator e crítico não apresente convergência garantida a uma política otimizada, como demonstra-se matematicamente, este modelo apresenta vínculos mais fortes com a pesquisa neurobiológica envolvendo tomada de decisão.

1.6.2 Aprendizagem por reforço e modelos do ambiente

Do ponto de vista computacional distinguem-se dois tipos de aprendizagem por reforço: (1) o aprendizado baseado em modelos do ambiente (*model based*) e (2) o aprendizado independente de modelos do ambiente (*model free*) (para uma revisão DAYAN, 2009). No primeiro tipo o agente desenvolve um modelo ou mapa cognitivo do meio ambiente e o atualiza constantemente, o que lhe permite otimizar o processo de tomada de decisão em termos de obtenção de reforços. No aprendizado

independente de modelos o que o agente aprende é, a cada estado, a expectativa de valor (o quão boa em termos de reforço total) de cada ação. Os casos discutidos acima se enquadram no aprendizado sem modelos, pois o agente não constrói um modelo de ambiente, que implicaria em descobrir e armazenar para cada estado a probabilidade de reforço e de transição para o estado seguinte, em função de cada ação. Por exemplo, no aprendizado sem modelos num labirinto, o agente aprende em cada ponto de decisão, para cada opção, quão distante, em média, está o reforço ou a saída, sem construir um mapa mental do labirinto. No caso de aprendizado baseado em modelos o agente construirá um mapa cognitivo do labirinto, que representará todos os estados e todas as probabilidades de transição entre eles, e usará este mapa para cada decisão. E após cada decisão o mapa será atualizado.

O aprendizado sem modelos tem a característica de que os valores armazenados associados a estados mais proximais não serem imediatamente sensíveis a mudanças nas contingências ou utilidades de estados mais distais. O novo valor de um estado proximal deve ser descoberto a partir de várias experiências e deve incorporar o novo valor de um estado distal conforme as equações acima; *i.e.* o novo valor de um estado distal “retrocede” gradativamente, na velocidade determinada pelo coeficiente de aprendizado η (eq. 5), para os valores dos estados proximais. No aprendizado baseado em modelos, onde as transições também podem ser probabilísticas, quando há uma modificação de uma contingência o agente modifica sua representação no mapa cognitivo e poderá utilizá-la imediatamente a partir dos estados anteriores, uma vez que, em cada estado, em cada decisão, o agente percorre mentalmente o mapa inteiro. Embora o aprendizado sem modelos exija uma prática intensa e seja bem menos sensível em caso de mudança de contingências, sua demanda computacional é menor do que a do aprendizado baseado em modelos. Neste é necessário que toda a árvore de opções seja apreendida e que o agente a percorra mentalmente a cada momento de decisão, tendo, portanto, uma exigência computacional muito maior. Se pensarmos que em agentes biológicos esta exigência computacional se traduz em demanda sobre a capacidade da memória operacional,

podemos ter uma ideia das limitações desta opção. No entanto, esta distinção é muito importante para os sistemas biológicos uma vez que reflete ou reproduz a diferença entre hábitos (associações entre estímulos e respostas) e ações orientadas a resultados (associações entre respostas e resultados, *goal directed actions*).

Do ponto de vista dos paradigmas comportamentais, os hábitos se caracterizam pela maior persistência da resposta do agente em face de degradações da contingência ou desvalorização do reforço. Já as associações resposta-resultado são muito sensíveis a estas mudanças, que rapidamente levam à diminuição da frequência ou extinção da resposta. Estes dois tipos de aprendizagem também têm bases neuroanatômicas diferentes; as ações habituais têm base no estriado dorsolateral (no rato) ou putamen (nos primatas) e regiões corticais associadas, enquanto que as associações resposta-resultado têm sua base anatômica no estriado dorsomedial (no rato) ou núcleo caudado (nos primatas) e áreas corticais conexas (DAW *et al.*, 2005).

Para a neurociência, a importância dos modelos de diferença temporal, que formalizam matematicamente o aprendizado sem modelos do ambiente, está na semelhança entre δ , o sinal de erro ou de violação de expectativa, e o sinal fásico dos neurônios dopaminérgicos do tronco encefálico. A associação entre dopamina e aprendizagem por reforço foi descoberta em meados dos anos 1990, no laboratório de Wolfram Schultz (1997). Este pesquisador registrou a atividade eletrofisiológica de neurônios dopaminérgicos de macacos durante o desempenho de tarefas simples de condicionamento instrumental ou Pavloviano e mostrou que períodos de intensa atividade dos neurônios dopaminérgicos (*bursts*) ocorriam, inicialmente, quando o animal recebia um reforço. Posteriormente, quando o animal já havia aprendido que um sinal sonoro permitia prever a liberação do reforço após determinado tempo, a referida atividade deslocava-se para o momento do sinal e não mais ocorria no momento da liberação do reforço. Como no algoritmo de aprendizagem por reforço, inicialmente, antes de qualquer aprendizado, quando não há qualquer previsão, um reforço desencadeia um sinal de erro, pois tanto V_t (o tom) como os estados posteriores V_{t+1} têm valor zero, o que leva a um valor positivo de δ_t , devido a

r_t , como se pode ver substituindo-se os valores na eq.(4) ($\delta_t = r_t + \gamma V_{t+1} - V_t$) (alternativamente podemos interpretar o reforço como V_{t+1} , o que resultará igualmente num sinal positivo, pois r_t e V_t são nulos). À medida que o tom passa a ser associado com a futura liberação de reforço, a imprevisibilidade passa a este. Nesta nova situação, o tom (agora já um estímulo condicionado) dá origem a um δ_t positivo, ou pelo reforço imediato a que está associado (r_t), ou, na ausência deste, pelo reforço futuro que sinaliza (V_{t+1}). Caso um reforço previsto não ocorra, haverá, no momento em que este era esperado, uma baixa na atividade dos neurônios dopaminérgicos, sinalizando assim um sinal negativo. Usando as equações acima teremos, nesta situação, $\delta_t > 0$ porque no momento t houve o sinal inesperado indicando reforço futuro e $\delta_{t+1} < 0$, no momento $t+1$ em que o reforço previsto não é liberado.

FIGURA 3 – COMPARAÇÃO ENTRE O MODELO ALGORÍTMICO (COLUNA DA ESQUERDA) E O REGISTRO ELETROFISIOLÓGICO DE NEURÔNIOS DOPAMINÉRGICOS (COLUNA DA DIREITA) DURANTE CONDICIONAMENTO PAVLOVIANO. OS PAINÉIS “A” E “D” ILUSTRAM O SINAL DE ERRO PARA O ESTÍMULO INCONDICIONADO NO INÍCIO DO TREINAMENTO, COM AS LINHAS PONTILHADAS MOSTRANDO A EVOLUÇÃO DESTESINAL NO QUINTO E NO DÉCIMO EPISÓDIOS DE TREINAMENTO. OS PAINÉIS “B” E “E” MOSTRAM A RESPOSTA APÓS O TREINAMENTO, AGORA DESLOCADA PARA O MOMENTO DO APARECIMENTO DO ESTÍMULO CONDICIONADO. OS PAINÉIS “C” E “F” MOSTRAM A SITUAÇÃO EM QUE UM REFORÇO ANUNCIADO POR UM ESTÍMULO CONDICIONADO NÃO OCORRE. NOS PAINÉIS “D” E “F” CADA PONTO REPRESENTA UM POTENCIAL DE AÇÃO E AS DIFERENTES LINHAS REPRESENTAM DIFERENTES EPISÓDIOS. OS HISTOGRAMAS MOSTRAM A SOMA DA ATIVIDADE REPRESENTADA ABAIXO (extraído de NIV, 2009).

A Figura 3 ilustra a atividade de neurônios dopaminérgicos durante a atividade de condicionamento (Figuras 3d, 3e e 3f) e, paralelamente, uma simulação computacional de aprendizagem por reforço usando o modelo matemático acima descrito (Figuras 3a, 3b e 3c). Na simulação a tarefa foi de condicionamento Pavloviano no qual o estímulo condicionado é seguido, dois segundos depois, pelo reforço. No registro eletrofisiológico apresentado por Schultz e col. (1997) numa tarefa análoga, o estímulo condicionado sinalizava a disponibilidade do reforço desde que o macaco fizesse o movimento de apanhá-lo. Deve-se, no entanto, notar que no momento em que um reforço esperado é omitido, temos um sinal negativo no modelo computacional, e uma pausa da atividade neuronal no registro eletrofisiológico. Estudos da atividade dos neurônios dopaminérgicos em primatas sugerem que os sinais negativos de erro são codificados pela duração da pausa da atividade destes neurônios (BAYER *et al.*, 2007).

A liberação fásica de dopamina em estruturas-alvo das projeções dopaminérgicas, tais como o núcleo *accumbens*, refletem sinais de erro correspondentes aos descritos pelos algoritmos de diferença temporal (DAY *et al.*, 2007). A maioria dos estudos humanos usando fMRI relata ativação com características de sinais de erro em áreas ricamente inervadas por fibras dopaminérgicas, como por exemplo o estriado e o córtex órbito-frontal (O'DOHERTY *et al.*, 2003; PESSIGLIONE *et al.*, 2006). Fato importante é que a presença ou ausência de atividade relacionada a sinais de erro no estriado distingue entre os que adquirem o desempenho ótimo da tarefa e os que não aprendem (SCHOENBERG *et al.*, 2007). Relevante na interpretação destes dados é o fato de que o sinal BOLD do fMRI tem uma correlação mais importante com os potenciais de campo locais (que refletem o *input* para uma área) do que com a taxa de disparos neuronais da área (que está na origem do *output* desta área).

Como referimos anteriormente, existe uma relação muito estreita entre os dados gerados pela neurociência e o modelo ator-crítico no caso de aprendizado da associação estímulo-resposta. Isto

corresponde muito de perto a um aprendizado por reforço sem modelos onde se associa a cada estado (por exemplo, estímulos ambientais) a preferência por uma ação. Em outras palavras, o modelo de ator-crítico aprende procedimentos habituais, ou associações sensório-motoras. Do ponto de vista anatômico, envolve o estriado dorsolateral, que nos primatas corresponde ao putâmem, e que é parte das alças córtico-estriato-tálamo-corticais envolvendo áreas sensório-motoras. Lesões do estriado dorsolateral em ratos levam a mudanças do comportamento governado pela associação estímulo-resposta (habitual) para comportamento governado pela associação entre a ação e a consequência (*goal directed behavior*) (YIN *et al.*, 2006a, 2006b). Este último tipo de comportamento está associado ao estriado dorsomedial e áreas corticais conexas. Lesões, em ratos, do estriado dorsomedial resultam na redução da sensibilidade do animal à desvalorização do reforço ou a degradação da contingência, características típicas do comportamento governado pela associação estímulo-resposta (YIN *et al.*, 2005). Como no modelo ator-crítico o ator aprende as associações estado-ação, é natural ligar o estriado dorsolateral ao ator. Já o crítico, que aprende o valor dos estados, deverá ser ligado a áreas que representem valor, que são áreas que têm ativação durante a expectativa do reforço, uma vez que o valor corresponde a expectativa da soma de reforços futuros. Para ter a função de crítico, uma área deve também receber e enviar projeções para o sistema dopaminérgico, de modo a poder aprender o valor dos estados e enviar este valor para treinar o ator respectivamente. O estriado ventral preenche todos estes requisitos, assim como o córtex órbito-frontal e a amígdala. Um estudo usando fMRI em humanos (O'DOHERTY *et al.*, 2004) pesquisou a separação entre ator e crítico comparando uma tarefa de condicionamento instrumental e uma tarefa de condicionamento Pavloviano paralelas. Ambas as tarefas envolviam o mesmo aprendizado de valor dos estados, mas apenas a tarefa de condicionamento instrumental envolvia o aprendizado de seleção de ações. O sinal BOLD no estriado ventral estava correlacionado a sinais de erro em ambas as tarefas, mas no estriado dorsal esta correlação existia apenas na tarefa instrumental. Estes resultados são consistentes com a projeção da divisão ator-crítico na divisão ventral / dorsal do estriado, pois em ambas as tarefas havia aprendizado

de valor de estados, mas apenas na tarefa instrumental havia aprendido do valor de pares de estado-ação. No entanto, outros estudos também dão suporte a modelos de aprendizagem-Q, onde não há necessidade de aprender valor de estados e a partir disto aprender o valor dos pares estado-ação, podendo estes últimos ser aprendidos diretamente (ROECH *et al.*, 2007).

Antes de finalizar este breve resumo, convém discutir algumas limitações no uso dos modelos computacionais de aprendizagem por reforço em neurociência (DAYAN; NIV, 2008; MAIA, 2009). Entre estas dificuldades está o fato de que os modelos de aprendizagem por reforço utilizam a desvalorização exponencial, o fator γ que aplica a mesma desvalorização a cada intervalo de tempo. No entanto, a pesquisa experimental tem mostrado que humanos, macacos, ratos e pombos usam a desvalorização hiperbólica. Como na desvalorização hiperbólica o fator de desvalorização não é constante, mas é uma função do tempo decorrido até o momento do reforço, um modelo que utilizasse desvalorização hiperbólica não poderia calcular o sinal de erro no momento necessário. Embora existam modelos que propõem soluções, ainda estamos longe de resolvê-lo. Outra dificuldade importante diz respeito ao encadeamento temporal rígido dos modelos de diferença temporal discutidos. Todo o sistema funciona baseado em previsões precisas de tempo para liberação do reforço, que são determinantes para o sinal de erro. Embora no laboratório possamos trabalhar com intervalos precisos, nos ecossistemas onde naturalmente vivem e aprendem os animais não podem contar com esta precisão. Além do mais, esta envolve várias representações discretas do estímulo até o momento do reforço que não parecem ter equivalente no tecido neural (SCHULTZ *et al.*, 1997). Para lidar com este problema O'Reilly e colaboradores (2007) propuseram o modelo que chamaram de valor primitivo valor apreendido (*primary value learned value*), em que através de um mecanismo mais complexo e fisiologicamente mais verossímil o sistema dopaminérgico torna-se robusto ao ruído temporal. Outra limitação importante diz respeito à alternância entre explorar uma única alternativa e fazer a exploração do meio ambiente para melhor amostragem das opções (*exploration versus exploitation*), fenômeno extremamente importante no comportamento animal. O modelo acima

apresentado, em que há o desenvolvimento gradual da associação entre um estímulo e uma ação, funciona em ambientes estacionários, nos quais a exploração de uma única alternativa é uma política eficiente. No entanto, em ambientes não estacionários este tipo de política é problemática, pois aí a melhor estratégia é a do comportamento exploratório que permita a descoberta e aprendizado de novas e melhores contingências de reforço. A pesquisa neurobiológica experimental, associada a estudos de modelagem, mostra que as variações da resposta tônica e fásica dos neurônios noradrenérgicos, situados no *locus coeruleus* (LC), quando associadas a um mecanismo de aprendizagem por reforço, são capazes de implementar um mecanismo de alternância de “exploration versus exploitation” conforme as condições de reforço. Numa tarefa de reversão de aprendizado, após a aquisição da associação inicial, quando há uma certeza sobre a estrutura do ambiente e consequentemente uma alta frequência de reforços, o LC funciona no modo fásico. Quando as contingências são invertidas e o nível de reforços cai, refletindo a incerteza sobre a estrutura do ambiente, a atividade tônica do LC aumenta e a atividade fásica diminui. Permanece assim até a nova associação ser adquirida, quando então a atividade do LC volta ao estado inicial (COHEN *et al.*, 2007). O mecanismo proposto envolve o ganho (coeficiente angular da função de ativação) dos neurônios das regiões frontais. Baixa atividade dos neurônios do LC (modo tônico), e consequentemente baixo ganho da ativação do córtex pré-frontal, é associada à ativação de muitas unidades de resposta, o que favorece o comportamento de exploração das diversas alternativas. A passagem para o modo fásico está associada ao fenômeno inverso, com a consequência de diminuir este comportamento (USHER *et al.*, 1999).

1.6.3 Modelos do circuito córtico-estriato-tálamo-cortical e a modulação dopaminérgica

Na aprendizagem por reforço mencionamos a importância da modulação dopaminérgica do estriado e córtex neste processo. Agora passaremos a um estudo mais detalhado das estruturas nervosas envolvidas e de como a sua fisiologia é descrita na literatura.

Franck e col. (HAZY *et al.*, 2007; COHEN; FRANK, 2009) propuseram um modelo de alças distintas parcialmente sobrepostas, que transmitem *inputs* excitatórios oriundos do córtex para o estriado onde fazem sinapse com neurônios espinhosos médios (*medium spiny neurons*, MSN) (Figura 4). Estes neurônios inibitórios (GABAérgicos) dividem-se em dois grupos que fazem projeções distintas para os núcleos de saída do circuito e, desta forma, exercem ações diferentes.

MSN que expressam receptores D1 formam a via direta ou facilitatória (*GO pathway*), monossináptica, que inibe diretamente os núcleos de saída do sistema, *i.e.*, o globo pálido interno (GPi) e a substância negra pars reticulada (SNr). Estas estruturas formadas por neurônios GABAérgicos tonicamente ativos projetam-se sobre a região correspondente do tálamo que, assim, é mantida inibida. O tálamo tem um sistema de retroalimentação positiva com o córtex, isto é, a atividade cortical é transmitida ao tálamo e de lá retorna ao córtex sustentando sua atividade. Desta maneira, um impulso excitatório que parte do córtex, passando pela via direta, inibirá o GPi/SNr, que consequentemente liberará de sua inibição tônica a parte do tálamo a que se projeta. Assim permitirá, nesta parte, a retroalimentação positiva com a área cortical com que está conectada.

FIGURA 4 – MODELO ESQUEMÁTICO DO SISTEMA CÓRTICO-ESTRIATO-TÁLAMO-CORTICAL

(extraído de MAIA; FRANCK, 2011).

MSN que expressam receptores dopaminérgicos D2 formam a via indireta (*NOGO pathway*), que, quando ativa, tem ação inibitória no processo de seleção de ação. Os MSN desta via não fazem sinapse diretamente no GPi/SNr, mas projetam-se para o globo pálido externo (GPe) levando à sua inibição. O GPe, tonicamente ativo, se projeta sobre as estruturas de saída do sistema, o GPi/SNr, mantendo-o inibido. O GPi/SNr é composto de neurônios GABAérgicos, também tonicamente ativos, que se projeta sobre o tálamo, ao ser inibido pelo GPe, libera a região talâmica correspondente para a alça de retroalimentação positiva com o córtex. Assim, quando um impulso proveniente do córtex transita pela via indireta há uma inibição do GPe, que então deixa de inibir a região correspondente do GPi/SNr. Esta região do GPe/SNr, não mais recebendo inibição do GPe, conseqüentemente inibe a região talâmica a que se projeta, e assim impede o funcionamento da alça de retroalimentação positiva com a região cortical correspondente. Além desta projeção, o GPe faz uma alça de retroalimentação com o núcleo subtalâmico (STN).

O STN envia projeções excitatórias difusas (não representadas na Figura 4) sobre o GPi/SNr, que equivalem a um sinal inibitório geral que dificulta ou impede qualquer ação. O STN recebe também projeções diretas do córtex formando a chamada via hiperdireta. Esta via hiperdireta, que leva à inibição generalizada do GPi/SNr, tem a função de prevenir a emissão de respostas quando há conflito, no córtex, entre as várias ações de valor semelhante competindo por expressar-se (o que portanto leva a uma atividade cortical maior) (FRANK, 2006). Convém ressaltar que enquanto a inibição da estrutura de saída do sistema pela via indireta (GPe) é focal, a inibição feita pela via hiperdireta é difusa. Desta maneira a via indireta inibe uma resposta específica enquanto que a via hiperdireta inibe todas as respostas.

Resumidamente, a via direta é excitatória, facilitando o impulso cortical que por ela transita e a via indireta é inibitória, tendo o efeito oposto. De capital importância para o funcionamento deste

sistema de seleção de ações é a modulação dopaminérgica. Os neurônios dopaminérgicos cujos corpos estão no tronco encefálico (área tegmental ventral – VTA e substância negra *pars compacta* – SNc) projetam-se para o estriado e córtex pré-frontal. No estriado, a ação da dopamina nas células que expressam receptores D1 aumenta o contraste, isto é, aumenta a atividade de células mais ativas enquanto diminuem a atividade de células pouco ativas. Isto tem como efeito geral a amplificação da relação sinal/ruído. Nos MSN os receptores D2 a dopamina é inibitória, qualquer que seja o nível de atividade da célula. Devido à expressão em grande quantidade de receptores D1 na via direta, o aumento da concentração dopaminérgica tem o efeito de facilitar a transmissão do sinal dos grupamentos neuronais corticais mais ativos através da alça córtico-estriado-tálamo-cortical correspondente. Ao mesmo tempo, pela ativação dos receptores D2, este aumento da concentração dopaminérgica inibe a transmissão do sinal através da via indireta. Uma queda na concentração de dopamina tem o efeito oposto. Desta maneira a ação considerada no córtex quando da queda de concentração dopaminérgica será inibida ou enfraquecida. A cada momento há um equilíbrio entre as duas vias que é determinado pela concentração dopaminérgica. Além deste efeito instantâneo na ativação dos MSN, a dopamina apresenta um importante efeito sobre a plasticidade sináptica, resultando em aprendizado. Quando um neurônio pré-sináptico e um pós-sináptico estão simultaneamente ativos há um aumento da intensidade da conexão sináptica entre estas duas células, fenômeno conhecido como lei de Hebb. Este aumento da conexão sináptica é potencializado pela dopamina resultando em potenciação de longa duração (LTP). Quando ocorre o oposto as conexões não mudam ou são enfraquecidas, resultando em depressão de longa duração (LTD). Na realidade o mecanismo primário de aprendizado nos gânglios basais ocorre via modulação dopaminérgica de células ativadas por *input* excitatório glutamatérgico córtico-estriatal. Com este processo, a ação ou representação cujas conexões com os MSN correspondentes sofreram LTP, na via excitatória ou na inibitória, terão sua execução respectivamente facilitada ou inibida de forma definitiva. A ação da dopamina no córtex será discutida posteriormente.

As descargas fásicas dos neurônios dopaminérgicos ocorrem quando diante de reforços inesperados ou da percepção de estímulos condicionados que anunciam reforços (para revisão SCHULTZ, 2010). Assim, este sistema tem importante função no aprendizado da escolha de ações associadas a reforços ou à percepção de estímulos condicionados e, inversamente, no aprendizado da não execução daquelas associadas a punições ou a ausência de reforços esperados. Uma vez ocorrido o aprendizado, nos núcleos da base, da facilitação das ações adaptativas em resposta a estímulos condicionados, a conexão das representações de ambos a nível cortical são fortalecidas (*e.g.*, entre as áreas sensoriais e áreas pré-motoras) em função da repetição da atividade concomitante de ambos através de um processo de aprendizado hebbiano. Processo oposto ocorre com ações não adaptativas. Isto é consistente com a observação que a integridade das projeções dopaminérgicas nos núcleos basais é essencial para aquisição, mas para a execução de respostas instrumentais já aprendidas (HAZY *et al.*, 2007).

Uma vez que os núcleos da base participam na seleção das respostas motoras simples, mas recebem e enviam projeções para áreas relacionadas à atividade cognitiva, é possível estender a arquitetura descrita acima para decisões não motoras. Podemos aplicar este esquema às representações da memória operacional, tanto no caso da decisão sobre as representações que aí entram (*gating*) para serem objeto de computações, como no caso da liberação de representações que irão influenciar outras computações em outras regiões encefálicas, por exemplo, outra representação de memória operacional, influência descendente (*top-down*) da percepção ou decisões motoras.

A Figura 5 apresenta uma versão ampliada do modelo que estamos discutindo. Nela está representada a separação entre o córtex frontal e regiões corticais posteriores, que se comunicam entre si por conexões córtico-corticais e também por meio de projeções para a mesma área do estriado dorsal. Com este modelo ampliado, que é baseado em evidências anatômicas e fisiológicas, podemos discutir melhor o aprendizado de associações sensório-motoras e de representações de memória operacional.

FIGURA 5 – MODELO ESPECIFICANDO COM CÓRTEX FRONTAL E POSTERIOR E CONEXÕES

(extraído de HAZY *et al.*, 2007).

Este aspecto da fisiologia requer ainda um mecanismo, *output gating*, não representado na Figura 5, que controle quando uma representação ativamente mantida no córtex pré-frontal (PFC) é enviada para outra região cortical. Este mecanismo, o *output gating*, é implementado por meio da arquitetura colunar da estrutura laminar frontal cortical e sua conectividade específica com os núcleos da base e tálamo (BROWN *et al.*, 2004). Resumidamente, as lâminas que dão origem aos *outputs* do PFC, especialmente a lâmina Vb, não emitem um sinal até que um limiar seja atingido. Em efeito este mecanismo não é diferente do mecanismo de *gating* motor. Este modelo, em suas diversas versões, quando implementado em redes neurais, foi capaz de reproduzir dados empíricos, tanto referentes a estados patológicos, tais como doença de Parkinson e transtorno de déficit de atenção e hiperatividade (THDA), como manipulações farmacológicas em indivíduos sadios (FRANK *et al.*, 2007b).

Rougier e colegas. (2005) desenvolveram adicionalmente este modelo incluindo a manutenção temporal de padrões de atividade neural no córtex pré-frontal, dependente da bi-estabilidade dos neurônios pré-frontais e da sua conectividade excitatória recorrente. A manutenção temporal está associada ao mecanismo de *gating* implementado pelos núcleos da base e projeções dopaminérgicas diretas ao córtex pré-frontal, ambos funcionando num esquema de aprendizado por reforço. O aumento da dopamina no córtex pré-frontal estabiliza as representações ativas enquanto sua

diminuição as desestabiliza e assim permite a emergência de novos padrões. As projeções pré-frontais modulam a atividade em diversas outras áreas corticais. O modelo incluindo estes mecanismos quando treinado em diversas tarefas desenvolveu representações similares a regras abstratas que são passíveis de generalização flexível para novas tarefas. No entanto, é importante ressaltar que a memória operacional não se limita apenas à manutenção de representações (CURTIS; LEE 2010).

O modelo de memória operacional envolvendo os núcleos basais e córtex pré-frontal (PBWM) deve ser capaz de seis operações básicas para poder executar essas tarefas:

1. rápida atualização de conteúdo;
2. manutenção de conteúdo em relação a distratores;
3. manter representações múltiplas e separadas;
4. atualização seletiva das diversas representações;
5. “output gating” independente de cada representação; e
6. apreender quando fazer a atualização da representação e quando emitir uma resposta.

Estas seis características podem ser melhor compreendidas se examinadas à luz de uma tarefa de memória operacional. Para isto analisaremos em maior detalhe a tarefa 1-2-AX (FRANK *et al.*, 2001; O'REILLY; FRANK, 2006; HAZY *et al.*, 2006), que é uma derivação da tarefa AX-CPT extensivamente usada em estudos com humanos. Na tarefa 1-2-AX apresenta-se uma sequência de letras, uma por vez, em sucessão (*e.g.*, A, X, B, Y) e solicita-se ao voluntário que detecte uma subsequência específica, por exemplo, um “A” seguido imediatamente de um “X”; o voluntário deve sinalizar a detecção por meio de um ato motor, *e.g.*, pressionar uma determinada tecla. As outras sequências possíveis devem ser sinalizadas por meio de outro ato motor, *e.g.*, pressionar outra tecla. O

desempenho desta tarefa requer uma forma simples de memória operacional, *i.e.*, a manutenção de uma letra até a apresentação da próxima (neste caso “A” e “X”); depois de emitida a resposta, essas informações deixam de ser relevantes. A sequência a ser detectada pode ser modificada por instrução prévia, fornecida por meio de números que aparecem na sequência de estímulos. No exemplo apresentado na Figura 6, o número 1 sinaliza que deve-se responder positivamente ao par “A e X”, ao passo que o número 2 sinaliza que deve-se responder positivamente ao par “B e Y”. Assim, nesta versão da tarefa há uma alça exterior (número apresentado) a ser mantida na memória operacional que determina qual é a sequência de letras, a alça interior, a ser detectada.

FIGURA 6 – REPRESENTAÇÃO DA TAREFA 1-2-AX (extraído de HAZY *et al.*, 2007).

Quando um estímulo é apresentado ele deve ser registrado e mantido na memória operacional durante o período necessário. O estímulo da alça externa (1 ou 2) deve ser mantido na memória até que outro da mesma categoria seja fornecido. Já os estímulos da alça interna (A, B, X, Y) devem ser atualizados a cada novo estímulo. Isto exige diferentes *loci* para representações de níveis diferentes, bem como um mecanismo responsável pelo *gating* do *input* e do *output* em condições que permitam o desempenho da tarefa. No esquema acima, quando o estímulo 2 aparece, uma atualização

da representação da alça externa deve ocorrer. Esta representação deve ser mantida concomitantemente à atualização por estímulos pertencentes à alça interna, e vice-versa. Assim, as operações 3 e 4 descritas acima, isto é, representações múltiplas separadas (uma para cada alça) e a atualização seletiva (da alça externa ou interna conforme o estímulo) são requeridas. Há uma influência mútua entre essas representações; isto é, enquanto uma representação da alça externa, 1 por exemplo, determina que é AX e não BY a sequência alvo, a apresentação de um A determina que o estímulo que leva a resposta é X e não Y. Este processo de controle de saída é dependente tanto dos sucessivos *inputs* quanto destes atingirem um determinado limiar para a ação (por exemplo, de duração e intensidade do estímulo que permita a detecção ou certeza).

As predições deste modelo foram testadas tanto em humanos sadios como em portadores de patologias envolvendo o sistema dopaminérgico tais como doença de Parkinson e transtorno de déficit de atenção e hiperatividade (Figura 7). Também a simulação do modelo em redes neurais produziu resultados semelhantes aos obtidos nas pesquisas com animais e populações humanas doentes ou sadias (para revisões HAZY *et al.*, 2007; FRANK *et al.*, 2007c; MAIA; FRANK, 2011). Na doença de Parkinson, caracterizada por degeneração dos neurônios dopaminérgicos da substância negra, há uma diminuição importante da dopamina, principalmente no estriado. O tratamento consiste em agentes dopaminérgicos, ou precursores da dopamina como levodopa ou agonistas dopaminérgicos D2.

FIGURA 7 – (A) TAREFA DE APRENDIZADO POR REFORÇO PROBABILÍSTICO. TREINAMENTO COM OS PARES MOSTRADOS EM “TRAIN” SENDO QUE OS NÚMEROS ENTRE PARÊNTESES INDICAM A PROBABILIDADE DE REFORÇO/NÃO REFORÇO. DURANTE A FASE DE TESTE, “TEST”, TODAS AS COMBINAÇÕES DE ESTÍMULOS SÃO APRESENTADAS SEM FEEDBACK PARA ESCOLHA, PARA TESTAR O APRENDIZADO, QUE SE MANIFESTARÁ PELA ESCOLHA DE A (MAIOR PROBABILIDADE DE REFORÇO) E NÃO ESCOLHA DE B (MENOR PROBABILIDADE DE REFORÇO) NOS DIVERSOS PARES NOVOS APRESENTADOS. (B) SIMULAÇÃO DA DOENÇA DE PARKINSON EM REDES NEURAIS. (C) E (D) ESTUDOS COM PACIENTES PARKINSONIANOS, SENDO QUE NO PRIMEIRO HÁ DIFERENÇAS IMPORTANTES EM RELAÇÃO AOS CONTROLES TANTO NO APRENDIZADO POSITIVO QUANTO NEGATIVO, JÁ NO SEGUNDO ESTUDO O RESULTADO APARECE APENAS PARA O APRENDIZADO NEGATIVO. (E) EFEITO NO DESEMPENHO DA TAREFA EM ESTUDANTES UNIVERSITÁRIOS SADIOS APÓS PLACEBO, CABERGOLINE E HALOPERIDOL, QUE AUMENTAM E DIMINUEM A CONCENTRAÇÃO DOPAMINÉRGICA NO ESTRIADO VIA RECEPTORES D2 PRÉ-SINÁPTICOS. (F) DESEMPENHO DE ADULTOS DO GRUPO CONTROLE E PORTADORES DE THDA EM USO OU ABSTENDO-SE DA MEDICAÇÃO (extraído de COHEN; FRANK, 2009).

O modelo prediz que as diversas alterações dopaminérgicas presentes nos pacientes tanto durante o uso de medicação como durante os períodos em que se abstêm destas se refletirão em alterações na tomada de decisão baseada em aprendizado por reforço. Quando pacientes parkinsonianos abstêm-se da medicação, seu aprendizado por reforço negativo (ausência de reforço)

deve ser melhor que seu aprendizado por reforço positivo (presença de reforço), porque os baixos níveis de dopamina levarão a uma ativação maior da via indireta e, ao mesmo tempo, dificultarão a ativação da via direta. Na situação em que estes pacientes estão fazendo uso da medicação, a situação se inverte com o aumento crônico da dopamina pré-sináptica disponível. A este quadro se deve acrescentar que a administração crônica de levodopa, que é a principal medicação usada no tratamento desta condição, aumenta as descargas fásicas de dopamina e a expressão de zif-268, um gene de expressão imediata (*immediate early gene*) ligado à plasticidade sináptica na via direta, mas não na via indireta (CARTA *et al.*, 2005). Assim, o modelo faz a predição de que a medicação aumenta a retroalimentação positiva na via direta, mas dificultará o aprendizado negativo na via indireta.

Entre as várias confirmações experimentais desse modelo citamos o estudo feito com pacientes parkinsonianos, testados nas condições de uso e abstinência da medicação, em uma tarefa de aprendizado por reforço probabilístico (Figura 7(a)) na qual algumas escolhas estavam associadas a maiores probabilidades de reforço positivo ou reforço negativo. Num dos estudos (FRANCK *et al.*, 2004) os pacientes que se abstiveram da medicação aprenderam melhor as alternativas associadas ao retorno negativo, enquanto que pacientes em uso da medicação apresentaram o padrão oposto (Figura 7c), embora em estudo semelhante (FRANCK *et al.*, 2007a) tenha sido encontrado apenas o prejuízo no aprendizado na condição de retroalimentação negativa nos pacientes em uso de medicação (Figura 7d). Quando se fez a simulação do aumento e diminuição do nível de dopamina na implementação do modelo em redes neurais, os resultados encontrados se assemelham muito aos resultados experimentais (Figura 7a). Resultados semelhantes foram obtidos com estudantes universitários jovens e saudáveis (Figura 7(e)) que foram testados na mesma tarefa nas condições de uso de baixas doses de agonistas e antagonistas dopaminérgicos (FRANK; O'REILLY, 2006). Neste estudo o antagonista D2, utilizado em dose baixa, teve seu efeito sobre os receptores pré-sinápticos levando a um aumento da liberação de dopamina. O agonista D2 em dose baixa tivera o efeito inverso. Na mesma linha foram os resultados em populações de adultos portadores de THDA (Figura 7(f)) em

tratamento com psicoestimulantes (FRANK *et al.*, 2007b), que bloqueiam o transportador de dopamina, levando a um aumento desta no estriado.

O modelo proposto por Houk (2005; HOUK *et al.*, 2007) é semelhante em sua fisiologia ao modelo do grupo de Franck (HAZY *et al.*, 2007; COHEN; FRANK, 2009) discutido acima; no entanto, o modelo inclui alças ligando o córtex cerebral aos núcleos e córtex cerebelares. Neste modelo, enquanto o sistema córtico-estriato-tálamo-cortical participa da seleção da ação ou representação de memória operacional com base em sua propensão ao reforço, as conexões cerebelares seriam responsáveis pela “amplificação e escultura” do padrão selecionado pelo sistema córtico-estriatal.

No modelo proposto por Gurney e colaboradores (2001), os núcleos da base são centrais na seleção da ação servindo como um árbitro entre as várias possibilidades, baseando a seleção de uma ação na saliência ou propensão à execução a ela associada. Esta informação, proveniente do córtex, transita pela via direta que causa uma inibição focal nos núcleos de saída do sistema, GPi/SNr. A excitação difusa destes gerada pelo STN, associada ao foco de inibição transmitido pela via direta, constitui uma rede funcionando em “centro-off /periferia-on” capaz de fazer a seleção. O equilíbrio entre estes dois canais seria garantido pela inibição transmitida pelo GPe (oriunda dos MSN que expressam receptores D2) sobre o STN e sobre o GPi/SNr. A função da modulação dopaminérgica neste modelo não está ligada ao reforço, mas à ocorrência de estímulos salientes no meio ambiente e teria como função central a identificação de quais aspectos do comportamento são importantes para causar eventos não previstos. O aprendizado no sistema seria o da relação entre ações do animal e eventos salientes no meio ambiente, a chamada descoberta de agência. Com este aprendizado o animal poderia posteriormente, a partir de computações a nível cortical, compor sequências comportamentais visando obtenção de reforço bem como evitar punições (PRESCOTT *et al.*, 2006; REDGRAVE *et al.*, 2008).

Para Cisek (2010), que estuda fundamentalmente a relação entre o sistema visual dorsal e tomada de decisão, o foco está fundamentalmente nas computações realizadas na via dorsal e suas conexões com o lobo frontal, sendo estas computações moduladas por outras estruturas, entre as quais os núcleos da base. O ponto central da teoria está na crítica ao modelo de representação herdado pela neurociência cognitiva que postula que uma decisão começa com uma representação sensorial de um estado do mundo, que é em seguida submetida a computações cognitivas que extraem sua estrutura. A partir daí são geradas opções de ação que são submetidas a um exame para escolha das potenciais ações, das quais a melhor é em seguida executada. Neste esquema, cada fase teria uma representação neural distinta, ou seja, trata-se de um modelo serial de tomada de decisão. A crítica de Cisek (2010) parte do princípio de que entre a área visual V1 e a área motora M1 a maioria das representações intermediárias não se adapta à classificação das categorias acima descritas, mas são ambíguas no sentido de que têm características de representação espacial, de valor associado a um movimento, como de diversas opções de parâmetros para um movimento. Na interação constante com o meio ambiente, o modelo serial não seria o melhor para descrever o processo de escolha da ação de um animal. No meio ambiente há uma constante competição entre as oportunidades disponíveis e demandas para agir. Assim, argumenta Cisek (2010), a arquitetura funcional básica do sistema nervoso evoluiu para mediar a interação em tempo real com o mundo, que exige que os animais continuamente especifiquem potenciais ações e façam escolham entre elas à medida que as condições do ambiente mudam. Neste contexto, propõe que o sistema nervoso processa informação sensorial para especificar, em paralelo, várias ações potenciais disponíveis naquele momento. Estas competem entre si em vista de processamento adicional, enquanto que informações adicionais são coletadas para influenciar esta competição, até que uma única resposta seja selecionada. Deste modo o sistema visual dorsal estaria na origem da especificação de várias ações que competem entre si no córtex fronto-parietal. Esta competição está sob influência de estruturas subcorticais como os gânglios da base e amígdala. É por estarem tão intimamente envolvidos neste processo, que os neurônios desta

região apresentam características tanto sensoriais, quanto de parâmetros motores e de valor. A importância deste modelo reside no questionamento do processamento serial na tomada de decisão e, consequentemente, na natureza das representações corticais envolvidas. É importante ressaltar que este modelo, ao ampliar nossa compreensão das representações corticais no processo de tomada de decisão, continua consistente com o modelo de Frank e colegas (para revisão mais recente COHEN; FRANK, 2009) e ao mesmo que o enriquece.

1.7 “Matching pennies”

O *matching pennies* é um jogo estratégico do tipo soma zero, pois em todas as situações possíveis neste jogo a soma dos ganhos e perdas de ambos os jogadores é sempre zero. Nele, dois jogadores escolhem simultaneamente uma das faces de uma moeda que deve ser colocada sobre a mesa com a face escolhida voltada para cima. Um dos jogadores, o *matcher*, ganha se as duas moedas exibirem a mesma face; o outro jogador, o *no-matcher*, ganha se as faces forem diferentes. O equilíbrio de Nash para este jogo corresponde a cada jogador produzir uma sequência de escolhas imprevisíveis, isto é, uma sequência em que a história anterior de escolhas e reforços não forneça qualquer informação sobre a probabilidade da sua próxima escolha. Em outras palavras, a sequência de escolhas deve ser, além de equiprovável, aleatória ou estocástica. No entanto, estudos com primatas humanos e não-humanos mostram que os jogadores frequentemente se afastam do equilíbrio de Nash apresentando correlação entre as escolhas e seus resultados. A provável causa é que para descobrir a estratégia mais eficiente durante o jogo, os jogadores estariam utilizando um algoritmo de aprendizagem por reforço. Como vimos anteriormente, neste tipo algoritmo de aprendizagem a escolha é feita, estocasticamente, baseada na expectativa de valor das alternativas; a alternativa que apresenta maior valor tem maior probabilidade de ser escolhida e, tanto maior a diferença, tanto maior será a

probabilidade da escolha da alternativa de maior valor. Uma vez feita a escolha, a alternativa escolhida tem seu valor atualizado proporcionalmente à diferença entre o reforço obtido e o reforço esperado. Evidentemente, um agente que utilize um algoritmo desta natureza apresentará desvio da estratégia normativa deste jogo segundo a teoria dos jogos. Podemos esperar que haja um desvio da estocasticidade normativa em direção da relação das escolhas com as escolhas prévias e seus resultados (para uma revisão LEE; SEO, 2007).

Barracough e colaboradores (2004) avaliaram o desempenho comportamental e a atividade neural do córtex pré-frontal dorsolateral de 3 macacos (*Macaca mulatta*) numa versão visual sacádica do *matching pennies*. Os animais tiveram, inicialmente, extensivo treinamento na tarefa. Cada jogada (Figura 8) iniciava com o animal fixando o olhar num quadrado amarelo que aparecia num monitor de computador por 0.5 segundo (*fore-period*). Após este período, dois círculos verdes idênticos eram apresentados, um de cada lado do quadrado amarelo na linha horizontal, com a mesma excentricidade, iniciando então o período de espera (*delay*), durante o qual o animal devia continuar fixando o quadrado central até que este desaparecesse; este período de espera também tinha duração de 0.5 segundo. Após o desaparecimento do quadrado amarelo o animal devia fazer a sua escolha, *i.e.*, fazer uma sacada (*move-hold*) para um dos círculos verdes laterais dentro de 1 segundo. Neste momento, o computador-oponente, que havia feito sua escolha no início da jogada, revelava-a (*feedback*) com a apresentação de um anel vermelho em torno do círculo correspondente ao lado escolhido. Se o animal houvesse escolhido o mesmo lado que o computador-oponente, então recebia uma gota de suco de frutas 0,5 segundo depois.

FIGURA 8 – REPRESENTAÇÃO ESQUEMÁTICA DA TAREFA ENVOLVENDO O JOGO “MATCHING PENNIES” UTILIZADA NO EXPERIMENTO DE BARRACLOUGH *et al.* (2004).

Os animais foram testados tendo o computador-oponente programado para utilizar sucessivamente três algoritmos diferentes para fazer sua escolha. No primeiro algoritmo, cada animal era testado enquanto o computador-oponente fazia a sua escolha utilizando a estratégia de equilíbrio de Nash para este jogo, *i.e.*, escolhia de forma aleatória e com igual frequência cada um dos lados. Esta estratégia de escolha, denominada algoritmo zero, foi usada para determinar o viés intrínseco de escolha que seria característico de cada animal. Vale ressaltar que, enquanto jogava contra este algoritmo, o animal obteria sempre a mesma probabilidade de reforço, isto é, 50%, independentemente da estratégia que utilizasse. Durante os testes com algoritmo zero houve um viés nas escolhas característico de cada animal, *i.e.*, a frequência de escolhas do lado direito foi de 70% para o animal “C”, 90,2% para o animal “E” e 33% para o animal “F”. Independentemente do viés nas escolhas, os animais receberam frequências de reforço semelhantes.

No estágio seguinte do experimento, os animais jogaram contra o computador-oponente que utilizava para fazer sua escolha o algoritmo 1. Neste algoritmo o computador mantinha na memória todas as escolhas feitas pelo animal naquela sessão e testava se havia algum viés sistemático na sequência de escolhas, isto sem considerar a obtenção de reforço. Por exemplo, o computador testava se após uma sequência de escolhas direita-direita-esquerda havia probabilidade de uma escolha à

direita que diferisse estatisticamente de 50%. Caso isto ocorresse, o computador escolheria a esquerda com a mesma probabilidade. Devemos ressaltar que no algoritmo 1, o fato de uma escolha do animal ter sido ou não reforçada não era considerado pelo computador-oponente no momento de computar sua escolha, de modo que qualquer influência que o reforço prévio tivesse na sequência de escolhas não era detectado pelo computador-oponente; conseqüentemente, essa influência não levava a uma diminuição da probabilidade de reforço. Nesta fase os animais tiveram uma sequência de escolhas muito mais próxima da equiprobabilidade. A frequência de escolha da direita foi de 48,9%, 49 e 51,1%. No entanto, houve um aumento da probabilidade do animal repetir o lado escolhido na jogada seguinte, se esta resultou em reforço ou, contrariamente, escolher o lado oposto, quando não obteve reforço. Esta estratégia conhecida como “ganhou-repete-perdeu-muda” (em inglês, *win-stay-lose-switch* – WSLS) foi usada com as frequências de 64.3%, 63,3% e 73.1% pelos macacos “C”, “E” e “F”, respectivamente. O uso desta estratégia aumentou gradualmente durante as sessões em que os animais jogavam contra o algoritmo 1, mesmo não havendo um aumento da frequência de reforço, o que mostra que esta estratégia corresponde a uma tendência espontânea. Além disto, o fato de ter havido um aumento do uso desta estratégia no curso do jogo contra o algoritmo 1 mostra que no início os animais produziram sequências exploratórias. Isto pode ser decorrência do fato que gerar sequências exploratórias e independentes das escolhas e reforços anteriores pode ser metabolicamente custoso.

Adotar a estratégia WSLS contra um oponente que registra tanto a escolha como o resultado desta escolha pode ser bastante desvantajoso, uma vez que o oponente utilizará esta informação em seu próprio benefício; neste caso, há uma queda na frequência de reforços para quem a utiliza. Para avaliar se os animais eram capazes de diminuir o uso da estratégia WSLS, o computador-oponente foi programado para computar os efeitos dos reforços nas escolhas dos animais. Por exemplo, o computador testava se após uma sequência de direita(+)-direita(-)-esquerda(+) (onde (+) e (-) representam escolha reforçada ou não, respectivamente) havia probabilidade significativamente

diferente de 50% de uma escolha à direita. Neste caso, o computador escolhia a esquerda com a mesma probabilidade. Esta nova estratégia de escolha do computador é chamada de algoritmo 2. Diante do algoritmo 2 o comportamento dos animais apresentou uma mudança significativa, *i.e.*, o uso da estratégia WSLS diminuiu gradualmente. Na média, os animais “C”, “E” e “F” utilizaram esta estratégia 54,8%, 53,4% e 56,5% das jogadas durante a interação com o algoritmo 2. A obtenção de reforço foi menor que durante o jogo contra o algoritmo 1.

Podemos usar a teoria de aprendizagem por reforço que descrevemos acima para estudar o desempenho dos animais na tarefa. Neste modelo, a função-valor de cada lado a ser escolhido é atualizada de acordo com o sinal de erro (*reward prediction error*) segundo a equação:

$$V_{t+1}(x) = V_t(x) + \phi[r_t - V_t(x)] \quad \text{eq. (9)}$$

onde $V_t(x)$ corresponde a expectativa de valor do lado “x” (x = esquerdo ou direito) na escolha do momento “t”, r_t corresponde ao reforço recebido no momento “t” e ϕ ao coeficiente de aprendizagem (*learning rate*), enquanto que $[r_t - V_t(x)]$ é o sinal de erro. É importante ressaltar que como nesta tarefa a passagem a outro estado não traz informações sobre estados subsequentes, o termo relativo a estes está ausente nas equações usadas na modelagem matemática do experimento.

Esta equação pode ser reescrita como:

$$V_{t+1}(x) = (1 - \phi)V_t(x) + \phi r_t \quad \text{eq.(10)}$$

Podemos reparametrizá-la como:

$$V_{t+1}(x) = \alpha V_t(x) + \Delta_t(x) \quad \text{eq. (11)}$$

utilizando $\alpha = 1 - \phi$ e o *input* exógeno $\Delta_t(x) = \phi r_t$. Um valor alto de alfa, denominado fator de decaimento (*decay factor*), indica que o resultado de uma escolha tem uma influência num grande número de escolhas posteriores e o resultado de cada escolha resulta numa atualização pequena no

valor daquele lado. Uma vez que a análise do comportamento de escolha dos animais revelou que a probabilidade de repetir uma escolha reforçada é significativamente diferente da probabilidade de mudar a escolha após não ter sido reforçado, para estimativa dos valores por máxima verossimilhança optou-se por dois parâmetros diferentes para o caso de o animal ser reforçado ou não numa jogada. Assim usamos Δ^+ para a situação em que o animal recebe reforço e Δ^- para a situação oposta. Na situação em que aquele lado não é escolhido na jogada usamos a notação Δ^0 . A probabilidade de escolha da direita (d) na jogada “t” é dada por:

$$P_t(d) = \exp(\beta \cdot V_t(d)) / [\exp(\beta \cdot V_t(d)) + \exp(\beta \cdot V_t(e))] \text{ eq. (12)}$$

que pode ser reescrita como:

$$P_t(d) = (1 + \exp(-\beta \cdot \Delta V_t))^{-1} \text{ eq. (13)}$$

onde $\Delta V_t = V_t(d) - V_t(e)$.

As estimativas dos parâmetros do modelo mostraram que durante o jogo contra o algoritmo 1 o valor de α foi baixo ($\alpha < 0,3$) e os valores de Δ foram assimétricos e levaram a atualizações importantes da função valor ($\Delta^+ > 0,5$ e $\Delta^- < -0,5$). Estes resultados são consistentes com um uso intenso da estratégia WSLS em que o resultado positivo ou negativo da última jogada influencia mais fortemente as probabilidades de escolha na próxima jogada, mas não simetricamente. Durante o jogo com o algoritmo 2 o valor do fator de desvalorização foi mais alto ($\alpha > 0,8$) para os animais “C” e “E” mas para o animal “F” foi sensivelmente mais baixo ($\alpha = 0,42$) e mais próximo de seu valor no jogo contra o algoritmo anterior. Os valores dos Δ levaram a mudanças menores da função valor; $\Delta^+ < 0,2$ e os valores de Δ^- maiores que no jogo contra o algoritmo 1, sendo que em um dos animais este valor foi positivo (animal “C”). Estes resultados são consistentes com o menor uso da estratégia WSLS observado no estudo, pois a obtenção de um reforço leva a uma atualização menor do valor daquela opção e, conseqüentemente, uma alteração menor nas probabilidades de escolha de cada lado.

O registro da atividade neural no córtex pré-frontal dorso lateral (BARRACLOUGH *et al.*, 2004) foi realizado com os animais “C” e “E” do estudo de Lee e colegas (2004), durante jogo contra o algoritmo 2. A atividade neural foi sugestiva de que neurônios pré-frontais podem estar envolvidos na integração temporal de sinais relacionados à escolha prévia do animal e seu resultado. Estes sinais seriam utilizados para atualização das funções de valor, cujas diferenças servem de base para a escolha do animal.

Estudos posteriores (SEO *et al.*, 2007, 2009; SEO; LEE, 2007, 2009; para revisão SEO; LEE, 2008) com outros animais, também jogando exclusivamente contra o algoritmo 2, avaliaram pormenorizadamente a atividade neural dos animais durante o jogo. O registro da atividade neural nos córtices pré-frontal dorso lateral (DLPFC) e cíngulo anterior dorsal (dACC) apresentam pelo menos três tipos diferentes de sinais que são relacionados às escolhas do animal e suas consequências (reforço ou não reforço), além das escolhas do oponente. No DLPFC os sinais relacionados às escolhas do animal e suas consequências são fortemente representados, enquanto que no dACC sinais relacionados aos reforços são dominantes. Em ambas as áreas, alguns neurônios também apresentam sinais relacionados às escolhas do oponente, mas este sinal era relativamente fraco. Todos os sinais decaem gradualmente no curso de duas a quatro jogadas.

Deve-se ressaltar que esses estudos utilizam uma tarefa que difere dos jogos existentes na realidade social humana. Primeiramente, porque neles o animal apenas ganha ou não ganha reforço, enquanto no jogo do *matching pennies* ganha-se ou perde-se uma quantia em dinheiro. Ademais, quantia em dinheiro corresponderia a um reforço condicionado. Seo e Lee (2009) tentaram modelar o jogo *matching pennies* de forma mais realista em três macacos rhesus; ao final de uma jogada, ganhar ou perder era acompanhado da recepção ou perda de um reforço condicionado (discos vermelhos que eram exibidos na tela do monitor e eram trocados por seis gotas do reforço quando o animal acumulava seis discos, ganhando ou perdendo um por jogada). O desempenho dos animais nesta forma de jogo foi semelhante ao desempenho na forma em que havia a liberação de um reforço incondicionado. O

registro da atividade neural trouxe como novidade que neurônios no córtex frontal dorsomedial possivelmente desempenham uma função única no ajustamento da estratégia de tomada de decisão baseada em ganhos e perdas de reforços condicionados.

A atividade neural na área LIP durante o jogo de *matching pennies* foi objeto de outro estudo. Neste Seo e col. (2009) encontraram que parte dos neurônios desta região codificava a diferença de valor entre as opções. Uma parte ainda maior de neurônios desta área tinha sua atividade relacionada à soma dos valores das funções relativas a cada opção. Este sinal poderia estar relacionado ao nível médio de reforço. Outros sinais presentes no córtex pré-frontal também estão presentes na área LIP, tais como as escolhas anteriores do animal.

1.8 Breve síntese

A literatura neuroeconômica em geral (KABLE; GLIMCHER, 2009) e sobre jogos em particular (KRUEGER *et al.*, 2008) se concentra em pesquisas com humanos e macacos, que são utilizados sempre que são usados procedimentos invasivos para registro de atividade neuronal. Neste caso os macacos são utilizados apenas como modelos do comportamento humano. Embora seja evidente a importância deste tipo de pesquisa voltada fundamentalmente ao comportamento humano, esta abordagem deixa de lado importantes questões. A tomada de decisão é uma questão central para todas as espécies, pois no organismo animal, pelo menos a partir de certo ponto da escala evolutiva, sistemas cognitivos se interpõem entre o sistema perceptivo e o sistema motor fazendo com que a partir das informações aferentes possa ser escolhida a resposta mais adaptativa ao organismo. Isto equivale a dizer que em cada situação o animal deve tomar uma decisão e, evidentemente, estas decisões são fundamentais para a sobrevivência.

Questões que dizem respeito à escolha entre ganhos de diferentes magnitudes liberados com prazos ou probabilidades diferentes, evidentemente não dizem respeito apenas à atividade financeira onde os agentes devem constantemente decidir sobre riscos e prazos associados a diferentes retornos. A atividade humana enfrenta questões deste há muito, por exemplo, as decisões dos agricultores a respeito do que plantar, pois as diversas culturas diferem em suscetibilidade a pragas e variações climáticas, prazo para colheita, bem como no valor nutricional ou mercantil do produto. Mesmo na vida cotidiana o homem enfrenta estas questões a todo o momento, como por exemplo, na escolha de um alimento pelo sabor, por um lado, e pelos riscos à saúde, por outro.

Evidentemente, a decisão econômica, significando a decisão sobre qual ação é mais adaptativa ao organismo, também é uma questão fundamental à maioria dos animais, como ressaltado acima. A literatura a respeito do comportamento e das bases neurais da tomada de decisão econômica dos roedores é imensa (para revisões DALLEY *et al.*, 2004; CARDINAL, 2006; FLORESCO; MAGYAR, 2006); também outras espécies como pombos foram pesquisados (SANABRIA; THRAILKILL, 2009). Estes estudos mostram que estes animais são capazes de tomar decisões que otimizam o retorno quando se trata de escolher entre reforços de magnitude diferentes liberados com prazos ou probabilidades diferentes, o que nos mostra que os tipos de tomada de decisão que encontramos em agentes financeiros têm origem evolutiva bastante antiga. Por outro lado, a modelagem destes comportamentos em roedores abre grandes possibilidades para a pesquisa, pois com estes animais podemos fazer pesquisas que são impossíveis, antiéticas ou antieconômicas de realizar com primatas, humanos ou não humanos. Roedores foram utilizados em muitos estudos envolvendo lesões com a finalidade de estudar o envolvimento de estruturas tanto corticais como subcorticais nos mais variados tipos de tomada de decisão econômica. Vários tipos de manipulações farmacológicas permitiram o estudo das funções de sistemas de neuromoduladores neste tipo de tomada de decisão. Também as mais variadas técnicas de biologia molecular têm sido aplicadas em roedores com a finalidade de estudar não apenas a função de certos genes e efeitos de seus polimorfismos, como também para

estudar a função de vias ou sistemas que expressam preferencialmente um determinado gene (HIKITA *et al.*, 2010). A riqueza de abordagens metodológicas que estão disponíveis para uso em roedores, associada à imensa base de dados disponível na literatura que permite comparação e análise, mostra a importância dos modelos de comportamento nestes animais.

No entanto, a quase totalidade destes estudos modela situações em que o animal interage com um agente que tem uma resposta fixa, seja ela a associação de uma magnitude de reforço a um prazo, probabilidade de liberação, punição ou esforço para obtenção. Em todas estas situações o animal descobre, a partir da experiência, a contingência que permanece fixa. Mesmo o trabalho de Zeeb e colaboradores (2009), embora envolvendo um jogo, pertence a esta categoria. Estes autores estudaram o desempenho de ratos numa versão do *Iowa Gambling Task* (BECHARA *et al.*, 1994) modificada para uso por roedores. O *Iowa Gambling Task* não é um jogo estratégico, no sentido de exigir do jogador o ajuste da estratégia a cada escolha feita, pois não há um oponente usando uma estratégia para responder às escolhas do outro jogador, como no caso visto acima do jogo *matching pennies*. Na versão do *Iowa Gambling Task* para ratos, chamada de rIGT, o animal deve fazer uma escolha entre quatro opções que fornecem reforços de diferentes magnitudes com diferentes probabilidades associados a um período de espera (durante o qual o animal está impossibilitado de jogar) cuja duração e probabilidade varia. Resumidamente, o reforço de maior magnitude é liberado com baixa probabilidade e associado a um período de espera mais longo e mais frequente. Nesta tarefa a duração da sessão é fixa, de maneira que os períodos de espera implicam num menor ganho total por sessão. Neste jogo nem a opção associada ao maior reforço nem aquela associada à menor punição são as que maximizam o ganho do animal, mas uma das intermediárias. Após experiência com a tarefa, os animais chegam a identificar esta opção e a escolhê-la com maior frequência sem, no entanto, fazer dela sua escolha exclusiva. Ressalte-se que nesta tarefa, ao contrário de paradigmas de tomada de decisão mais simples, o animal tem que integrar as variáveis da magnitude e probabilidade de reforço e de punição para encontrar a melhor escolha. Novamente, tal como humanos, os roedores

conseguem identificar a melhor estratégia e basear suas escolhas neste conhecimento, mostrando assim, mais uma vez, a importância deste tipo de modelo para a pesquisa científica em geral. Por outro lado, este tipo de resultado lança uma luz sobre a cognição animal e sobre a evolução dos primórdios da tomada de decisão racional. No entanto, como já ressaltamos acima, as decisões econômicas em todos os sentidos vão muito além da computação da magnitude, prazo, probabilidade de ganho e punição, pois envolvem fundamentalmente as decisões em que outro agente racional está também envolvido. São situações tão variáveis que vão desde a disputa com membros da própria espécie pela oportunidade reprodutiva ou alimentação até as complexas decisões tomadas pelos agentes no mercado financeiro global de hoje, passando pela complexa luta pela sobrevivência nos ecossistemas naturais onde diferentes espécies se enfrentam na luta pela sobrevivência, situações estas em que os resultados de uma escolha dependem também da escolha de outros agentes envolvidos na situação e do grau de previsibilidade de cada escolha. Como já vimos, estas situações podem ser modeladas matematicamente pela teoria dos jogos, o que evidencia a fundamental importância da pesquisa neste campo. No entanto, como ressaltado anteriormente, com exceção de uma pesquisa utilizando pombos, todas as demais pesquisas, como vimos, com jogos, foram feitas com humanos ou macacos, nenhuma com ratos.

2. OBJETIVO

Em função da importância das interações estratégicas no comportamento animal e da ausência de estudos deste tipo de comportamento em roedores, julgamos ser oportuno estudar a capacidade para este tipo de interação nestes animais, utilizando como tarefa um jogo estratégico simples como o *matching pennies*, já utilizado com macacos. Por outro lado, uma vez que constatamos que os primatas não humanos utilizam um algoritmo de reforço para descobrir e atualizar a estratégia neste jogo (para uma revisão SEO; LEE, 2008), julgamos importante estudar como os roedores o fazem na mesma situação. Dada a importância do sistema dopaminérgico no aprendizado por reforço (para revisões, BROMBERG-MARTIN *et al.*, 2010; SCHULTZ, 2010) decidimos também estudar os efeitos de agentes farmacológicos que alterem o funcionamento deste sistema neste comportamento.

Nesta pesquisa trabalhamos com dois grupos de animais. O primeiro grupo de ratos foi testado em duas versões diferentes do jogo para estudarmos se seu desempenho se aproximava do desempenho normativo da teoria dos jogos e também a sua resposta a mudanças de estratégia do oponente. Na segunda parte de nossa pesquisa, com o segundo grupo de animais, estudamos a relação do sistema dopaminérgico com a estratégia de jogo.

3. EXPERIMENTOS

3.1 Experimento I. Desempenho de ratos no jogo estratégico de “matching pennies”

Neste experimento foram testadas duas hipóteses:

1. se ratos são capazes de uma interação estratégica simples como aquela necessária para jogar o *matching pennies*, aproximando-se do desempenho normativo descrito pela teoria dos jogos; e
2. se estes animais são capazes de responder a mudanças de estratégia do oponente no sentido de maximizar a probabilidade de obtenção de reforço.

Assim, ratos foram treinados para jogar o *matching pennies* utilizando uma caixa de Skinner, em que as escolhas eram feitas através de focinhas em buracos dispostos especialmente para esta finalidade.

3.1.1 Material e Métodos

Sujeitos

Foram utilizados onze ratos Wistar machos, originários do biotério do Instituto de Biociências da Universidade de São Paulo. No início do treinamento, que durou em média 20 dias, os animais tinham 75-90 dias de vida e pesavam 250-300 g. Os animais ficaram acondicionados em caixas de polietileno medindo 40 x 33 x 16 cm, dois ou três em cada caixa, em biotério com ciclo de luz invertido (luzes acesas das 19:00 às 7:00 horas).

As sessões de treinamento e registro de dados ocorriam seis dias por semana, com início às 14:00 horas. Logo após a sessão de treinamento ou experimental os animais eram reconduzidos ao biotério onde recebiam 18 g de ração cada, sendo este o único alimento recebido. Água era fornecida *ad libitum*.

Aparelho

O aparelho consiste de uma caixa operante de acrílico branco medindo 26 x 26 x 26 cm. composta por dois módulos internos incluindo um conjunto de três buracos e um bebedouro (Figura 9, painel esquerdo) e três módulos externos incluindo o mecanismo de acionamento do bebedouro, o mecanismo de funcionamento das lâmpadas e fotocélulas que estão inseridas nos buracos, e um recipiente contendo o reforço (Figura 9, painel direito).

Os três buracos são tubos quadrados (Figura 9) que medem 2,5 cm de lado e 4,5 cm de comprimento cada e estão dispostos horizontalmente lado a lado. Na parte de trás de cada um dos buracos está localizado um LED branco de alto brilho, regulado para produzir uma luminosidade de 200 Lux na entrada do buraco. A 0,5 cm da entrada de cada buraco há um sensor conectado a um microcomputador que registra interrupções de um feixe de luz infravermelha quando o animal insere o focinho no mesmo. Imediatamente abaixo do buraco central está localizado o bebedouro, 1 cm acima do assoalho, onde o animal recebe o reforço (20 µl de solução de sacarose a 5%). Esse bebedouro é constituído de uma caixa de acrílico branco inserida na parede que contém os buracos, com abertura para dentro da caixa experimental, medindo 5,0 cm de largura, 3,5 cm de altura e 3,3 cm de profundidade, contendo a 1,0 cm da parede da caixa experimental um orifício de 8 mm diâmetro onde se disponibiliza uma cuia. Esta cuia está conectada a uma haste flexível que, por sua vez, é acionada por um motor; em cada acionamento a haste puxa a cuia para dentro de um recipiente contendo a solução de sacarose, o que leva ao seu enchimento com essa solução que é então disponibilizada para os animais. Para iluminação interna da caixa e para parte da sinalização durante a tarefa, um LED branco, de alto brilho (também regulado para produzir luminosidade de 200 Lux) está colocado no

centro do teto da caixa. A caixa operante está inserida em uma caixa de madeira compacta com 50 cm de lado. Desta forma, durante as sessões os animais ficavam num ambiente cuja única iluminação era fornecida pela luz do teto da caixa. Além disto, os ruídos ambientes eram atenuados. As lâmpadas, as fotocélulas e o motor do bebedouro estão ligados a uma interface e essa a um microcomputador que os aciona e promove o monitoramento e registro do desempenho dos animais através de um programa computacional, especialmente desenvolvido para esse fim.

FIGURA 9 – A FOTO À ESQUERDA MOSTRA A PAREDE INTERNA DA CAIXA COM OS BURACOS USADOS PARA FOCINHAR (ACIMA) E O CONSOLE COM A CUIA ONDE O ANIMAL RECEBE O REFORÇO (ABAIXO). À DIREITA, VÊ-SE A PORÇÃO EXTERNA DOS 3 TUBOS QUADRADOS E LOGO ABAIXO AS CONEXÕES DA CAIXA COM O COMPUTADOR. MAIS ABAIXO VÊ-SE O MOTOR E RESPECTIVO MECANISMO USADO PARA ACIONAR A CUIA QUE DISPENSA O REFORÇO E O RESERVATÓRIO (CAIXA DE ACRÍLICO BRANCA) DE SOLUÇÃO USADA COMO REFORÇO.

Procedimento Comportamental

Cada sessão era composta de 200 jogadas. Com a luz do teto acesa, cada jogada iniciava-se com o acendimento do buraco central, que permanecia aceso durante 5 segundos aguardando que o animal fizesse sua focinhada. Se o animal não a fizesse neste período, o buraco apagava-se e a luz do

teto ficava apagada durante 5 segundos adicionais (período de “time-out”), após os quais nova jogada era iniciada. Feita a focinhada no buraco central, soava uma campainha, a luz do fundo deste buraco se apagava, ao mesmo tempo em que se acendiam as luzes dos fundos dos buracos laterais. O animal tinha então mais 5 segundos para fazer sua escolha e focinhar um dos buracos laterais. Feita esta focinhada, o computador revelava sua escolha mantendo acesa por mais 0.5 segundo a luz do buraco escolhido, enquanto que a luz do lado não escolhido apagava-se. A escolha do computador era feita de acordo com o algoritmo 1 ou 2 de Lee e colaboradores. (2004) que serão descritos a seguir. Se o animal não fizesse sua escolha em até 5 segundos após o som da campainha ocorria um período de *time-out* com 5 segundos de duração, após o qual se iniciava nova jogada.

Dois algoritmos diferentes foram usados, um em cada etapa do experimento. Cada algoritmo foi escrito tendo em vista a detecção e exploração de vieses estatísticos no comportamento de escolha dos ratos. Para isto tinha em sua memória a história de escolhas e reforços do animal na sessão anterior e, a estes dados, acrescentava as escolhas e resultados da sessão em curso.

Algoritmo 1

No algoritmo 1 eram explorados apenas os vieses relacionados à sequência de escolhas, independentemente destas serem ou não reforçadas. Primeiramente, a probabilidade de que o animal escolheria um determinado buraco para focinhar, bem como o conjunto de probabilidades condicionais de que o animal escolheria um determinado lado, dadas as suas escolhas nas “N” jogadas anteriores ($N = 0$ a 4) eram estimadas. Então, para cada uma destas cinco probabilidades, o computador testava a hipótese nula de que o animal tinha feito sua escolha aleatoriamente com iguais probabilidades e, conseqüentemente, dos lados escolhidos nas jogadas anteriores. Quando esta hipótese nula não era rejeitada para nenhuma das probabilidades calculadas, o computador fazia sua escolha aleatoriamente com probabilidade 0,5. Caso contrário, o computador selecionava o lado de acordo com a probabilidade condicional com o maior desvio de 0,5 que fosse estatisticamente significativa (teste

binomial, $p < 0,05$). Por exemplo, se houvesse uma ocorrência significativamente mais frequente, digamos 75%, de escolha à direita após uma sequência de escolhas como direita – esquerda – esquerda e se esta probabilidade condicional desviasse mais de 0,5 que qualquer outra que atingisse significância estatística no teste, então o computador escolhia a esquerda com probabilidade 75%.

Algoritmo 2

No algoritmo 2, tanto os vieses relacionados à sequência de escolhas, como os relacionados à sequência de escolhas e seus resultados eram explorados pelo computador. Assim, além das hipóteses testadas no algoritmo 1, o segundo algoritmo testava as probabilidades condicionais de que o animal faria sua escolha dadas as “N” escolhas anteriores e seus resultados ($N = 1$ a 4). O computador fazia os mesmos testes, agora com 9 hipóteses em lugar de 5, e escolhia de acordo com a mesma regra.

Variáveis comportamentais analisadas

O desempenho dos animais foi estudado através de diversas variáveis, descritas abaixo, que focam diferentes aspectos da estratégia do animal e da sequência produzida.

Análise dos dados

Foram estudadas diversas variáveis intrínsecas à estratégia de escolha do animal, entre as quais citamos a probabilidade de reforço, efeito do reforço sobre as escolhas seguintes e diversos índices provenientes da teoria da informação que nos permitem estudar a estocasticidade e correlação interna das escolhas com as escolhas anteriores e seus resultados.

Para estudo da estrutura interna da sequência de escolhas foram calculadas diversas medidas de entropia informacional e informação mútua entre as escolhas. Em termos não matemáticos podemos definir a entropia informacional como a medida de quão aleatória é uma sequência. Assim, numa sequência de A e B a maior aleatoriedade ocorrerá quando tivermos 50% de A e 50% de B. Este

tipo de medida bastante simples apresenta sérios limites na informação que nos fornece sobre a estrutura da sequência. Imaginemos que a sequência tenha 100 símbolos. Neste caso se 50 símbolos seguidos forem A e os outros 50 forem B teremos a medida máxima de entropia, mas teremos uma sequência bastante previsível. Para contornar este problema podemos medir a frequência dos diversos pares possíveis: AA, AB, BA e BB. Medindo a frequência destes pares dentro da sequência teremos uma diferença importante entre o exemplo acima (uma sequência de 50 A seguida de uma de 50 B) e uma sequência gerada pelo jogo de uma moeda em que cara equivalesse a A e coroa a B. Medindo-se os pares teríamos no primeiro caso uma sequência de 49 AA seguida de um par AB, seguido de uma sequência de 49 BB. No caso da sequência oriunda do jogo com a moeda, teríamos uma distribuição mais próxima de equiprobabilidade dos quatro pares acima. Portanto, para estudarmos a estrutura interna de uma sequência a entropia descrita no segundo exemplo é mais apropriada. Esta entropia se chama de entropia de primeira ordem. Podemos também estudar sequências de 3 símbolos seguidos, por exemplo ABB, BAB etc., que seria entropia de segunda ordem. Na realidade podemos estudar entropias informacionais de ordem “N”.

A entropia informacional é um índice que tem valor máximo quando a aleatoriedade da sequência é máxima, isto é, um índice que assume seu valor máximo quando os diversos blocos estudados tiverem uma distribuição equiprovável e que assume valor zero quando a sequência é completamente previsível. Ela é calculada pela fórmula:

$$H = - \sum_{i=1}^k p_i \log_2 p_i$$

eq. (14)

onde H é a entropia da sequência estudada e p_i é a probabilidade de ocorrência de cada um dos k blocos de símbolos na sequência estudada, por exemplo, se temos apenas quatro possibilidades equiprováveis (AA, AB, BA e BB), $k=4$ e a entropia máxima será 2 bits. Uma vez que trabalhamos com sequências finitas e utilizamos a frequência relativa dos símbolos em lugar de sua probabilidade temos

um viés na estimativa da entropia da fonte geradora da sequência (os animais no nosso caso). Para corrigirmos este viés é necessário uma correção do valor calculado (MILLER, 1955). Assim, com a correção a fórmula acima fica:

$$H = - \sum_{i=1}^k \hat{p}_i \log_2 \hat{p}_i + \frac{k-1}{1.3863 N} \quad \text{eq. (15)}$$

Neste termo de correção N é a extensão da sequência em número de pares (no caso da sequência de 100 símbolos temos 99 pares, 98 triplas etc.).

Outro índice que utilizamos é a informação mútua, que permite medir a previsibilidade probabilística de um símbolo quando conhecemos os símbolos que o antecederam. Assim na primeira sequência que descrevemos acima vemos que a informação mútua é bastante alta, pois temos 50 A e 50 B e as transições de A para A e de B para B correspondem à quase totalidade da sequência, havendo uma única exceção que é a transição AB central. Então, um A é seguido por outro A em 49 dos 99 pares, enquanto que dos outros 50 pares, 49 são da transição B para B e apenas um par é da transição A para B. Portanto, ao encontrarmos um A ou um B nos temos uma alta probabilidade de saber o próximo símbolo, fazendo então com que a informação mútua seja alta. Já na sequência gerada aleatoriamente nada sabemos sobre a probabilidade de um símbolo quando sabemos seu antecedente; assim a informação mútua é zero. A informação mútua é calculada pela equação abaixo, já mostrando o termo de correção relativo a sequências finitas.

$$I = \sum_i^r \sum_j^c \hat{p}_{ij} \log_2 \frac{\hat{p}_{ij}}{\hat{p}_i \hat{p}_j} - \frac{(r-1)(c-1)}{1.3863 N} \quad \text{eq. (16)}$$

Nesta equação, p_i e p_j são as probabilidades dos eventos, por exemplo a probabilidade de um A na posição $n-1$ e a probabilidade do mesmo símbolo na posição n , enquanto que p_{ij} representa a probabilidade da dupla AA na sequência. No denominador temos então a probabilidade daqueles dois

símbolos ocorrerem juntos ao acaso (dada pelo produto das probabilidades), no numerador temos a ocorrência efetiva da dupla na sequência. Se a ocorrência efetiva for diferente da probabilidade de ocorrência ao acaso, então, em nosso exemplo, a presença de A na posição $n-1$ nos diz alguma coisa sobre a probabilidade da ocorrência de A na posição n . Como no caso da entropia informacional, necessitamos de uma correção para sequências finitas. No termo de correção r corresponde ao número de eventos de entrada. No caso de estudarmos a informação mútua entre o símbolo anterior e o próximo, temos dois eventos de entrada, se, no entanto, estivéssemos estudando a informação mútua entre dois símbolos anteriores (posições $n-1$ e $n-2$) e o símbolo na posição n o valor de r seria 3. A variável c corresponde aos eventos de saída, no caso apenas 2 (os dois símbolos possíveis na posição n). O valor de N no denominador tem o mesmo significado que na equação da entropia informacional.

Em nosso estudo calculamos a entropia informacional das escolhas, índice que chamamos de H_1 . Este índice nos permite estudar os afastamentos da equiprobabilidade, sem referência ao lado para o qual ocorre. Outros índices pesquisados foram o H_2 em que estudamos a distribuição das transições das escolhas $n-1$ a n (sequências DD, DE, ED e EE) e o H_3 em que consideramos as sequências das três últimas escolhas (DED, EDD etc.).

A informação mútua foi calculada entre a escolha n e a escolha $n-1$ e seu resultado, por exemplo, entre a escolha da D quando o animal tinha escolhido E na jogada anterior e tinha recebido reforço. Chamamos a este índice de MI_1 . Foi também calculada a informação mútua entre a escolha na jogada n e as escolhas nas jogadas $n-1$ e $n-2$ e seus resultados. Chamamos este índice de MI_2 . Finalmente calculamos MI_3 : a informação mútua entre a escolha na jogada n e as escolhas nas jogadas $n-1$, $n-2$ e $n-3$ e seus resultados. Quando há completa independência de uma escolha em relação às escolhas e resultados anteriores a informação mútua é zero. Ao contrário, quando uma escolha é totalmente determinada pelas escolhas anteriores, a informação mútua será 1 (no caso de duas opções, onde a entropia máxima será 1).

Para análise da estratégia de jogo dos animais utilizamos os seguintes índices:

1. **Probabilidade de reforço.**
2. **Probabilidade de “win-stay-lose-switch” (WSLS).** Frequência da estratégia de repetir na jogada seguinte a mesma opção quando esta foi reforçada e mudar quando não foi reforçada.
3. Duas medidas de **informação mútua**. Aplicando a equação (16) calculamos a informação mútua entre a opção do animal numa jogada e as N escolhas anteriores e seus resultados (N = 1 e 3), denominadas respectivamente de **MI_1** e **MI_3**.
4. Três medidas de **entropia informacional**, em que medimos a aleatoriedade (equiprobabilidade) dos blocos compostos apenas pelas escolhas laterais. No primeiro índice **H_1** estudamos apenas a distribuição da última escolha, em **H_2** estudamos a aleatoriedade nas duplas de escolhas laterais (em outras palavras, quão equiprováveis são as duplas DD, DE, ED e EE). No terceiro, **H_3**, estudamos quão equiprováveis são as triplas de escolha lateral, por exemplo DDE, EDD etc.

A análise estatística foi feita utilizando-se o teste não paramétrico de Wilcoxon para amostras pareadas. Este teste foi escolhido em função do tamanho pequeno da amostra, que impossibilita uma verificação segura da normalidade dos dados.

Além dos índices acima descritos foram estudados vários modelos de aprendizagem por reforço cujos parâmetros foram estimados pela maximização da função de verossimilhança (PAWITAN, 2001). Estes modelos foram comparados utilizando-se o BIC (*Bayesian Information Criteria*) com a finalidade de identificar qual deles tem melhor ajuste aos dados e que, portanto, seria a melhor aproximação das computações executadas pelos animais para desempenho da tarefa. Cinco modelos de aprendizagem por reforço, utilizados na literatura a respeito do jogo de *matching pennies*, foram ajustados e comparados. O primeiro modelo é o de Lee *et al.* (2004), que foi exposto acima quando da

descrição do experimento deste grupo. Neste modelo os parâmetros a serem estimados pela maximização da função de verossimilhança são o coeficiente de aprendizado α e os valores de Δ^+ e Δ^- que representam o quanto se acrescenta à expectativa de valor de um lado escolhido, em caso de recebimento ou não de reforço, respectivamente. Este modelo é definido pela equação (11) a seguir.

$$V_{t+1}(x) = \alpha V_t(x) + \Delta_t(x) \quad \text{eq. (11)}$$

Na equação acima $\Delta_t(x) = \Delta^+$ em caso de obtenção de reforço e $\Delta_t(x) = \Delta^-$ no caso contrário. O segundo modelo estudado é uma variação do primeiro em que temos apenas dois parâmetros, o coeficiente de aprendizado α e $\Delta_t(x)$, com a restrição que $\Delta^- = -\Delta^+$. O terceiro modelo (COHEN; RANGANATH, 2007) que tem os valores de Δ fixados em 1 e - 1, também utiliza dois parâmetros; o mesmo coeficiente de aprendizado α e o parâmetro η . Este último parâmetro multiplica a diferença entre o reforço recebido e a expectativa de reforço, isto é, define o quanto a violação da expectativa de reforço contribuirá para a atualização dos valores das opções. O quarto modelo é o utilizado por Lee e Seo (2007) e trabalha apenas com dois parâmetros: o coeficiente de aprendizagem α e a temperatura inversa β (termo oriundo da física estatística e relacionado à aleatoriedade de um sistema). Enquanto que α define o quanto o valor de cada opção é atualizado em função do recebimento ou não de reforço, β define o quanto a diferença entre os valores das opções influencia a escolha, ou em outras palavras, o quanto uma escolha é independente dos resultados anteriores, ou seja, estocástica. O último modelo estudado tem apenas um parâmetro que é a probabilidade de o animal repetir em uma jogada a escolha anterior do oponente. Neste modelo a estratégia do animal consiste em repetir a jogada do oponente com uma probabilidade fixa e esta é o parâmetro a ser estimado (LEE; SEO, 2007).

Os códigos de análise para otimização numérica foram escritos em linguagem R (R DEVELOPMENT CORE TEAM, 2011) tendo utilizado o algoritmo “BFGS” implementado nas funções de otimização deste aplicativo computacional tendo ainda sido verificada a igualdade dos resultados

com os fornecidos por outros algoritmos alternativos implementados tal como o de Nelder-Mead. Estimativas de erro padrão são obtidas a partir da matriz hessiana, esta obtida também numericamente. Modelos foram comparados utilizando-se o BIC ("Bayesian Information Criteria"), em que o valor da verossimilhança maximizada é penalizado em função do número de parâmetros do modelo e tamanho da amostra permitindo assim a comparação dos diferentes modelos.

A função de verossimilhança foi construída para o desempenho dos animais no jogo contra cada algoritmo, como um todo. O modelo de melhor ajuste foi escolhido com base no BIC (*Bayesian Information Content*), sendo escolhido o modelo que apresentou menor valor deste índice.

3.1.2. Resultados (do Experimento I)

Jogando contra o algoritmo 1 os animais fizeram 2.400 escolhas e tiveram como mediana da frequência de reforço 0,4813 (Tabela 1). Esta probabilidade de reforço é significativamente menor que 50% (teste de Wilcoxon para amostras pareadas, $W=3$, $p=0,0049$). Durante o jogo contra este algoritmo a mediana do uso da estratégia de WSLS foi de 0,5629, o que é significativamente maior que 50% (teste de Wilcoxon para amostras pareadas, $W=0$, $p=0,0001$).

TABELA 1 – RESULTADOS DO EXPERIMENTO 1. WSLS, PROBABILIDADE DE USO DA ESTRATÉGIA DE REPETIR A ESCOLHA DO Oponente NA PRÓXIMA JOGADA; Δ WSLS DIFERENÇA ENTRE A PROBABILIDADE DE REPETIR A ESCOLHA DO LADO QUANDO HOUVE REFORÇO E DE ESCOLHER O OUTRO LADO APÓS A ESCOLHA NÃO SER REFORÇADA; H_1, ENTROPIA INFORMACIONAL DA SEQUÊNCIA DE ESCOLHAS, NÃO INCLUINDO A ESCOLHA DO Oponente, H_2 E H_3 A MESMA ENTROPIA CONSIDERANDO OS BLOCOS DAS DUAS E TRÊS ÚLTIMAS ESCOLHAS, RESPECTIVAMENTE; M_1, A INFORMAÇÃO MÚTUA ENTRE A ESCOLHA DO ANIMAL E A SUA ESCOLHA E A DO Oponente NA JOGADA ANTERIOR; M_3, A INFORMAÇÃO MÚTUA ENTRE A ESCOLHA DO ANIMAL E O BLOCO DAS 3 ESCOLHAS ANTERIORES DO ANIMAL E DO Oponente. N.S.= NÃO SIGNIFICATIVO.

	Reforço	WSLS	Δ WSLS	H_1	H_2	H_3	MI_1	MI_3
ALG_1	0,4813	0,5629	0,1349	0,9921	1,977	2,955	0,0461	0,0811
ALG_2	0,4807	0,5392	0,0746	0,9984	1,986	2,975	0,0153	0,0551
p-valor	n.s.	0,0020	0,0049	n.s.	0,0029	0,0049	0,0029	0,0029

Os animais fizeram 6.000 escolhas jogando com o algoritmo 2. Nesta segunda etapa, a mediana da probabilidade de reforço caiu para 0,4807, também significativamente inferior a 50% (teste de Wilcoxon para amostras pareadas, $W=0$, $p=0,0001$). A mediana do uso de WSLS no segundo algoritmo baixou para 0,5392, mantendo-se ainda significativamente superior a 50% (teste de Wilcoxon para amostras pareadas, $W = 0$, $p = 0,0001$).

Comparando a probabilidade de uso de WSLS entre os dois algoritmos, constatamos que no algoritmo 2 esta é significativamente menor (Figura 10) (teste de Wilcoxon para amostras pareadas, $W=1$, $p=0,002$). Também no jogo contra o segundo algoritmo o Δ WSLS foi significativamente menor, *i. e.*, no jogo contra o algoritmo 1 a estratégia de repetir a escolha reforçada foi usada com mais frequência que a estratégia de abandonar o lado não reforçado, indicando uma maior assimetria entre os efeitos do reforço e do não reforço na tomada de decisão nesta etapa do experimento.

Embora a probabilidade de reforço tenha diminuído na segunda parte do experimento, esta diferença não atingiu significância (teste de Wilcoxon para amostras pareadas, $W = 15$, $p > 0,05$).

Examinando se houve tendência a escolher um lado mais que outro, ou seja, se a entropia de ordem zero (H_1) foi significativamente diferente do valor máximo, encontramos que é menor tanto no jogo contra o algoritmo 1 (teste de Wilcoxon para amostras pareadas, $W = 6$, $p=0,0137$), como no jogo contra o algoritmo 2 (teste de teste de Wilcoxon para amostras pareadas, $W = 1$, $p=0,0020$). O desvio da equiprobabilidade não diferiu nas duas fases do experimento (teste de Wilcoxon para amostras pareadas, $W = 24$, $p>0,05$). As entropias H_2 e H_3 tiveram aumento no jogo contra o algoritmo 2 comparadas ao jogo contra o algoritmo 1 (teste de Wilcoxon para amostras pareadas, $W = 2$, $p=0,0029$ e $W= 3$, $p=0,0049$, respectivamente), indicando que os blocos de escolhas de duplas e triplas foram mais aleatórios no jogo contra o segundo algoritmo.

FIGURA 10 – EVOLUÇÃO DO USO DA ESTRATÉGIA DE WSLS DA PRIMEIRA PARA A SEGUNDA FASE DO EXPERIMENTO (* $P < 0.05$)

Para avaliarmos o efeito do reforço nas escolhas nos dois algoritmos comparamos os índices de informação mútua descritos acima. Nos dois casos (MI_1 e MI_3) o índice diminuiu, indicando que quando passamos do primeiro algoritmo para o segundo as escolhas tornam-se mais independentes das escolhas anteriores e seus resultados (teste de Wilcoxon para amostras pareadas, $W = 2$, $p=0,0029$, para os dois índices). Comparando os valores de MI_1 com MI_3 encontramos no jogo contra ambos os algoritmos que o valor de MI_3 maior que o de MI_1 (teste de Wilcoxon para amostras pareadas, $W = 0$, $p<0,001$), indicando que os efeitos das escolhas e resultados anteriores não se limitam aos da jogada imediatamente anterior.

O modelo que apresentou melhor ajuste foi o de Lee e colegas (2004) descrito detalhadamente na sessão 1.7. Analisando os parâmetros usados para modelar o comportamento de escolha dos animais encontramos que α foi significativamente maior (mediana de 0,9281 *versus* 0,5905; teste de Wilcoxon para amostra pareadas, $W=10$, $p=0,0420$) no jogo contra o algoritmo 2 (Figura 11).

FIGURA 11 – VALOR DO COEFICIENTE DE APRENDIZADO α PARA CADA UM DOS ANIMAIS

Os valores de Δ^+ e Δ^- (delta_(+) e delta_(-) na Figura 12, respectivamente) foram ambos menores quando os animais jogaram contra o algoritmo 2, embora apenas no caso de Δ^+ esta diferença tenha atingido o nível de significância (teste de Wilcoxon para amostras pareadas, $W=4$, $p=0.0068$ para o caso de Δ^+ e $W=33$, $p=n.s.$ para o caso de Δ^-). Nos dois algoritmos houve diferença significativa entre os valores de Δ^+ e Δ^- , sendo sempre $\Delta^+ > \Delta^-$ (teste de Wilcoxon para amostras

pareadas, $W=0$, $p<0,001$ em ambos os casos). No jogo contra ambos os algoritmos os valores de Δ^+ foram sempre positivos, já os valores de Δ^- foram tanto negativos quanto positivos.

Os valores de Δ^+ e Δ^- (delta_(+) e delta_(-) na Figura 12, respectivamente) foram ambos menores quando os animais jogaram contra o algoritmo 2, embora apenas no caso de Δ^+ esta diferença tenha atingido o nível de significância (teste de Wilcoxon para amostras pareadas, $W=4$, $p=0.0068$ para o caso de Δ^+ e $W=33$, $p = \text{n.s.}$ para o caso de Δ^-).

FIGURA 12 – VALORES DOS PARÂMETROS DELTA NO JOGO CONTRA OS ALGORITMOS 1 E 2

(* $P < 0.05$ E + $P < 0.001$).

Para a estimativa da temperatura inversa, β na equação (13) utilizamos o modelo de Seo e Lee (2007), uma vez que este parâmetro não é recuperável no modelo de Lee e colegas (2004) por aparecer num produto com a diferença de outros parâmetros (Δ^+ e Δ^-).

$$P_t(d) = (1 + \exp(-\beta \Delta V_t))^{-1} \text{ eq. (13)}$$

É importante ressaltar que se tratam de modelos diferentes que supõem que o comportamento de escolha dos animais é controlado por parâmetros diferentes, embora com o mesmo nome. Neste segundo modelo o parâmetro α (coeficiente de aprendizado) recebe uma interpretação diferente da do modelo de Lee e colegas (2004), pois determina o comportamento na relação com β (temperatura inversa) e não com os parâmetros Δ^+ e Δ^- , que neste segundo modelo são constantes. Também, na notação adotada em Lee e colegas (2004), adotamos a convenção de reescrever α como $1 - \alpha$. Isto faz com que esperemos o crescimento de α do primeiro para o segundo algoritmo no modelo usado em Lee e colegas (2004) e a evolução inversa no modelo usado em Seo e Lee (2007). Neste modelo não encontramos diferença entre os valores de α e β no jogo contra os algoritmos 1 e 2 (teste de Wilcoxon para amostras pareadas, $W = 29$ e 30 respectivamente, $p = \text{n.s.}$ para ambos os casos). A Figura 13 mostra a relação dos valores entre estes parâmetros, sendo que o aumento de um é compensado pela diminuição do outro. Este mesmo tipo de gráfico foi obtido por Seo e Lee (2007b).

FIGURA 13 – RELAÇÃO ENTRE O COEFICIENTE DE APRENDIZADO (II) E A TEMPERATURA INVERSA (II). OS DADOS REFERENTES AOS DOIS ALGORITMOS FORAM AGREGADOS

3.1.3 Discussão

Neste experimento estudamos a estratégia de jogo dos ratos para abordarmos duas questões fundamentais. A primeira diz respeito à capacidade destes animais jogarem o jogo do “matching pennies” que se manifesta pela obtenção de um percentual de reforço próximo do nível ótimo previsto na teoria dos jogos, isto é, 50%. Isto só é possível se os animais forem capazes de utilizar uma estratégia que aproxime a estratégia ótima que consiste em uma sequência equiprovável e estocástica de escolhas. A segunda questão diz respeito à capacidade dos animais de responderem a mudanças de estratégia do oponente com mudanças na estrutura de sua sequência de escolhas. Esta segunda

questão nos permite lançar uma luz sobre o mecanismo usado pelos animais para adotar uma estratégia para este jogo. Os resultados mostraram que roedores são capazes de aproximar a probabilidade de reforço ótima da teoria dos jogos, e que respondem a diferentes estratégias utilizadas pelo oponente. Também, esta mudança de estratégia em resposta ao oponente indica que um algoritmo de aprendizagem por reforço está na base da estratégia dos animais neste jogo.

Os resultados mostram que os ratos são capazes de um desempenho muito próximo ao de jogadores ideais previstos pela teoria dos jogos. No algoritmo 1 a mediana da probabilidade de reforço foi de 48,3%, que, embora significativamente menor que 50%, aproxima-se bastante do valor normativo. Já no algoritmo 2 este índice ficou em 47,7%, diferença não significativa em relação ao algoritmo 1. Estes resultados se aproximam bastante dos obtidos por Lee e colegas (2004) com seu experimento envolvendo macacos. No algoritmo 1 seus animais obtiveram uma probabilidade de reforço de 48,9%, 49,1% e 49,5%, enquanto que no jogo contra o algoritmo 2 a probabilidade de reforço ficou em 47,6%, 46,9% e 47,8%, respectivamente. Deve-se salientar que apesar dos macacos do experimento de Lee e colegas (2004) terem tido uma interação muito maior com o oponente, cada um fazendo em torno de 50.000 escolhas contra o algoritmo 1 e em torno de 37.000 contra o algoritmo 2, os resultados foram muito próximos dos do nosso experimento com ratos. O fato de o computador ter em sua memória para cálculos uma base maior de dados pode ter compensado esta maior experiência, pois em nosso experimento, em que cada sessão havia 200 escolhas, a base de dados usada pelo computador atingia 400 escolhas ao final da sessão (uma vez que a sessão anterior era também usada para os cálculos), enquanto que no experimento de Lee e colegas (2004) cada sessão tinha mais de 1.000 escolhas. Outro fator interessante a ser considerado na comparação dos resultados dos dois trabalhos é a questão da diferença de capacidade cognitiva entre as duas espécies utilizadas. Os resultados obtidos indicam que para esta tarefa esta diferença cognitiva não é tão importante como para outras. De fato, Soltani e colaboradores (2006) simularam os resultados do trabalho de Lee e colaboradores (2004) com uma rede neural composta de apenas 2.600 unidades.

Isto pode ser indicativo de que a geração de sequências aleatórias é um comportamento bastante antigo filogeneticamente e bastante difundido entre as espécies e que não dependeria de estruturas neurocognitivas muito complexas. Isto se deveria a sua importância para a sobrevivência nos mais diversos ecossistemas. A esse respeito, vale ressaltar que em ambos os experimentos a probabilidade de reforço foi significativamente inferior a 50%, indicando que nos dois casos as escolhas feitas foram previsíveis pelo computador oponente, que explorou isto em detrimento dos animais. O fato de os animais, nos dois experimentos, terem sempre feito um uso de WSLS maior que o acaso indica que esta previsibilidade está ligada à história dos reforços recebidos, fato este que aponta para o uso de um algoritmo de aprendizagem por reforço no processo de tomada de decisão.

O segundo ponto analisado neste experimento é como os animais reagiram à mudança de estratégia pelo oponente. No algoritmo 1 o computador usava somente a história de escolhas dos animais para prever a próxima escolha do animal e decidir o lado que escolheria naquela jogada. Por não utilizar a informação a respeito dos reforços obtidos pelos animais deixava aberta a possibilidade de utilização da estratégia de repetir na próxima jogada a escolha reforçada ou de mudar a escolha quando não havia obtido reforço; a chamada estratégia de ganha-repete-perde-muda (“win-stay-lose-switch strategy”, WSLS). Já no segundo algoritmo o computador levava em consideração a informação dos reforços obtidos em cada escolha, de modo que a utilização da estratégia de WSLS resultava em menor probabilidade de reforço. Como foi apontado anteriormente, em ambos os algoritmos os animais utilizaram a estratégia de WSLS com uma probabilidade significativamente superior a 50% (mediana de 56,29% e 53,92% para o algoritmos 1 e 2, respectivamente). Este decréscimo significativo na probabilidade de uso de WSLS quando jogando contra o algoritmo 2 indica que os animais passaram a fazer suas escolhas com uma maior independência dos reforços obtidos quando o oponente passou a utilizar os dados relativos a esses reforços nos cálculos em que baseava sua decisão.

É interessante comparar novamente o desempenho dos ratos com o dos macacos do trabalho de Lee e colegas (2004). Os macacos tiveram um uso de WSLS no algoritmo 1 bastante

superior aos ratos de nosso experimento: 65%, 73% e 63% *versus* 56,3% em nosso experimento. Já no algoritmo 2 esta diferença foi bem menor: 55%, 53% e 56% *versus* 53,9%. Este uso menor desta estratégia provavelmente deve-se ao menor número de jogadas contra o algoritmo 1 em nosso experimento, o que pode ter levado nossos animais explorá-la em menor escala em função de uma menor experiência com o algoritmo. Embora o número de jogadas contra a algoritmo 2 também tenha sido bem menor em nosso experimento esta diferença de exposição foi menos importante. Podemos levantar a hipótese de que existiria um limiar de exposição ao oponente a partir do qual o comportamento de escolha mudaria pouco. Este limiar teria sido cruzado no caso do algoritmo 2 em nosso experimento, mas não no caso do algoritmo 1.

Estes dados levantam a questão da geração de sequências aleatórias. O comportamento de um jogador ideal neste jogo é a geração de sequências completamente aleatórias, isto é, independente da escolha anterior e independente de ter ou não recebido reforço. Isto equivale a uma sequência independente das escolhas anteriores do animal e do computador. É um fato conhecido que mesmo humanos têm muita dificuldade tanto de reconhecer, como de gerar sequências aleatórias (RAPOPORT; BUDESCU, 1997). O uso de WSLS, numa frequência acima do acaso, aponta na direção do uso de um algoritmo de aprendizagem por reforço, por parte dos animais, para desenvolver a estratégia de jogo, dando maior valor aos lados escolhidos que foram reforçados nas jogadas anteriores. Também o aumento da informação mútua entre a escolha e as jogadas anteriores reforça esta ideia. O fato de haver um crescimento do valor de MI_3 comparado a MI_1 mostra que não somente o resultado da jogada imediatamente anterior, mas também os resultados de jogadas mais remotas eram integrados no processo de tomada de decisão. Estes dados são consistentes com o modelo de aprendizado por reforço descrito na sessão 1.6, pois caso o animal, em vez de integrar o valor dos resultados de várias jogadas sucessivas, apenas escolhesse, com uma probabilidade fixa, o lado escolhido pelo oponente na jogada anterior MI_3 não seria maior que MI_1. Esta estrutura de resultados é consistente com um baixo coeficiente de aprendizagem (*learning rate*) nas equações de

aprendizado por reforço discutidas acima (ϕ na eq. 9), pois quanto menor é o valor deste parâmetro maior é o peso relativo dos resultados das jogadas mais antigas em relação à imediatamente anterior. No jogo contra o algoritmo 2, quando os animais passaram a ser reforçados com menor frequência pela utilização da estratégia de WSLS, estes diminuíram o seu uso e consequentemente os índices de informação mútua calculados foram menores. Estes, como no jogo contra o algoritmo 1, tiveram valores crescentes de MI_1 para MI_3, tendo as mesmas implicações. Na comparação dos valores da informação mútua das duas fases do experimento, a diminuição do valor dos índices no algoritmo 2 indica que a atualização do valor de cada buraco foi menos importante que no jogo contra o algoritmo 1, pois as escolhas e resultados anteriores têm menor efeito na escolha seguinte. Este resultado é consistente com um coeficiente de aprendizagem (ϕ na eq. 9) menor na segunda fase do experimento. Para maior clareza, convém lembrar que neste modelo reparametrizamos a eq. 9 introduzindo o coeficiente de aprendizado α (tal que $\alpha = 1 - \phi$) e os parâmetros Δ^+ e Δ^- . Com esta reparametrização chegamos à eq. 11 ($V_{t+1}(\mathbf{x}) = \alpha V_t(\mathbf{x}) + \Delta_t(\mathbf{x})$). Isto fez com que no jogo contra o algoritmo 2 o valor de α aumentasse em relação ao algoritmo 1.

As conclusões desta análise das sequências de escolhas dos animais e do oponente computadorizado do ponto de vista da teoria da informação são bastante coerentes com os resultados das estimativas de parâmetros dos modelos matemáticos feitas por maximização da função de verossimilhança. Entre os modelos estudados o que melhor se ajustou aos dados foi o de Lee e colegas (2004). Comparando as duas fases do experimento, encontramos um aumento do coeficiente de aprendizado α e valores menores de Δ^+ na segunda fase. Estes resultados, consistentes com as análises que fizemos utilizando a teoria da informação, mostram que quando jogavam contra o algoritmo 2 os animais atualizaram o valor das opções mais lentamente por darem menor peso ao resultado da escolha mais recente. Outro ponto importante a ressaltar é que, nesta análise, como na análise utilizando os conceitos da teoria da informação, o modelo de jogo que consistiria em repetir, com uma probabilidade fixa, a escolha do computador na jogada anterior foi rejeitado. Apesar das

semelhanças qualitativas com os resultados de Lee e colegas (2004) nossos resultados apresentaram diferenças quantitativas: a diferença dos coeficientes de aprendizado nas duas fases do experimento foram menores em nosso caso, resultado consistente com o uso menos frequente, em nosso experimento, da estratégia de WSLS no jogo contra o algoritmo 1. Também os valores de Δ^- foram frequentemente positivos em nossos resultados, contrastando com valores quase sempre negativos nos resultados do experimento de Lee e colegas (2004). De qualquer maneira, para a tomada de decisão o modelo mostra que o que é computado é a diferença entre os valores.

Resumindo, este experimento mostra claramente que, em interações estratégicas, roedores são capazes de aproximar o desempenho normativo descrito pela teoria dos jogos e que para isto utilizam um algoritmo de aprendizagem por reforço. Este resultado corresponde à primeira demonstração da capacidade de roedores de jogarem o jogo *matching pennies*, abrindo novas possibilidades de utilizar este tipo de abordagem experimental em estudos sobre as bases neurobiológicas de processos de tomada de decisão.

3.2 Experimento II. Os efeitos de agonistas e antagonistas dopaminérgicos na estratégia de jogo dos animais

Como mostrado no Experimento I, ratos abordam o jogo de *matching pennies* utilizando um algoritmo de reforço. Há consenso na literatura sobre o envolvimento do sistema dopaminérgico neste tipo de aprendizado (ver sessão 1.6.3). Resumidamente, sempre que o animal se depara com um reforço não previsto há uma descarga fásica dos neurônios dopaminérgicos situados no tronco encefálico, que é proporcional à magnitude e probabilidade de reforço. A mesma resposta também ocorre quando diante da presença imprevista de um estímulo condicionado associado a um reforço.

Estas descargas levam a alterações sinápticas que estão na base do aprendizado. Como a tarefa do *matching pennies* envolve reforço probabilístico seguido a uma ação, é natural esperar mudanças na estratégia de jogo dos animais induzidas tanto por agonistas inespecíficos, como por agonistas e antagonistas específicos dos receptores D1 e D2. Estes dois tipos de receptores têm expressão e função diferenciadas nos circuitos córtico-estriato-tálamo-corticais envolvidos no processo de aprendizado de tomada de decisão. Portanto, esperamos que alterações induzidas farmacologicamente do sistema como um todo e de partes específicas do circuito produzam alterações na estratégia de escolha dos animais. Assim, nesta segunda fase estudamos os efeitos da modulação dopaminérgica sobre o desempenho dos animais no jogo do *matching pennies*.

Partindo da hipótese que agentes dopaminérgicos que alteram o equilíbrio entre as via direta e indireta, devem produzir a nível comportamental um aumento ou diminuição do efeito do reforço sobre as escolhas seguintes alterando o valor de WSLS ou de Δ WSLS e de outros índices correlatos oriundos da teoria da informação, nossas hipóteses foram:

(1) que agonistas D1 aumentariam os efeitos dos reforços recebidos nas escolhas subsequentes, causando um maior uso da estratégia de repetir na próxima jogada a escolha reforçada que se manifestaria por um aumento de Δ WSLS; no caso dos antagonistas D1 sendo esperada a ausência ou uma diminuição ou deste efeito em relação à situação controle;

(2) que agonistas/antagonistas D2 teriam efeitos diferentes dependendo dos receptores atingidos; se a dose utilizada fosse pequena o suficiente para estimular/inibir apenas os receptores pré-sinápticos ou igualmente pré e pós-sinápticos (os efeitos esperados em cada caso serão discutidos na apresentação dos resultados)

(3) agonistas inespecíficos, que causam um aumento da liberação de dopamina o efeito esperado seria estimulação da via direta e inibição da via indireta tendo como efeito um aumento no valor de Δ WSLS.

3.2.1 Material e Método

Sujeitos

Foram utilizados 10 ratos Wistar machos provenientes do biotério do Setor de Ciências Biológicas da Universidade Federal do Paraná. Os animais tinham entre 75 e 90 dias e pesavam 250-300 g no início do treinamento, que durou em média 20 dias. A acomodação, alimentação e ciclo de luz era idêntica à dos animais do Experimento 1.

Aparelho

Os animais foram testados nos mesmos aparelhos utilizados no Experimento 1.

Procedimento comportamental

Neste experimento, as sessões envolveram 300 jogadas. O computador dispunha de um registro de todas as escolhas e resultados da sessão anterior, mas utilizava apenas as últimas 300 escolhas. Em outras palavras, a cada nova jogada, correspondente à escolha do animal e seu resultado, os dados da mais antiga jogada eram abandonados e os dados da nova jogada eram incluídos; assim, a base utilizada para cálculo contava sempre com as 300 últimas escolhas do animal.

Neste experimento os animais jogaram apenas contra o algoritmo 2.

Fármacos utilizados

Depois que os animais atingiram um nível assintótico de desempenho na tarefa *matching pennies*, jogando contra o algoritmo 2, foram submetidos a testes envolvendo a administração de fármacos que alteram a função dopaminérgica. Para testar os efeitos da manipulação da função dopaminérgica utilizamos cinco substâncias, incluindo dois antagonistas específicos (SCH 23990 e

eticloprida), dois agonistas específicos (SKF 87291 e bromocriptina) e um agonista inespecífico (dl-metanfetamina), nas doses apresentadas na tabela abaixo.

TABELA 2 – SUBSTÂNCIAS, TIPO DE ATIVIDADE E DOSES ADMINISTRADAS DURANTE O EXPERIMENTO 2.

Atividade	Substância	Dose 1	Dose 2
Agonista D1	SKF 81297	0.3 mg/kg	0.6 mg/kg
Agonista D2	Bromocriptina	1.25 mg/kg	2.50 mg/kg
Antagonista D1	SCH 23330	0.001 mg/kg	0.004 mg/kg
Antagonista D2	Eticloprida	0.01 mg/kg	0.04 mg/kg
Agonista Inesp.	dl-Metanfetamina	1mg/kg	1.5 mg/kg

Com exceção da metanfetamina, todas as drogas foram adquiridas junto a Sigma Aldrich. A metanfetamina foi gentilmente cedida pelo Departamento de Psicobiologia da UNIFESP – Escola Paulista de Medicina e foi analisada no Laboratório de Química Orgânica da Universidade Federal do Paraná, tendo seu grau de pureza avaliado como sendo superior a 95%.

Com exceção da bromocriptina, todas as drogas foram dissolvidas em salina fisiológica. A bromocriptina foi primeiramente dissolvida em dimetil sulfoxido e posteriormente esta solução foi diluída meio a meio em salina. Todas as drogas foram administradas por via intraperitoneal, sempre no volume de 1 ml de solução por kg de peso do animal. A metanfetamina e a bromocriptina foram administradas 40 minutos antes da sessão experimental; as demais drogas foram administradas 15 minutos antes da sessão experimental. As doses escolhidas correspondem a doses que na literatura produziram efeitos em tarefas de tomada de decisão (ver ZEEB *et al.*, 2009; ST ONGE; FLORESCO, 2009; WINTER *et al.*, 2009).

O efeito desses fármacos foi avaliado utilizando-se uma estratégia experimental que seguiu um ciclo de seis dias para cada fármaco. Isso incluiu três dias de estabilização, seguidos de uma sessão realizada após a administração de veículo. No dia seguinte o animal era exposto a uma sessão sob efeito da droga a ser estudada. No dia seguinte à administração da droga não havia qualquer tipo de atividade. Cada sessão experimental tinha duração de 45 minutos ou 300 escolhas, o que ocorresse antes. Esse esquema foi repetido para cada uma das drogas e doses. A sequência da administração das drogas foi SCH 23390, eticloprida, SKF 81297, bromocriptina e metanfetamina, sempre com a primeira dose sendo a menor.

Análise dos dados

O desempenho dos animais após aplicação de cada droga foi comparado ao seu próprio desempenho após a injeção do veículo no dia imediatamente anterior. Esta comparação foi feita utilizando-se o teste não paramétrico de Wilcoxon para amostras pareadas. Este teste foi escolhido em função do tamanho pequeno da amostra, que impossibilita uma verificação segura da normalidade dos dados.

Foram estudadas diversas variáveis, algumas intrínsecas à estratégia de escolha do animal e outras relacionadas às propriedades da emissão da resposta motora. Para o estudo da estratégia de escolha do animal utilizamos as mesmas variáveis que no Experimento 1, acrescidas de um índice que permite estudo de variações na composição interna de WSLS, que chamamos de Δ WSLS, que será explicado a seguir, e mais um índice de medida de informação mútua, a informação mútua entre a escolha do animal e suas duas escolhas anteriores com seus resultados, calculados com a fórmula descrita acima (eq. (16)). Também incluímos a medida da probabilidade de repetir a escolha anterior. Para estudo da emissão da resposta motora focamos nos tempos de reação e nos tipos de erro cometidos pelo animal.

Para estudarmos o efeito dos agentes farmacológicos no desempenho dos animais utilizamos mesmos os mesmos índices do Experimento 1, acrescidos dos seguintes:

- 1 Mais uma medida de informação mútua; aplicando a eq. (4) calculamos a informação mútua entre a opção do animal numa jogada e as 2 escolhas anteriores e seus resultados; denominamos este índice de **MI_2**.
- 2 **Probabilidade de repetição**: a proporção de jogadas em que o animal simplesmente repete a jogada anterior, independentemente de reforço.
- 3 Erro de **precocidade**: a proporção em que animal, ao não receber reforço, não aguarda o início da próxima jogada, mas faz um movimento de focinhar ou o buraco central ou um dos laterais.
- 4 Abandonos. Dividimos a situação de abandono em três grupos: (1) quando os animais abandonam o jogo por uma única jogada, o **abandono_1**, (2) abandonos de duas a quatro jogadas **abandono_2_4** e (3) abandonos de cinco ou mais jogadas, que chamaremos de **abandono_5**.
- 5 **Numero de Jogadas** na Sessão. O número de jogadas totais completadas em cada sessão.
- 6 **Tempo de reação**: o tempo decorrido entre o início da focinhada no buraco central e o início da focinhada no buraco lateral.

3.2.2 Resultados (do Experimento II)

Os resultados são apresentados a seguir levando-se em consideração o tipo de alteração fisiológica promovida pelos fármacos utilizados.

3.2.2.1 Agonistas Dopaminérgicos Específicos

3.2.2.1.1 SKF 81297

TABELA 3 – PRINCIPAIS ALTERAÇÕES PRODUZIDAS SKF 81297 NA DOSE DE 0,3 MG/KG SOBRE OS ÍNDICES DE ESTRATÉGIA DE JOGO DOS RATOS

	WSLS	Δ WSLS	MI_1	MI_2	MI_3
SALINA	0,5483	- 0,0379	0,0067	0,0189	0,0754
SKF 81297	0,5772	- 0,0047	0,0214	0,0302	0,0860
P-VALOR	0,0117	N.S.	0,0039	0,0488	N.S.

Como agonista dos receptores D1 utilizamos o **SKF 81297** em duas doses. Na dose mais baixa, de 0,3 mg/kg de peso, encontramos alterações na estratégia do animal, quando comparamos os efeitos da droga com a sessão controle (Figura 14 e 15); há um aumento significativo do uso da estratégia de WSLS (teste de Wilcoxon para amostras pareadas, $W = 2$, $p = 0,0117$).

A estrutura interna deste índice não mudou, pois não houve diferença significativa entre a probabilidade de escolher o lado anteriormente reforçado (WS) e a probabilidade de escolher o outro lado quando não reforçado na última jogada (LS) (teste de Wilcoxon para amostras pareadas, $W = 18$, $p > 0,05$). Vale ressaltar que tanto sob efeito da droga como na condição controle, a diferença entre WS e LS não é significativamente maior que zero (teste de Wilcoxon para amostras pareadas, $W=16$ e $W=25$, respectivamente, $p > 0,05$, em ambos os casos).

FIGURA 14 – EFEITO DO SKF 81297 NA DOSE 0,3 MG/KG NO USO DA ESTRATÉGIA DE WSLS ($p < 0.005$)

Na mesma linha, houve aumento dos índices de informação mútua MI_1 (teste de Wilcoxon para amostras pareadas, $W = 1$, $p = 0,0039$) e MI_2 (teste de Wilcoxon para amostras pareadas, $W = 8$, $p = 0,0488$). Já a informação mútua entre a escolha e as três escolhas anteriores, MI_3, e seus resultados não apresentou variação (teste de Wilcoxon para amostras pareadas, $W = 22$, $p > 0,05$).

FIGURA 15 – EFEITOS DO SKF 81297 NA DOSE DE 0,3 MG/KG SOBRE (A) MI_1, (B) MI_2, (C) MI_3 E (D) Δ WSLS (* $p < 0.05$)

Este quadro de variação do valor dos três índices da informação mútua entre a escolha feita e as escolhas anteriores e seus resultados leva à conclusão de que os efeitos desta droga na estratégia de escolha estendem-se até a segunda escolha anterior, não ultrapassando este ponto. A probabilidade de reforço não apresentou diferença em relação à sessão controle uma vez que, para fazer a previsão da próxima escolha do animal, o computador usava também dados da sessão anterior, juntamente com dados da sessão atual. Assim, em função da base de dados usada, as alterações de estratégia provocadas pela droga não foram suficientes para produzir uma variação significativa da frequência de reforço. Com esta dose houve também um aumento significativo do índice

abandonos_2_4 (abandonos da tarefa por 2 a 4 jogadas), (teste de Wilcoxon para amostras pareadas, $W = 0$, $p = 0,0078$). Em função de falha de programação não dispomos dos dados referentes à sessão em que os animais receberam a dose de 0,6 mg/kg de SKF 81297.

3.2.2.1.2 Bromocriptina

O agonista D2 bromocriptina na dose de 1,25 mg/kg não alterou significativamente os índices relativos à estratégia. Obtivemos, no entanto, um aumento do tempo de reação (teste de Wilcoxon para amostras pareadas, $W = 0$, $p = 0,0020$) e uma diminuição das emissões precoces de resposta (teste de Wilcoxon para amostras pareadas, $W = 3$, $p = 0,0098$). Com a dose de 2,5 mg/kg de peso encontramos alterações importantes na estratégia de escolha (Figura 16 e 17). Embora não tenha havido aumento na frequência do uso da estratégia de WSLS (teste de Wilcoxon para amostras pareadas, $W = 14$, $p > 0,05$), a composição deste índice alterou-se. O valor de Δ WSLS exibiu uma diminuição significativa (teste de Wilcoxon para amostras pareadas, $W = 8$, $p = 0,0488$).

TABELA 4 – PRINCIPAIS ALTERAÇÕES PRODUZIDAS PELA BROMOCRIPTINA NAS DOSES DE 1,25 E 2,50 MG/KG SOBRE OS ÍNDICES DE ESTRATÉGIA DE JOGO DOS RATOS

	WSLS		Δ WSLS		% REPETIÇÃO	
	BR 1.25	BR 2.50	BR 1.25	BR 2.50	BR 1.25	BR 2.50
DMSO-SALINA	0.5750	0.5756	0.0176	- 0.0110	0.50500	0.4950
BROMOCRIPTINA	0.5817	.05903	- 0.0321	- 0.0531	0.4799	0.4701
p-VALOR	N.S.	N.S.	N.S.	0.0488	N.S.	0.0391

FIGURA 16 - EFEITOS DA BROMOCRIPTINA 2,5 MG/KG: (A) SOBRE A REPETIÇÃO DE LADO ESCOLHIDO
(B) SOBRE A FREQUÊNCIA DO USO DE WSLS (* P < 0.05)

Isto mostra que os animais aumentaram a utilização de LS (mudar de lado após a escolha não reforçada) em detrimento de WS (repetir a escolha que foi reforçada). Na mesma linha, encontramos uma diminuição da repetição do lado escolhido (teste de Wilcoxon, $W = 5$, $p = 0,0391$) e uma tendência ao aumento de H_1 (teste de Wilcoxon, $W = 6$, $p = 0,0547$) indicando uma tendência ao aumento da equiprobabilidade dos lados escolhidos. Foi observado também um aumento do tempo de reação (teste de Wilcoxon, $W = 0$, $p = 0,0020$), uma diminuição da emissão precoce de respostas (teste de Wilcoxon, $W = 3$, $p = 0,0098$) e um aumento dos abandonos de cinco ou mais jogadas (teste de Wilcoxon, $W = 0$, $p = 0,0313$).

FIGURA 17 - EFEITO DA BROMOCRIPTINA 2.5 MG/KG SOBRE AS PROBABILIDADES DO ANIMAL REPETIR A ESCOLHA REFORÇADA VERSUS EVITAR A ESCOLHA NÃO REFORÇADA NA PRÓXIMA JOGADA (* P < 0.05)

3.2.2.2 Antagonistas Dopaminérgicos Específicos

3.2.2.2.1 SCH 23390

O antagonista dopaminérgico D1, o SCH 23390 na dose de 0,001 mg/kg, levou a um aumento de valor de duas variáveis que se aproxima do nível de significância: probabilidade de uso de WSLS (teste de Wilcoxon para amostras pareadas, $W = 10$, $p=0,0820$) e de H₂ (teste de Wilcoxon para amostras pareadas, $W = 9$, $p=0,0620$). Do ponto de vista da emissão de respostas, não encontramos qualquer diferença significativa. Sob efeito do SCH 23390 na dose 0,004 mg/kg os animais não apresentam qualquer alteração significativa da estratégia de escolha. Com esta dose aparecem diferenças importantes na emissão de respostas. Isto é, há um aumento generalizado de

abandonos da atividade. Quando consideramos o abandono de apenas uma jogada temos um aumento significativo (teste de Wilcoxon para amostras pareadas, $W = 0$, $p = 0,0078$) e, no caso de abandonos de duas a quatro jogadas e de mais de cinco jogadas, temos um aumento que se aproxima do nível de significância (teste de Wilcoxon para amostras pareadas, $W = 0$, $p = 0,0625$, em ambos os casos). Houve também uma diminuição significativa das respostas precoces (teste de Wilcoxon para amostras pareadas, $W = 2$, $p = 0,0059$) e um aumento significativo dos tempos de reação (teste de Wilcoxon para amostras pareadas, $W = 4$, $p = 0,0137$). Estes índices indicam que o bloqueio dos receptores D1 levou a uma menor motivação pela atividade.

3.2.2.2.2 Eticloprida

Com o antagonista específico D2 eticloprida na dose de 0,01 encontramos apenas uma variação significativa, um aumento de H_2 (teste de Wilcoxon para amostras pareadas, $W = 5$, $p = 0,0195$). Na dose de 0,04 mg/kg não houve qualquer alteração. Em relação à emissão de resposta não encontramos nenhuma variação significativa com nenhuma das duas doses.

3.2.2.3 Agonista Dopaminérgico Inespecífico

Como agonista dopaminérgico inespecífico foi utilizada a dl-metanfetamina nas doses de 1,0 e 1,5 mg/kg de peso. Com a dose menor constatamos um aumento significativo em ΔWSL (teste de Wilcoxon para amostras pareadas, $W = 7$, $p = 0,0371$), sem alterações significativas de WSL (teste de Wilcoxon para amostras pareadas, $W = 17$, $p = n.s.$), sugerindo que os animais aumentaram o uso da estratégia de repetir a escolha reforçada em detrimento da estratégia de mudar a escolha do lado quando não reforçados. Quando calculada a porcentagem de numa jogada repetir a escolha anterior, independentemente do reforço, também constatamos um aumento significativo (teste de Wilcoxon para

amostras pareadas, $W = 5$, $p = 0,0391$). A entropia das triplas de escolhas, H_3 , exibiu diminuição significativa (teste de Wilcoxon para amostras pareadas, $W = 2$, $p = 0,0059$), indicando uma maior

TABELA 4 – PRINCIPAIS ALTERAÇÕES PRODUZIDAS PELA METANFETAMINA NAS DOSES DE 1,0 E 1,50 MG/KG SOBRE OS ÍNDICES DE ESTRATÉGIA DE JOGO DOS RATOS

	WSLS		Δ WSLS		% REPETIÇÃO	
	MET 1.0	MET 1.5	MET 1.0	MET 1.5	MET 1.0	MET 1.5
SALINA	0.5633	0.5523	0.0163	0.0186	0.5050	0.5084
METANFETAMINA	0.5533	0.5167	0.1053	0.1245	0.5452	0.5552
p-VALOR	N.S.	0.0117	0.0371	0.0391	0.0391	0.0391

monotonia na sequência de lados escolhidos quando considerados três a três. Estes três índices apontam para uma tendência a perseverar nas respostas emitidas, tendência que também foi revelada na composição do índice de *win-stay-lose-switch* (Δ WSLS), *i.e.*, houve maior repetição da escolha reforçada (WL) em detrimento da mudança de lado após escolha não reforçada (LS). A tendência a perseverar deve ser considerada juntamente com o peso atribuído ao reforço na jogada anterior quando decidindo a próxima escolha. Houve também um aumento de MI_3 , a informação mútua entre a escolha e as três escolhas anteriores e seus resultados (teste de Wilcoxon para amostras pareadas, $W = 0$, $p = 0,0020$), evidenciando também a importância atribuída aos reforços anteriores na escolha, mas apenas quando considerados os três anteriores conjuntamente. Não houve alterações significativas no campo da emissão das respostas. Quando os animais receberam a dose de 1,5 mg/kg houve alterações mais importantes. Isto é, além do aumento em Δ WSLS (teste de Wilcoxon para amostras pareadas, $W = 5$, $p = 0,0391$), houve também uma diminuição da utilização de WSLS (teste de Wilcoxon para amostras pareadas, $W = 2$, $p = 0,0117$) (Figura 19).

FIGURA 19 – EFEITOS DA METANFETAMINA NA DOSE DE 1,5 MG/KG: (A) NO USO DA CESTRATÉGIA DE WSL E (B) NA COMPOSIÇÃO DESTE ÍNDICE (DIFERENÇA ENTRE WS E LS). ($p > 0.05$)

Este último resultado, um paradoxo em se tratando do efeito de um agonista dopaminérgico, se esclarece quando consideramos a evolução das entropias de sequências de escolhas que não consideram o reforço. Aqui encontramos uma diminuição tanto de H_2 (teste de Wilcoxon para amostras pareadas, $W = 2$, $p = 0,0117$) como de H_3 (teste de Wilcoxon para amostras pareadas, $W=0$, $p = 0,0039$), mostrando que as frequências de blocos de lados escolhidos dois a dois (DD, ED, etc.) e três a três (DDE, EDE, etc.) se afastou mais da equiprobabilidade. Consistente com estes resultados observou-se, novamente, um aumento significativo da repetição do lado escolhido (teste de Wilcoxon para amostras pareadas, $W = 5$, $p = 0,0391$) (Figura 20).

Os animais também apresentaram uma tendência ao afastamento da equiprobabilidade da escolha, isto é, à diminuição de H_1 (teste de Wilcoxon para amostras pareadas, $W = 7$, $p = 0,0742$). Resumidamente, todos estes índices indicam comportamento de perseveração do lado escolhido. Na dose de 1,5 mg/kg encontramos um aumento do tempo de reação (teste de Wilcoxon para amostras pareadas, $W = 1$, $p = 0,0078$) e dos abandonos_2_4 (teste de Wilcoxon para amostras pareadas, $W = 1$, $p = 0,0078$).

FIGURA 20 – EFEITO DA METANFETAMINA NA DOSE DE 1.5 MG/KG SOBRE A REPETIÇÃO DO LADO ESCOLHIDO NA JOGADA SEGUINTE. (* $p < 0.05$)

Em conjunto, estes resultados mostram que os animais exibiram fundamentalmente um comportamento de escolha tendendo à estereotipia, repetindo frequentemente o lado escolhido, independentemente do reforço. Esta tendência já se mostrou de forma menos evidente sob efeito da dose de 1 mg/kg.

3.2.3 Síntese dos Resultados

A seguir apresentamos os resultados do Experimento II em forma de tabelas para facilitar uma visão de conjunto. Na Tabela 5 encontram-se apenas os dados relativos à estratégia de jogo dos animais e na Tabela 6 os dados referentes à emissão de respostas.

TABELA 5 – RESUMO DOS EFEITOS DAS DROGAS UTILIZADAS NO EXPERIMENTO II NA EMISSÃO DE RESPOSTA EM RELAÇÃO A SITUAÇÃO CONTROLE. EM AZUL: VALOR DE P QUANDO HOUVE AUMENTO SIGNIFICATIVO ($P < 0.05$) NO VALOR DA VARIÁVEL; EM VERMELHO: VALOR DE P QUANDO HOUVE DIMINUIÇÃO SIGNIFICATIVA ($P < 0.05$); E EM CINZA N.S.: VARIAÇÃO NÃO SIGNIFICATIVA ($P > 0.05$) (QUANDO A VARIAÇÃO SE APROXIMOU, MAS NÃO ATINGIU O NÍVEL DE SIGNIFICÂNCIA O VALOR DE P FOI INFORMADO, SEGUIDO DE + OU - PARA INDICAR CRESCIMENTO OU DIMINUIÇÃO NO VALOR DA VARIÁVEL EM RELAÇÃO À SITUAÇÃO CONTROLE).

DROGA p-valor	WSLS	Δ WS	H_1	H_2	H_3	M_1	M_2	M_3	% Rep
SKF 81297 0.3 MG/KG	0.0117	N.S.	N.S.	N.S.	N.S.	0.0039	0.0488	N.S.	N.S.
BROMOCRIPTINA 1.25 MG/KG	N.S.	N.S.	N.S.	N.S.	N.S.	N.S.	N.S.	N.S.	N.S.
BROMOCRIPTINA 2.5 MG/KG	N.S.	0.0488	0.547+	N.S.	N.S.	N.S.	N.S.	N.S.	0.0391
SCH 23390 0.001 MG/KG	0.0840+	N.S.	N.S.	0.0620+	N.S.	N.S.	N.S.	N.S.	N.S.
SCH 23390 0.004 MG/KG	N.S.	N.S.	N.S.	N.S.	N.S.	N.S.	N.S.	N.S.	N.S.
ETICLOPRIDA 0.01 MG/KG	N.S.	N.S.	N.S.	0.0195	N.S.	N.S.	N.S.	N.S.	N.S.
ETICLOPRIDA 0.04 MG/KG	N.S.	N.S.	N.S.	N.S.	N.S.	N.S.	N.S.	N.S.	N.S.
METANFETAMINA 1.0 MG/KG	N.S.	0.0371	N.S.	N.S.	0.0059	N.S.	N.S.	0.0020	0.0391
METANFETAMINA 1.5 MG/KG	0.0017	0.0391	0.0742-	0.0117	0.0039	N.S.	N.S.	0.0547-	0.0391

TABELA 6 – RESUMO DOS EFEITOS DAS DROGAS UTILIZADAS NO EXPERIMENTO II NA EMISSÃO DE RESPOSTA EM RELAÇÃO A SITUAÇÃO CONTROLE. EM AZUL: VALOR DE P QUANDO HOUVE AUMENTO SIGNIFICATIVO ($P < 0.05$) NO VALOR DA VARIÁVEL; EM VERMELHO: VALOR DE P QUANDO HOUVE DIMINUIÇÃO SIGNIFICATIVA ($P < 0.05$); E EM CINZA N.S.: VARIAÇÃO NÃO SIGNIFICATIVA ($P > 0.05$) (QUANDO A VARIAÇÃO SE APROXIMOU, MAS NÃO ATINGIU O NÍVEL DE SIGNIFICÂNCIA O VALOR DE P FOI INFORMADO, SEGUIDO DE + OU – PARA INDICAR CRESCIMENTO OU DIMINUIÇÃO NO VALOR DA VARIÁVEL EM RELAÇÃO À SITUAÇÃO CONTROLE). N_JOG; NÚMERO DE JOGADAS FEITAS NA SESSÃO, TR: TEMPO DE REAÇÃO, PREC: NÚMERO DE ERROS DE MOVIMENTOS PRECOCES, AB_1, AB_2_4 E AB_5: ÍNDICES DE ABANDONO DO JOGO DURANTE UMA, DUAS A QUATRO E MAIS DE CINCO JOGADAS.

DROGA p-valor	N_JOG	TR	PREC	Ab_1	Ab_2_4	Ab_5
SKF 81297 0.3 MG/KG	N.S.	N.S.	N.S.	N.S.	0.0078	N.S.
BROMOCRIPTINA 1.25 MG/KG	N.S.	0.0020	0.0098	N.S.	N.S.	N.S.
BROMOCRIPTINA 2.5 MG/KG	N.S.	0.0020	0.0098	N.S.	N.S.	0.0313
SCH 23390 0.001 MG/KG	N.S.	N.S.	N.S.	N.S.	N.S.	N.S.
SCH 23390 0.004 MG/KG	N.S.	0.0137	0.0059	0.0078	0.0625+	0.0625+
ETICLOPRIDA 0.01 MG/KG	N.S.	N.S.	N.S.	N.S.	N.S.	N.S.
ETICLOPRIDA 0.04 MG/KG	N.S.	N.S.	N.S.	N.S.	N.S.	N.S.
METANFETAMINA 1.0 MG/KG	N.S.	N.S.	N.S.	N.S.	N.S.	N.S.
METANFETAMINA 1.5 MG/KG	N.S.	0.0078	N.S.	0.0547+	0.0078	N.S.

3.2.3.1 Discussão

A demonstração, feita no experimento I, de que os animais utilizam um algoritmo de aprendizagem por reforço para uma aproximação da estratégia ótima no jogo da *matching pennies* foi o ponto de partida para este segundo experimento. Basicamente, neste modelo de aprendizagem para este jogo, o reforço obtido num dos lados aumenta a expectativa de reforço subsequente naquele lado vindo, desta forma, aumentar a probabilidade de escolha daquele lado, conforme modelo matemático descrito no tópico 1.6. Como é sabido, este aumento ou diminuição da expectativa de reforço ou seja, a atualização da função valor de um estado (neste caso o lado) ou de uma ação (no caso, escolher direita ou esquerda) é feito pelo sistema dopaminérgico através do sinal de erro de previsão de reforço

(δ na eq. (4)) transmitido pelas descargas fásicas dos neurônios dopaminérgicos do tronco encefálico (para revisão SCHULTZ, 2010). Nossa hipótese foi de que alterações produzidas pela estimulação ou bloqueio farmacológico dos receptores D1 e/ou D2 alterariam o equilíbrio entre a via direta e indireta e consequentemente a resposta do circuito córtico-estriato-tálamo-cortical. Também computações em áreas corticais pré-frontais são influenciadas pelo sinal de erro de previsão de reforço e, desta forma, seriam alteradas por, pelo menos, algumas destas drogas. Disto resultariam modificações no efeito que o reforço positivo ou negativo teria sobre as escolhas seguintes.

Os resultados obtidos confirmaram parcialmente nossa tese, mostrando, no entanto, um quadro mais complexo.

O agonista dos receptores dopaminérgicos D1, o SKF 81297, na dose de 0,3 mg/kg, produziu um aumento de WSLs em relação à situação-controle sem alterar a estrutura interna do índice, Δ WSLS (*i.e.* WS - LS). Estes dois resultados indicam que a droga levou a um aumento tanto da repetição da escolha do lado reforçado (*win-stay*, WS), como da mudança de lado após não receber reforço (*lose-switch*, LS). Este aumento da frequência de abandono do lado não reforçado é à primeira vista um dado não esperado, por ser dependente da atividade da via indireta, que não tem relação com receptores D1. No entanto, é importante lembrar, a estimulação dos receptores D1 não leva sempre a estimulação da via direta, mas funciona como um filtro aumentando a atividade das células mais ativas e diminuindo a atividade das células menos ativas (HERNÁNDEZ-LÓPEZ *et al.*, 1997). Assim, quando uma escolha não é seguida de reforço e, consequentemente, é seguida de uma queda da concentração dopaminérgica há uma menor ativação da via direta pela dopamina e a estimulação dos receptores D1 (feitas pelo SKF 81297 de modo tônico) se faz sobre células menos ativas, levando-as à inibição maior do que na situação controle. Esta maior inibição em relação à situação controle associada à liberação da via indireta pela queda da dopamina deslocará o equilíbrio em direção a via indireta, levando a um maior uso da estratégia de mudar de lado após não receber reforço (*lose-*

switch). Foi também observado um aumento da informação mútua entre a escolha do animal em uma jogada e as escolhas do animal e do computador na jogada anterior (MI_1) e nas duas anteriores (MI_2), indicando que os efeitos da droga se estendem a até duas escolhas retroativamente, não ultrapassando, porém, este limite, uma vez que o índice relativo às três últimas (MI_3) não exibiu variação significativa. Não foi observada tendência a perseverar no lado escolhido, mostrando que o elemento essencial neste caso é a escolha reforçada e não o lado em si. Este fato nos leva a concluir que o efeito do agonista dopaminérgico D1 se estende, nesta tarefa, até a segunda escolha anterior e seu resultado, não chegando a aumentar o efeito sobre o processo de escolha de jogadas mais distantes.

O bloqueio dos receptores D1 com SCH 23390 revelou uma tendência a aumento do uso de WSLs com a dose menor. Na dose maior não houve variação significativa.

Resumidamente, o que encontramos durante a vigência do efeito do SKF 81297, comparado à situação-controle, é um claro aumento do efeito, tanto da presença como da ausência, do reforço numa jogada sobre as escolhas seguintes. O antagonista D1 SCH 23390 produziu um efeito muito mais modesto, apenas uma tendência ao aumento de WSLs que não chega a atingir o nível de significância e sem nenhum efeito sobre os índices de informação mútua.

A seguir comentamos dados da literatura sobre o estriado e a sua modulação pelo sistema dopaminérgico com a finalidade de comparar outros resultados com os nossos. Hikida e colegas (2010), utilizando manipulação genética que permitiu bloqueio seletivo e reversível da via direta e indireta do circuito córtico-estriatal, estudaram as funções dos receptores estriatais D1 e D2, que se expressam segregadamente na via direta e indireta, respectivamente. Estes pesquisadores constataram que cada uma das vias tem uma função diferenciada e seletiva no comportamento diante de reforços e de estímulos aversivos. O bloqueio da via direta, na qual se expressam os receptores D1, causa déficits seletivos no aprendizado baseado em reforço, enquanto que o bloqueio da via indireta

compromete a formação rápida de memórias que permitem evitar estímulos aversivos. Goto e colegas (2005), estudando os efeitos diferenciais da estimulação dos receptores D1 e D2 no núcleo *accumbens*, mostraram que os receptores D1 estão na base do aprendizado por reforço ligado a aferentes provenientes da região hipocampal, enquanto que os receptores D2 estão ligados ao controle executivo mediado pelas projeções pré-frontais (aprendizado *versus set shifting* no estudo destes autores). Os efeitos do SKF 81297 em nosso experimento se colocam nesta mesma linha: a distribuição espacial (buraco direito *versus* buraco esquerdo) dos reforços na última e penúltima escolha passa a ter uma importância maior na estratégia dos animais. Outra característica importante da fisiologia dos receptores D1 é a sua resposta não linear à estimulação crescente. No caso das funções da memória operacional existe uma intensidade ótima de estimulação destes receptores; desvios tanto para mais como para menos produzem déficit no desempenho das tarefas. Vijayraghavan e colegas (2007) estudaram a atividade de neurônios pré-frontais individuais de macacos durante o desempenho de uma tarefa envolvendo resposta óculo-motora postergada (*oculomotor delayed response*). Aplicações iontoforéticas de agonistas dopaminérgicos D1, entre os quais SKF 81297, durante o desempenho da tarefa, geraram um efeito em forma de U invertido na sintonização (*tuning*) espacial dos neurônios: baixos níveis de estimulação do receptor D1 levavam a um aumento do *tuning* através da redução mais intensa do disparo neuronal para as direções não preferidas, enquanto que altos níveis de estimulação deste receptor suprimiam a atividade para todas as direções, eliminando desta maneira o *tuning* espacial. Se o SKF 81297 tiver um efeito semelhante sobre as representações dos lados e ações recentemente associadas ao recebimento de reforço que sobre o *tuning* espacial no experimento de Vijayraghavan e colegas (2007), então haveria um realce dos estados e ações recentemente reforçados em relação aos não reforçados ou ruído de fundo, que se traduziria em efeitos na probabilidade de escolha nas jogadas subsequentes. Infelizmente não dispomos dos resultados da dose maior desta droga na política de escolhas adotada pelos animais para uma discussão da linearidade ou não de seus efeitos nesta tarefa.

Os experimentos que registraram a atividade neural de macacos durante o jogo do *matching pennies* encontraram representações neurais sustentadas de variáveis importantes para o algoritmo de reforço, entre as quais os lados escolhidos recentemente e os reforços obtidos (BARRACLOUGH *et al.*, 2004; SEO *et al.*, 2007; SEO; LEE, 2009; SEO *et al.*, 2009; para revisão SEO; LEE, 2008). Esta atividade relacionada ao reforço não está presente apenas no córtex, mas também no estriado como mostra o estudo de Histed e colegas. (2009), utilizando outro tipo de tarefa. Estes autores, estudando a atividade neuronal individual em ambas as áreas, encontraram atividade sustentada e persistente dependente do resultado das respostas anteriores do animal e que esta atividade era importante na determinação da próxima ação do animal. Ambas as áreas recebem importante projeção dopaminérgica que nelas têm efeito semelhante (WILLIAMS; CASTNER, 2006). Estudos (GOTO; GRACE, 2005) sobre os efeitos da modulação dopaminérgica no núcleo *accumbens* mostram que os impulsos de origem no subículo ventral, transmitindo informações contextuais e relacionadas ao aprendizado relativo a reforços, são fortemente modulados por receptores dopaminérgicos D1. Este processo potencializa a transmissão vSub-Nac e assim o aprendizado. Desta maneira mostra-se que o *input* dopaminérgico realmente afeta a projeção vSub-Nac de uma maneira importante para o comportamento dirigido por reforço. No modelo proposto por Frank (para revisões COHEN; FRANK, 2009; HAZY *et al.*, 2007) a via direta, ou estriato-nigral, é fundamentalmente modulada pela dopamina através dos receptores D1 e, sob *input* dopaminérgico, que ocorre tanto no momento de reforço quanto no momento da percepção de preditores de reforço, aumenta a ativação de representações corticais e sua manutenção na memória operacional. Também neste modelo, a repetição da experiência de um preditor seguido de um reforço leva ao aprendizado através potencialização de longa duração (LTP) que ocorre nas conexões entre as representações de ambos na via direta.

Apesar de não ter sido robusto o suficiente para atingir o nível de significância nem se refletir em aumento dos índices de informação mútua, podendo desta forma ser apenas fruto do acaso, o aumento do uso de WSLs após a dose menor do antagonista D1 SCH 23390 merece uma discussão

pelo fato de nos permitir discutir melhor a complexidade do tema das alterações farmacológicas da modulação dopaminérgica de funções cognitivas. Apesar de que, como discutiremos mais adiante, a maior parte de nossos resultados serem interpretáveis dentro do quadro de referência da modulação dopaminérgica dos circuitos córtico-estriato-tálamo-corticais e ativação de representações de memória operacional no córtex pré-frontal, devemos também considerar a possibilidade de alterações de outras computações realizadas em várias regiões corticais. Desta forma, a atividade cortical envolvida na integração dos resultados de várias jogadas anteriores, a computação e atualização constante de um coeficiente de aprendizado e a transformação destas informações em uma escolha (para uma revisão SEO; LEE, 2008) terão provavelmente suas diversas etapas moduladas diferentemente pela dopamina. Como exemplo da modulação dopaminérgica diferenciada de diferentes áreas corticais, citamos o trabalho em que Winstanley e colegas (2006) estudaram a modulação dopaminérgica e serotoninérgica dos córtices pré-frontal medial e órbito-frontal durante o desempenho da tarefa de tomada de decisão envolvendo desvalorização pelo atraso (*delay discounting*) em ratos. Os níveis de dopamina e serotonina foram medidos nas duas regiões citadas usando microdiálise *in vivo*. Foram usados grupos-controle para examinar as variações do nível dos neuromoduladores em animais na situação de resposta instrumental e recebimento de reforço, sem, no entanto, realizar escolha intertemporal. Restringindo-nos à dopamina, os resultados mostraram que os níveis do metabólito dopaminérgico DOPAC aumentaram no córtex órbito-frontal dos animais que realizaram a tarefa de desconto temporal, mas não no OFC de animais do grupo-controle. Já os níveis de DOPAC no córtex pré-frontal medial aumentaram nos dois grupos. Estes dados sugerem um delicado e preciso balanço regional da modulação dopaminérgica, que também varia de acordo com a tarefa. A ruptura deste balanço regional, ocasionada pela aplicação sistêmica da droga em nosso experimento, pode ter levado a este resultado paradoxal de termos aumento dos efeitos da situação de reforço por um antagonista dopaminérgico. Esta hipótese, no entanto, precisa ser testada experimentalmente. Considerações semelhantes de especificidade regional se aplicam à inervação dopaminérgica do estriado. Embora do

ponto de vista de vista morfológico o estriado apresente uma estrutura bastante homogênea com um circuito básico que se repete como um motivo fundamental, funcionalmente as regiões estriatais estão envolvidas em tipos diferentes de comportamento motivado (para revisão YIN *et al.*, 2006a). De um modo geral podemos dizer que o estriado ventral está ligado à aquisição e expressão de certas respostas apetitivas pavlovianas e contribui para o controle motivacional do desempenho instrumental. O estriado dorsal é necessário para aquisição e expressão de ações instrumentais, sendo o estriado dorsomedial fundamental para ações controladas pelas contingências de resposta-resultado (*response-outcome*) e o estriado dorsolateral das ações habituais ou controladas pela contingência estímulo-resposta (*stimulus-response*). Estas três regiões apresentam diferenças importantes quanto à inervação dopaminérgica (WICKENS *et al.*, 2007). Embora as diversas regiões estriatais recebam fibras dopaminérgicas de diferentes grupos de células, parece que a característica do sinal é homogênea. No entanto, há um gradiente na densidade de inervação dopaminérgica, com a região dorsolateral recebendo a maior densidade que as regiões ventromediais. Estas diferenças têm efeitos importantes na precisão temporal do sinal, com regiões tendo alta densidade de inervação ocorrendo uma transmissão mais rápida do sinal. Outras diferenças existem neste campo, por exemplo, os terminais dopaminérgicos colibram glutamato no NAc, mas não nas outras regiões do estriado (STUBER *et al.*, 2010). Todos estes aspectos servem para ilustrar a complexidade da análise dos resultados obtidos. A tarefa de nosso experimento envolve o aprendizado por reforço sem modelos do ambiente, que é funcionalmente ligado ao estriado dorsolateral onde há maior densidade de inervação dopaminérgica. Deste modo, as drogas usadas tiveram a ação máxima nesta região de interesse.

O efeito mais importante causado pelo agonista D2, bromocriptina, é uma mudança da composição do índice de WSLS sem nenhuma variação na frequência de uso da estratégia de WSLS. Os animais aumentaram a frequência da mudança de lado após uma escolha não reforçada em detrimento da repetição de uma escolha que havia sido reforçada na última jogada, i.e. apresentaram uma diminuição no valor de Δ WSLS. Em outras palavras, o aprendizado principal do animal passa a

acontecer com a situação de não ser reforçado, ou seja, com o fracasso e não com a situação de obtenção de reforço. Estes resultados apresentam um paralelo com os obtidos por Franck e colegas (2004) estudando o desempenho de pacientes parkinsonianos numa tarefa de aprendizado probabilístico (Figura 7(c)). Os pacientes tiveram seu desempenho, tanto sob efeito de medicamentos como após a sua suspensão, comparado com o desempenho de controles normais. (Os pacientes foram testados na vigência dos efeitos de medicamentos e 24 horas após sua suspensão). A medicação antiparkinsoniana induz uma hiperdopaminergia leve, que impede as flutuações negativas do sinal fásico dopaminérgico e, conseqüentemente, interfere com a perfeita sinalização do sinal de erro negativo. Assim, os pacientes, sob efeito dos medicamentos, aprendiam mais com o reforço positivo do que com o não reforço e, desta maneira, apreendiam a escolher a melhor entre todas as opções, mas tinham dificuldade em aprender a evitar a pior delas. Já, quando testado sem medicação, e, portanto numa situação de hipodopaminergia, encontramos o efeito oposto. No que diz respeito a evitar a pior escolha, o aprendizado dos pacientes é também superior ao dos controles sadios. Neste estudo os autores interpretam que na relativa ausência de dopamina no estriado a flutuação negativa do sinal de erro será mais intensa, permitindo uma maior ativação da via indireta e, conseqüentemente maior aprendizado com o não reforço. Esta via tem sua atividade de aprendizado dependente dos receptores D2, de maneira que, quando há uma descarga dopaminérgica fásica, estes receptores são ativados e inibem os neurônios em cujas membranas residem. A repetição do processo leva à depressão de longa duração (LTD) na via indireta, que é inibitória. O resultado comportamental é que a ação que antecede este aumento da atividade dopaminérgica terá uma menor probabilidade de vir a ser inibida. Quando há uma queda da concentração de dopamina por ausência de um reforço esperado ocorre o processo inverso; liberada da inibição causada pela dopamina através da estimulação dos receptores D2, a atividade na via indireta permitirá potenciação de longa duração (LTP) através de aprendizado hebbiano. A consequência comportamental será que a ação que antecede a queda da concentração dopaminérgica terá a probabilidade de repetição diminuída, em outras palavras esta ação

será inibida (para revisão COHEN; FRANK, 2009). Da mesma maneira que em nosso experimento, encontramos uma maior sensibilidade à ausência do reforço de uma ação. No entanto, em nosso experimento, não tivemos uma depleção dopaminérgica, mas o uso de uma droga que causa estimulação dos receptores D2, que intuitivamente nos levaria a pensar num aumento da sensibilidade ao reforço, uma vez que causaria um aumento da inibição da via indireta, que por sua vez levaria a um deslocamento do equilíbrio entre as duas vias em direção da via direta. Mas a situação semelhante à hipodopaminergia pode ser explicada quando consideramos a fisiologia dos receptores D2 com mais detalhes. Estes receptores existem em duas isoformas moleculares distintas (USIELLO *et al.*, 2000; DE MEI *et al.*, 2009) que apresentam funções diferentes *in vivo*. Enquanto a forma longa (D2RL) tem ação fundamentalmente pós-sináptica, a forma curta (D2RS) tem ação de auto-receptor (WU *et al.*, 2002), participando do controle da liberação de dopamina, e mais fundamentalmente controlando a recaptação da dopamina através de sua ação sobre a proteína transportadora de dopamina (DE MEI *et al.*, 2009). Também se deve considerar a ação dos D2R sobre os interneurônios (LOVINGER; SURMEIER, 2006); sabe-se que estes receptores estão presentes nos interneurônios colinérgicos, onde inibem a liberação de ACh, responsável por aumento da excitabilidade dos MSN estriatais, tendo assim também importante função tanto na tomada de decisão a cada momento, como, com a repetição do processo, na formação de LTD (WANG *et al.*, 2006). O resultado comportamental de aprendizado maior do fracasso que do sucesso e seu paralelo com o estudo de Frank e colegas (2004) nos sugere que a estimulação dos D2R pré-sinápticos dos neurônios dopaminérgicos levou a um aumento da recaptação de dopamina induzindo assim uma hipodopaminergia farmacológica e consequentemente uma diminuição do sinal de erro positivo e um aumento do sinal de erro negativo (*dopamine dip*). No entanto, é importante ressaltar, no jogo do *matching pennies*, ao contrário do aprendizado na tarefa de Frank e colegas (2004), os efeitos dopaminérgicos nas sinapses córtico-estriatais devem ser fundamentalmente transitórios e de aumento e diminuição da excitabilidade da via direta e indireta e não de formação de LTP e LTD. Isto se deve ao fato que no *matching pennies* os reforços não tem

associação consistente com nenhum dos lados, ao passo que na tarefa de aprendizado probabilístico de Frank e colegas (2004) existem associações consistentes entre estímulos e reforços.

Com a eticloprida, antagonista dos receptores D2, obtivemos apenas uma variação significativa. Constatamos um aumento da H_2 , mostrando que a distribuição das sequências dos pares de lados escolhidos (e.g., DD, DE etc., sem considerar os reforços) foi mais homogênea. O efeito mostra uma maior independência das escolhas apenas da escolha do lado imediatamente anterior. À primeira vista, pensando a partir do modelo da depleção dopaminérgica (FRANK *et al.*, 2004), o esperado era encontrarmos mudança em outros índices, principalmente aqueles relacionados aos reforços obtidos anteriormente. Sendo a via indireta modulada pelos receptores D2, um bloqueio destes causaria uma liberação desta via, *i.e.*, esta via se tornaria mais ativa exercendo assim um efeito inibitório maior. Assim, quando de um reforço, a atividade da via direta, que levaria a uma maior probabilidade de repetir a escolha do lado associado ao reforço, entra em competição com a via indireta liberada pelo bloqueio do receptor. Como resultado, teríamos uma diminuição da importância do reforço na escolha seguinte em relação à condição-controle. No caso de um reforço negativo e consequente queda na concentração dopaminérgica do estriado o efeito do sinal de erro negativo seria ainda maior levando ao abandono do lado não reforçado com mais frequência. Mas como testamos nossos animais sob o efeito de agentes bloqueadores dos receptores D2 e não de depleção dopaminérgica, é importante comparar nossos resultados com os do trabalho de FRANK e O'REILLY (2006) em que voluntários humanos sadios foram testados na mesma tarefa após receberem drogas que afetam o sistema dopaminérgico. Neste trabalho voluntários sadios receberam doses pequenas de um antagonista D2 (haloperidol) e de um agonista D2 (cabergoline). O desempenho na tarefa foi semelhante ao dos pacientes parkinsonianos testados durante a vigência do efeito dos medicamentos para o haloperidol e semelhante ao desempenho dos pacientes testados após a suspensão da medicação para o cabergoline. Tal resultado é facilmente interpretável a partir da hipótese de que as baixas doses usadas agiram sobre os receptores pré-sinápticos e causaram ou aumento da liberação

de dopamina (haloperidol) ou diminuição (cabergoline). Os efeitos descritos acima para a bromocriptina se acomodam perfeitamente neste quadro interpretativo. No caso da eticloprida, um antagonista de receptores D2, a ausência de alterações nos índices que medem os efeitos do reforço nas escolhas posteriores pode ser interpretada, dentro deste quadro, como resultando de um bloqueio tanto dos receptores pré com pós-sinápticos. O bloqueio pré-sináptico levaria a um aumento da liberação de dopamina. No entanto, o bloqueio dos receptores D2 pós-sinápticos levaria a um aumento da atividade inibitória da via indireta que, desta forma, compensaria o aumento da atividade da via direta pela maior estimulação dos receptores D1. Esta explicação, no entanto, necessita de teste experimental em que os animais recebam doses crescentes de eticloprida, iniciando em doses inferiores a menor dose usada em nosso experimento. Com as doses mais baixas esperaríamos encontrar os efeitos comportamentais do bloqueio pré-sináptico (aumento da concentração de dopamina) seguido, nas doses maiores dos efeitos encontrados em nosso trabalho. Complementando esta discussão, não podemos afastar a hipótese de que, com as doses utilizadas, nossos resultados tenham origem também em alterações de computações em regiões corticais. Em nosso caso, as diferentes hipóteses sobre a origem dos efeitos observados (*estriatal versus cortical*) podem ser testadas experimentalmente com a observação do desempenho dos animais sob efeito de administração local no córtex e no estriado de diferentes doses. O estudo de Winter e colegas (2009), em que se comparou os efeitos do bloqueio dos receptores na flexibilidade comportamental, ilustra experimentalmente a complexidade inerente à interpretação de experimentos envolvendo agentes farmacológicos, onde ausência de efeito observado não implica ausência de ação fisiológica da substância estudada. Bloqueadores dos receptores D1 e D2 foram administrados sistemicamente e, localmente, no córtex órbita-frontal e pré-frontal medial. No caso do bloqueio dos receptores D1 com SCH 23390, o efeito do bloqueio sistêmico e local foi idêntico; no caso após a mudança de contingência, o animal a detecta, mas não consegue direcionar seu comportamento por ela, retorna a agir baseado na contingência anterior. Como era de se esperar (GOTO; GRACE, 2005), não faz o

aprendizado da condição de reforço, apesar de ser capaz de detectá-la. Já no caso do bloqueio dos receptores D2 com sulpiride, apenas a infusão local no mPFC produziu este efeito no desempenho da tarefa, levantando os autores à hipótese de que na infusão sistêmica o bloqueio D2 do OFC mascarou os efeitos deste bloqueio no mPFC. Outra hipótese a ser considerada na explicação do perfil dos efeitos encontrados em nosso experimento pode dever-se a questões de dose. Embora em outros experimentos de tomada de decisão (e.g. ZEEB *et al.*, 2009; ST ONGE; FLORESCO, 2009) as doses utilizadas tenham tido efeito importante poderia ser possível que neste tipo de tarefa fossem necessárias doses diferentes. Novamente, apenas a pesquisa experimental poderá resolver esta questão.

Os efeitos da dl-metanfetamina se concentraram fundamentalmente numa tendência à perseveração, *i.e.*, a repetir a escolha de um lado na próxima jogada independentemente de ter havido ou não reforço. Isto se manifesta por um aumento do índice de repetição do lado escolhido na jogada anterior bem como por um aumento de Δ WSLS que ocorre simultaneamente a uma diminuição da probabilidade de uso de WSLS. Tal como no caso da anfetamina, a metanfetamina causa um aumento de liberação de dopamina e noradrenalina (DOCHERTY, 2008; KRAEMER; MAURER, 2002) além de glutamato no núcleo *accumbens* e córtex pré-frontal (SCHOBLOCK *et al.*, 2003b). Os efeitos que encontramos têm paralelos nos resultados encontrados por Loh e colegas (1993). Estes autores, estudando os efeitos de várias drogas sobre a perseveração em ratos, encontraram que a anfetamina produziu um aumento significativo da tendência a repetir um determinado movimento do ponto de vista de coordenadas egocêntricas mesmo em doses baixas. Estes dados estão na mesma linha de nossos resultados, com a diferença de que neste estudo os animais sempre recebiam o reforço, independentemente da escolha ou estratégia de escolha utilizada. Em nosso experimento, como não houve nestas sessões com metanfetamina variação na probabilidade de reforço, podemos dizer que os animais tampouco eram penalizados pela estratégia de perseveração. Resultados semelhantes foram obtidos por estes autores com outras drogas de abuso (nicotina e heroína), mas não com outras drogas

psicotrópicas sem potencial aditivo (escopolamina e haloperidol). Também Zeeb e colegas (2009) estudando o desempenho de ratos numa versão do *Iowa Gambling Task* encontraram que a anfetamina produziu, entre outros efeitos, uma tendência à perseveração. As ações da anfetamina e metanfetamina são semelhantes, mas não idênticas. Schoblock *et al.* (2003a) constataram que apesar de curvas de desempenho semelhantes, as duas drogas tem ponto de melhor desempenho na tarefa com doses diferentes. Também encontraram uma maior tendência da metanfetamina à estereotipia.

Embora outros estudos utilizando psicoestimulantes tenham encontrado comportamento perseverativo em outras tarefas, não encontramos uma explicação dentro de nosso quadro de referência, pois encontramos, com uma substância que aumenta a liberação de dopamina, uma mudança de estratégia que reflete a perda de importância dos reforços nas escolhas posteriores. O fato do comportamento de escolha perseverativo ter sido encontrado em drogas sem ação direta sobre o sistema dopaminérgico, tais como nicotina e heroína (LOH, *et al.*, 1993) sugere que este tipo de comportamento pode ter sua origem nos efeitos da droga sobre outros circuitos, ou que os efeitos farmacológicos mais complexos que simples liberação de dopamina (DOCHERTY, 2008; KRAEMER; MAURER, 2002; SCHOBLOCK *et al.*, 2003b) levam a uma desorganização dos circuitos responsáveis pelo aprendizado por reforço.

Para finalizar é interessante fazer uma comparação geral dos resultados acima comentados com o único trabalho na literatura (ZEEB *et al.*, 2009) que envolve o desempenho de ratos em uma tarefa de jogo. No entanto, como já foi discutido, este jogo, uma versão para ratos do tradicional *Iowa Gambling Task* (IGT), não é um jogo estratégico, pois as respostas do oponente computadorizado não variam em função das opções do animal, como no *matching pennies*. Na versão para ratos do IGT, denominada *rat gambling task* (rGT), o animal tem que escolher entre quatro opções que fornecem reforços de magnitudes diferentes associados a punições de duração e probabilidade diferentes. Os animais têm um tempo limitado em cada sessão de jogo, de maneira que os períodos de punição (*time-out*) reduzem efetivamente a possibilidade de ganho total. As opções são compostas de maneira que

opções que fornecem os maiores reforços estão associadas às punições mais longas e com maior probabilidade. Desta maneira, optar consistentemente pela opção de maior reforço resultará num ganho líquido menor ao final da sessão. Também a opção associada à punição de menor duração fornece o menor reforço e, escolhida consistentemente, leva a um menor ganho total. Outra opção, em que a magnitude, duração e probabilidades estão na faixa intermediária, quando consistentemente escolhida, leva ao maior ganho líquido na sessão. Os animais receberam os mesmos agonistas e antagonistas dopaminérgicos que usamos em nosso trabalho, mas como agonista dopaminérgico inespecífico receberam d-anfetamina, e foram testados com três doses de cada um dos compostos. Além disto, receberam 8-OH-DPAT, um agonista dos receptores serotoninérgicos 5-HT_{1A}. Os ratos testados nesta tarefa descobrem a melhor estratégia, que consiste em escolher com maior frequência a opção associada ao maior ganho líquido. Mas quando comparamos os efeitos das drogas encontramos diferenças muito marcantes em relação aos efeitos no desempenho do *matching pennies*. Além de uma tendência à perseveração a anfetamina induz uma aversão ao risco, *i.e.*, os animais passam a escolher a opção de menor reforço associada à punição de menor duração com menor frequência, isto mesmo levando a um ganho menor a longo prazo. Os agonistas D1 e D2, bem como o antagonista D1, não produzem qualquer efeito. A eticloprida, antagonista D2, na dose mais baixa, leva a uma maior frequência na escolha da melhor opção. Em nosso experimento não há uma opção melhor que outra, mas uma estratégia ótima, que é a independência de cada escolha das anteriores e seus resultados. A eticloprida, em nosso experimento, não levou a qualquer alteração da estratégia de escolha. Ao contrário do rIGT, no *matching pennies* o SKF 81297 levou a uma queda do desempenho devido ao aumento significativo de vários índices que refletem a dependência das escolhas da história de escolhas e reforços anteriores. A metanfetamina, no *matching pennies*, levou a um comportamento perseverativo paralelo à perseveração causada pela anfetamina no rIGT. Em ambos os casos trata-se de uma deterioração da estratégia de escolha. Estas diferenças refletem os diferentes efeitos dos neuromoduladores em diferentes tarefas. As mesmas drogas usadas com variações da mesma tarefa

apresentam diferentes resultados. Assim, no trabalho de Zeeb e colegas (2009) havia também grupos-controle, nos quais o tempo ou a probabilidade de punição eram mantidos constantes em todas as opções. O que se verificou nestes grupos foi que houve mudanças no efeito das drogas; por exemplo, o SKF 81297, que não havia causado qualquer mudança significativa no desempenho do rIGT, na condição controle, em que a duração da punição era mantida constante para todas as quatro opções, aumentou a escolha da opção de maior reforço com maior probabilidade de punição (esta de duração igual em todas as opções), e isto apenas na dose maior. Entre outras coisas, estes resultados são indicativos de que o matching pennies, cujas alterações de estratégia puderam quase que totalmente ter uma explicação plausível dentro do quadro de referência da modulação dopaminérgica dos circuitos córtico-estriato-tálamo-corticais e manutenção de representações de memória operacional no córtex pré-frontal, exija uma atividade computacional mais simples. Isto também se reflete no fato de que uma rede neural bastante simples (SOLTANI *et al.*, 2006) tenha conseguido reproduzir a estratégia de jogo dos animais de Lee e colegas (2004). Esta hipótese necessita teste experimental que pode ser feito através da comparação dos efeitos destes fármacos utilizados aplicados no estriado com os efeitos dos mesmos fármacos com aplicação intra-cortical.

Quando consideramos as variações na emissão de respostas motoras causadas pelas drogas vemos um padrão coerente. Aumento no tempo de reação está associado ou à diminuição das respostas precoces ou aumento de algum índice de abandono da tarefa ou a ambos. Encontramos a associação destas três variáveis como consequência da hipodopaminergia induzida pela bromocriptina e do bloqueio dos receptores D1 pelo SCH 23390. Estes resultados são consistentes com a extensão dos modelos de aprendizagem por reforço (NIV, 2007) que atribui à concentração tônica da dopamina a codificação da média de reforços. Esta média de reforços determina o nível ótimo de resposta que é depende de um equilíbrio entre o custo de emissão da resposta rápida e o benefício da rápida aquisição de reforço. O bloqueio dos receptores D1 pelo SCH 23390 tem um paralelismo funcional com uma queda da concentração tônica da dopamina, pois causa uma diminuição continuada da

estimulação da via direta, semelhante a que ocorre com a queda da dopamina causada pela estimulação pré-sináptica dos receptores D2 pela bromocriptina. O dado isolado do aumento dos abandonos de 2 a 4 jogadas causado pelo SKF 81297 e os aumentos do tempo de reação e dos abandonos de 2 a 4 jogadas causados pela metanfetamina não são explicáveis dentro do quadro de referência da concentração tônica da dopamina a codificação da média de reforços (NIV, 2007). Aqui também os efeitos da metanfetamina diferem dos esperados dentro do quadro teórico da modulação dopaminérgica do aprendizado por reforço; isto é indicativo de que as ações desta droga, pelo menos nas doses que utilizamos e fora do quadro de uso crônico, impedem as computações necessárias para este tipo de aprendizagem.

3.3 Discussão Geral

3.3.1 Discussão

Os resultados dos experimentos acima descritos demonstram, pela primeira vez, que os roedores também são capazes de interação estratégica, pelo menos na forma mais simples exigida pelo jogo do *matching pennies*. Neste tipo de jogo, os roedores tiveram um desempenho muito próximo ao desempenho de macacos testados na mesma tarefa, isto apesar da forma diferente em que ocorreu o jogo (escolha visual sacádica para os macacos *versus* focinhada para os ratos) e das diferenças na intensidade de experiência com a tarefa (muito menor em nosso experimento). Também, como os macacos de Lee e colegas (2004), os ratos foram capazes de responder a mudanças de estratégia do oponente, adaptando sua estratégia no sentido de tentar impedir uma diminuição na probabilidade de reforços. A estimativa, por maximização de função de verossimilhança, dos parâmetros dos modelos matemáticos de aprendizagem por reforço revelou que o modelo que apresentou o melhor ajuste foi aquele em que os parâmetros eram o coeficiente de aprendizagem e valores diferenciados para

reforços positivos e negativos. Fato importante, este modelo, que prevê a integração contínua dos valores das ações/opções anteriores, teve um ajuste aos dados superior ao de um modelo mais simples que prevê apenas o efeito do resultado da última jogada na escolha seguinte. Também neste caso nossos resultados foram consistentes com outros estudos realizados com macacos (SEO; LEE, 2007b).

Estes resultados semelhantes, em espécies animais distintas e distantes do ponto de vista evolutivos, mostram a importância das interações estratégicas para a sobrevivência nos diversos ecossistemas. No enfrentamento entre predador e presa, a capacidade de produzir respostas não previstas pelo oponente aumenta a probabilidade de sobrevivência e, inversamente, a reduz na situação oposta. Na sociedade esta situação é exemplificada tanto pela interação entre polícia e criminoso, como pelo caso do goleiro e do artilheiro no momento do pênalti. Consequentemente a evolução selecionou aqueles que desenvolveram circuitos neurais capazes das computações necessárias para este tipo de comportamento (SANABRIA; THRAILKILL, 2009). Como este tipo de comportamento pode ser simulado em redes neurais com pequeno número de elementos e arquitetura simples (SOLTANI *et al.*, 2006), podemos supor que mesmo sistemas nervosos primitivos podem gerar tais respostas. Entre as características deste sistema simulado está um mecanismo que forneça um sinal de violação de expectativa de reforço que tenha a capacidade de alterar as conexões sinápticas entre representações de estados e ações. No modelo, o sinal de violação de expectativa evidentemente foi baseado no sistema dopaminérgico dos sistemas biológicos. Os vários dados da literatura implicando o sistema dopaminérgico no aprendizado por reforço (SCHULTZ, 2010; BROMBERG-MARTIN, 2010), que está na base da estratégia usada pelos animais para jogar o MP, nos levaram a explorar os efeitos de alterações deste sistema induzidas farmacologicamente.

Partindo da hipótese que agentes dopaminérgicos que alteram o equilíbrio entre as via direta e indireta, devem produzir a nível comportamental um aumento ou diminuição do efeito do reforço

sobre as escolhas seguintes alterando o valor de WSLs ou de Δ WSLS e de outros índices correlatos oriundos da teoria da informação, nossas hipóteses foram:

(1) que agonistas D1 aumentariam os efeitos dos reforços recebidos nas escolhas subsequentes, causando um maior uso da estratégia de repetir na próxima jogada a escolha reforçada que se manifestaria por um aumento de Δ WSLS; no caso dos antagonistas D1 sendo esperada a ausência ou uma diminuição ou deste efeito em relação à situação controle;

(2) que agonistas/antagonistas D2 teriam efeitos diferentes dependendo dos receptores atingidos; se a dose utilizada fosse pequena o suficiente para estimular/inibir apenas os receptores pré-sinápticos ou igualmente pré e pós-sinápticos (os efeitos esperados em cada caso serão discutidos na apresentação dos resultados)

(3) agonistas inespecíficos, que causam um aumento da liberação de dopamina o efeito esperado seria estimulação da via direta e inibição da via indireta tendo como efeito um aumento no valor de Δ WSLS.

Os efeitos encontrados, embora apresentando diferenças dos esperados, são bastante consistentes com a participação da modulação dopaminérgica na utilização de um algoritmo de reforço na atualização constante da estratégia de jogo. Encontramos um quadro mais complexo, em que além do aumento do uso da estratégia de WSLs sem alterações significativas em Δ WSLS causado pelo agonista D1, encontramos um desvio da estratégia em direção a evitar escolhas associadas à situação de não reforço causada pelo agonista D2. Este último efeito se deve à dupla função dos receptores D2 (USIELLO *et al.*, 2000; MEI *et al.*, 2009), pré e pós-sináptica, sendo que a dose utilizada produziu efeitos compatíveis com ação pré-sináptica na regulação da recaptção da dopamina. O efeito fisiológico do aumento da recaptção foi equivalente a uma hipodopaminérgica se traduzindo por um aprendizado maior diante da situação de não reforço, resultado semelhante ao de pacientes parkinsonianos não medicados (FRANK *et al.*, 2004). Os antagonistas D1 e D2 não apresentaram

efeitos sobre a estratégia de escolha em relação aos reforços anteriores, possivelmente em função de uma complexidade maior do que a suposta pelo nosso modelo de trabalho e/ou por termos utilizado doses em os efeitos pré e pós-sinápticos se anularam, ou que não foram ativas nesta tarefa, embora o fossem em outras. A metanfetamina causou um comportamento de perseveração das escolhas, *i.e.*, a estratégia passou a privilegiar a repetição da escolha anterior, independentemente do fato de ter havido reforço não. Este resultado consistente com outros estudos (*e. g.*, ZEEB *et al.*, 2009; SCHOBLOCK *et al.*, 2003a) pode ser devido a uma menor especificidade da metanfetamina pelo sistema dopaminérgico (KRAEMER; MAURER, 2002; MELEGA *et al.*, 1995)

A demonstração de que ratos podem ser usados como modelos em estudos de tomada de decisão estratégica, e de que este tipo de decisão está sob influência da modulação dopaminérgica das redes corticais pré-frontais e das alças córtico-estriato-tálamo-corticais, abre novas perspectivas de pesquisa. A literatura a respeito de tomada de decisão com roedores é imensa, o que permite colocar em perspectiva os dados e lançar sobre eles novas luzes, como por exemplo, foi o caso da comparação do MP com o rIGT (ZEEB *et al.*, 2009). Também contamos, para pesquisa com roedores, com um número muito grande de técnicas, desde as mais tradicionais, como lesões, até as mais recentes, como a utilização de animais transgênicos, que ainda não foram utilizadas em estudos de tomada de decisão estratégica e aguardam esta utilização. A isto se acrescenta que um modelo de comportamento com roedores permite toda uma gama de pesquisas, que por fatores tanto éticos como econômicos não podem ser com primatas.

Por outro lado, a demonstração dos efeitos da modulação dopaminérgica sobre a tomada de decisão estratégica amplia a gama de comportamentos dependentes da dopamina e aponta para novas possibilidades de pesquisa, tanto da neuroeconomia em geral como da neuroeconomia dos transtornos mentais. Por exemplo, estudos de aplicação tópica de fármacos permitiriam a abordagem da função da modulação dopaminérgica nas diversas áreas corticais e subcorticais durante a tomada de decisão estratégica, bem como a importante questão da contribuição do aprendizado por reforço

baseado em modelos (*model-based*) e sem modelos (*model-free*) (DAW *et al.*, 2005; DAYAN, 2009), na aprendizagem inicial e posteriormente na atualização gradual da estratégia no jogo do *matching pennies*. O incipiente campo da neuroeconomia dos transtornos neuropsiquiátricos, em que pela primeira vez podemos estudar de modo quantitativo interações interpessoais, comparar os desempenhos de diversos grupos patológicos entre si e com controles sadios e relacionar estes resultados comportamentais a dados de atividade neural (revisão: KISHIDA *et al.*, 2010), pode ser enriquecido por pesquisas utilizando o jogo do *matching pennies* em quadros que apresentam alterações da função dopaminérgica em sua fisiopatologia. Como exemplo, citamos o caso do TDAH, em que se poderia estudar as características de sensibilidade ao reforço na estratégia de jogo dos subtipos impulsivo-hiperativo, desatento e misto (AMERICAN PSYCHIATRIC ASSOCIATION, 2000) e como estas eventuais diferenças se relacionam a marcadores eletrofisiológicos (FRANK *et al.*, 2005). Outros quadros a serem abordados são o jogo patológico e as dependências de substâncias psicoativas, principalmente de psicoestimulantes.

O jogo do *matching pennies* pode também ser utilizado em estudos sobre os efeitos de polimorfismos genéticos presentes nos sistemas neuromoduladores, inclusive o dopaminérgico, cujos efeitos já foram estudados em outros tipos de tarefas, mas não em tarefas estratégicas (FRANK; FOSSELA, 2011).

4 CONCLUSÃO

Concluindo podemos resumir os resultados de nossa pesquisa nos seguintes tópicos:

- (1) A capacidade de interações estratégicas com outro agente inteligente, na disputa por um recurso, pelo menos na forma exigida por um jogo simples como o *matching pennies*, está também presente em roedores, que, em nosso experimento, foram capazes de desenvolver uma estratégia de jogo que se aproxima da estratégia normativa descrita pela teoria dos jogos. Este desempenho foi muito semelhante ao de macacos na mesma tarefa.
- (2) A estratégia de jogo no caso do *matching pennies* é modulada pelo sistema dopaminérgico. Os nossos resultados são consistentes com a atualização contínua do valor de cada lado (ou da ação de escolher o lado direito ou esquerdo) depois de cada jogada, através da utilização de um algoritmo de aprendizagem por reforço, em que o sinal de erro de previsão (ou de violação de expectativa) é transmitido pelo sistema dopaminérgico.
- (3) Estes resultados demonstram a possibilidade da utilização de roedores como modelo em estudos de tomada de decisão estratégica em neuroeconomia e, também, a importância do sistema dopaminérgico para este tipo de decisão.

REFERÊNCIAS

BARRACLOUGH, D. J.; CONROY, D.; LEE, D. Prefrontal cortex and decision making in a mixed-strategy game. **Nat. Neurosci**, n. 7, p. 404-410, 2004.

BASSO, M. A.; WURTZ, R. H. Modulation of neuronal activity in superior colliculus by changes in target probability. **J Neurosci**, n. 18, p. 7519-34, 1998.

BAYER, H. M.; LAU, B.; GLIMCHER, P. W. Statistics of midbrain dopamine neuron spike trains in the awake primate. **J. Neurophysiol**, n. 98, p. 1428-1439, 2007.

BECHARA, A.; DAMASIO, A. R.; DAMASIO, H.; ANDERSON, S. W. Insensitivity to future consequences following damage to human prefrontal cortex. **Cognition**, v. 50, n. 1-3, p. 7-15, 1994.

BICKEL, W. K.; YI, R.; KOWAL, B. P.; GATCHALIAN, K. M. Cigarette smokers discount past and future rewards symmetrically and more than controls: is discounting a measure of impulsivity? **Drug Alcohol Depend**, n. 96, p. 256-62, 2008.

BROMBERG-MARTIN, E. S.; MATSUMOTO, M.; HIKOSAKA, O. Dopamine in motivational control: rewarding, aversive, and alerting. **Neuron**, v. 68, n. 5, p. 815-34, 2010.

CAMERER, C. F. Behavioral Studies of Strategic Thinking in Games. **Trends Cogn. Sci.**, n. 7, p. 225-231, 2003.

CAMERER, C. F. Behavioral Game Theory and the Neural Basis of Strategic Choice. In: GLIMCHER, P. W.; CAMERER, C. F.; FEHR, E.; POLDRACK, R. A. **Neuroeconomics**: decision making and the brain. New York: Academic Press, 2009.

CARDINAL, R. N. Neural systems implicated in delayed and probabilistic reinforcement. **Neural Netw**, v. 19, n. 8, p. 1277-301, 2006.

CARTA, A.R.; TRONCI, E.; PINNA, A.; MORELLI, M. Different responsiveness of striatonigral and striatopallidal neurons to L-DOPA after a subchronic intermittent L-DOPA treatment. **Eur J Neurosci**, n. 21, v. 5, p. 1196-1204, 2005.

CISEK, P.; KALASKA, J. F. Neural mechanisms for interacting with a world full of action choices. **Annu Rev Neurosci**, n. 33, p. 269-98, 2010.

COHEN, M. X.; FRANK, M. J. Neurocomputational models of basal ganglia function in learning, memory and choice. **Behav Brain Res**, n. 199, p. 141-56, 2009.

COHEN, M. X.; RANGANATH, C. Reinforcement learning signals predict future decisions. **Neurosci**, v. 27, n. 2, p. 371-8, 2007.

CROMWELL, H. C.; SCHULTZ, W. Effects of expectations for different reward magnitudes on neuronal activity in primate striatum. **J Neurophysiol**, v. 89, n. 5, p. 2823-38, 2003.

CURTIS, C. E.; LEE, D. Beyond working memory: the role of persistent activity in decision making. **Trends Cogn Sci**, n. 14, p. 216-22, 2010.

DALLEY, J.W.; CARDINAL, R. N.; ROBBINS, T. W. Prefrontal executive and cognitive functions in rodents: neural and neurochemical substrates. **Neurosci Biobehav Rev** v. 28, n. 7, p. 771-784, 2004.

D'ARDENNE, K.; MCCLURE, S. M.; NYSTROM, L. E.; COHEN, J. D. BOLD responses reflecting dopaminergic signals in the human Ventral Tegmental Area. **Science**, n. 319, p. 1264-1267, 2008.

DAW, N. D.; NIV, Y.; DAYAN, P. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. **Nat Neurosci**, n. 12, p. 1704-11, 2005.

DAY, J. J.; ROITMAN, M. F.; WIGHTMAN, R. M.; CARELLI, R. M. Associative learning mediates dynamic shifts in dopamine signaling in the nucleus accumbens. **Nat Neurosci**, n. 8, p. 1020-1028, 2007.

DAYAN, P. Goal-directed control and its antipodes. **Neural Netw**, n. 22, v. 3, p. 213-219, 2009

DAYAN, P.; NIV Y. Reinforcement learning: the good, the bad and the ugly. **Curr Opin Neurobiol**, n. 18, v. 2, p. 185-196, 2008.

DE MEI, C.; RAMOS, M.; IITAKA, C.; BORELLI, E. Getting specialized: presynaptic and postsynaptic dopamine D2 receptors. **Curr Opin Pharmacol**, n. 9, p. 53-58, 2009.

den OUDEN, H. E.; DAUNIZEAU, J.; ROISER, J.; FRISTON, K. J.; STEPHAN, K. E. Striatal prediction error modulates cortical coupling. **J Neurosci**, n. 30, p. 3210-9, 2010.

DOCHERTY, J.R. Pharmacology of stimulants prohibited by the World Anti-Doping Agency (WADA). **Br J Pharmacol**, n. 154, v. 3, p. 606-622, 2008.

DOYA, K. Reinforcement learning: Computational theory and biological mechanisms. **HFSP J**, n. 1, p. 30-40, 2007.

FLORESCO, S. B.; MAGYAR, O. Mesocortical dopamine modulation of executive functions: beyond working memory. **Psychopharmacology (Berl)**, v. 188, n. 4, p. 567-85, 2006.

FRANK, M. J. Dynamic dopamine modulation in the basal ganglia: a neurocomputational account of cognitive deficits in medicated and nonmedicated Parkinsonism. **J Cogn Neurosci**, n. 17, p. 51-72, 2005.

_____. Hold your horses: a dynamic computational role for the subthalamic nucleus in decision making. **Neural Netw**, n. 19, p. 1120-36, 2006.

FRANK, M. J.; FOSSELLA, J. A. Neurogenetics and Pharmacology of Learning, Motivation, and Cognition. **Neuropsychopharmacol**, n. 36, p. 133-152, 2011.

FRANK, M. J.; LOUGHRY, B.; O'REILLY, R. C. Interactions between frontal cortex and basal ganglia in working memory: a computational model. **Cogn Affect Behav Neurosci**, v. 1, n. 2, p. 137-60, 2001.

FRANK, M. J.; O'REILLY, R. C. A mechanistic account of striatal dopamine function in human cognition: psychopharmacological studies with cabergoline and haloperidol. **Behav Neurosci** v.120, n.3, p. 497-517, 2006.

FRANK, M. J.; SAMANTA, J.; MOUSTAFA, A. A.; SHERMAN, S. J. Hold your horses: impulsivity, deep brain stimulation, and medication in parkinsonism. **Science**, n. 318, p. 1309-12, 2007a.

FRANK, M. J.; SANTAMARIA, A.; O'REILLY, R. C.; WILLCUTT, E. Testing computational models of dopamine and noradrenaline dysfunction in attention deficit/hyperactivity disorder. **Neuropsychopharmacology** n. 32, v. 7, p. 1583-1599, 2007b.

FRANK, M. J.; SCHERES, A.; SHERMAN, S. J. Understanding decision-making deficits in neurological conditions: insights from models of natural action selection. **Philos Trans R Soc Lond B Biol Sci.** n. 362, p. 1641-1654, 2007c.

FRANK, M. J.; SEEBERGER, L. C.; REILLY, R. C. By Carrot or by Stick: Cognitive Reinforcement Learning in Parkinsonism. **Science**, n. 306, p. 1940-1943, 2004.

FRISTON, K. The free-energy principle: a rough guide to the brain? **Trends Cogn Sci**, n. 13, p. 293-301, 2009.

GLIMCHER, P. W. **Decisions, Uncertainty and the Brain**: the science of neuroeconomics. Cambridge: The MIT Press, 2003a.

_____. The neurobiology of visual-saccadic decision making. **Annu. Rev. Neurosci**, n. 26, p. 133-179, 2003b.

GLIMCHER, P. W.; CAMERER, C. F.; FEHR, E.; POLDRACK, R. A. Introduction: A Brief History of Neuroeconomics. In: _____. **Neuroeconomics**: decision making and the brain. New York: Academic Press, 2009.

GOTO, Y.; GRACE, A. A. Dopaminergic modulation of limbic and cortical drive of nucleus accumbens in goal-directed behavior. **Nature Neuroscience**, n. 8, p. 805-812, 2005.

GURNEY, K.; PRESCOTT, T. J.; REDGRAVE, P. A computational model of action selection in the basal ganglia. I. A new functional anatomy. **Biol Cybern**, n. 84, p. 401-10, 2001.

HAZY, T. E.; FRANK, M. J.; O'REILLY, R. C. Banishing the homunculus: making working memory work. **Neuroscience**, n. 139, p. 105-118, 2006.

_____. Towards an executive without a homunculus: computational models of the prefrontal cortex/basal ganglia system. **Philos Trans R Soc Lond B Biol Sci**, n. 362, p. 1601-13, 2007.

HERNÁNDEZ-LÓPEZ, S.; BARGAS, J.; SURMEIER, D. J.; REYES, A.; GALARRAGA, E. D1 receptor activation enhances evoked discharge in neostriatal medium spiny neurons by modulating an L-type Ca^{2+} conductance. **J Neurosci** n. 17, v. 9, p. 3334-3342, 1997.

HIKIDA, T.; KIMURA, K.; WADA, N.; FUNABIKI, K.; NAKANISHI, S. Distinct roles of synaptic transmission in direct and indirect striatal pathways to reward and aversive behavior. **Neuron**, v. 66, n. 6, p. 896-907, 2010.

HISTED, M. H.; PASUPATHY, A.; MILLER, E. K. Learning substrates in the primate prefrontal cortex and striatum: sustained activity related to successful actions. **Neuron**, 63, p. 244-53, 2009.

HOUK, J. C. Agents of the mind. **Biol Cybern**, n. 92, p. 427-37, 2005.

HOUK, J. C.; BASTIANEN, C.; FANSLER, D.; FISHBACH, A.; FRASER, D.; REBER, P. J.; ROY, S. A.; SIMO, L. S. Action selection and refinement in subcortical loops through basal ganglia and cerebellum. **Philos Trans R Soc Lond B Biol Sci**, n. 362, p. 1573-83, 2007.

KABLE, J. W.; GLIMCHER, P. W. The neural correlates of subjective value during intertemporal choice. **Nat. Neurosci**, n. 10, p. 1625-1633, 2007.

_____. The Neurobiology of Decision: Consensus and Controversy. **Neuron**, n. 63, p. 733-745, 2009.

KING-CASAS, B.; CAMERER, C. F.; QUARTZ, S. R.; MONTAGUE, P. R. Getting to Know You: Reputation and Trust in a Two-Person Economic Exchange. **Science**, n. 308, p. 78-82, 2005.

KING-CASAS, B.; SHARP, C.; LOMAX-BREAM, L.; LOHRENZ, T.; FONAGY, P.; MONTAGUE, P. R. The rupture and repair of cooperation in borderline personality disorder. **Science**, n. 321, p. 806-10, 2008.

KISHIDA, K. T.; KING-CASAS, B.; MONTAGUE, P. R. Neuroeconomic approaches to mental disorders. **Neuron** v. 67, n. 4, p. 543-554, 2010.

KRAEMER, T.; MAURER, H. H. Toxicokinetics of amphetamines: metabolism and toxicokinetic data of designer drugs, amphetamine, methamphetamine, and their N-alkyl derivatives. **Ther Drug Monit**, n. 24, v. 2, p. 277-289, 2002.

KRUEGER, F.; GRAFMAN, J.; MCCABE, K. Neural correlates of economic game playing. **Philos Trans R Soc Lond B Biol Sci**, v. 363, n. 1511, p. 3859-74, 2008.

LAU, B.; GLIMCHER, P. W. Value representations in the primate striatum during matching behavior. **Neuron**, n. 58, p. 451-463, 2008.

LEE, D. Game theory and neural basis of social decision making. **Nat Neurosci**, n. 11, p. 404-9, 2008.

LEE, D.; CONROY, M. L.; MCGREEVY, B. P.; BARRACLOUGH, D. J. Reinforcement learning and decision making in monkeys during a competitive game. **Cogn. Brain Res**, n. 22, p. 45-58, 2004.

LEE, D.; SEO, H. Mechanisms of reinforcement learning and decision making in the primate dorsolateral prefrontal cortex. **Ann N Y Acad Sci**, n. 1104, p. 108-22, May 2007.

LOH, E. A.; SMITH, A. M.; ROBERTS, D. C. Evaluation of response perseveration of rats in the radial arm maze following reinforcing and nonreinforcing drugs. **Pharmacol Biochem Behav**, n. 44, p. 735-40, 1993.

LOVINGER, D. M.; SURMEIER, D. J. Dopaminergic control of corticostriatal long-term synaptic depression in medium spiny neurons is mediated by cholinergic interneurons. **Neuron**, n. 50, p. 443-452, 2006.

MAIA, T. V. Reinforcement learning, conditioning, and the brain: Successes and challenges. **Cogn Affect Behav Neurosci**, n. 9, p. 343-64, 2009.

MAIA, T. V.; FRANK, M. J. From reinforcement learning models to psychiatric and neurological disorders. **Nat Neurosci**, v. 14, n. 2, p. 154-62, 2011.

MEI, C. D.; RAMOS, M.; IITAKA, C.; BORRELLI, E. Getting specialized: presynaptic and postsynaptic dopamine D2 receptors. **Current Opinion in Pharmacology**, n. 9, p. 53-58, 2009.

MELEGA, W. P.; WILLIAMS, A. E.; SCHMITZ, D. A.; DISTEFANO, E. W.; CHO, A. K. Pharmacokinetic and pharmacodynamic analysis of the actions of d-amphetamine and d-methamphetamine on the dopamine terminal. **J Pharmacol Exp Ther**, n. 274, p. 90-96, 1995.

MILLER, G. A. Note on the bias of information estimates. In: QUASTLER, H. (Ed.). **Information Theory in Psychology**. Glencoe: Free Press, 1955. p. 95-100.

MONTAGUE, P. R.; LOHRENTZ T. To detect and correct: norm violations and their enforcement. **Neuron**, n. 56, p. 14-8, 2007.

NASH, J. F. Equilibrium points in N-person games. **Proc. Nat. Acad. Sci. USA**, n. 36, p. 48-49, 1950.

NIV, Y. Cost, benefit, tonic, phasic: what do response rates tell us about dopamine and motivation? **Ann N Y Acad Sci** n. 1104, p. 357-376, 2007.

NIV, Y. Reinforcement learning in the brain. **J Math. Psychol**, n. 53, p. 139-154, 2009.

NIV, Y; MONTAGUE, R. Theoretical and Empirical Studies of Learning. In: GLIMCHER, P. W.; CAMERER, C. F.; FEHR, E.; POLDRACK, R. A. **Neuroeconomics**: decision making and the brain. New York: Academic Press, 2009.

NOWAK, M. A.; NOWAK, M. A.; PAGE, K. M.; SIGMUND, K. Fairness Versus Reason in the Ultimatum Game. **Science**, n. 289, p. 1773-1775, 2000.

O'DOHERTY, J. P.; DAYAN, P.; FRISTON, K.; CRITCHLEY, H.; DOLAN, R. J. Temporal difference models and reward-related learning in the human brain. **Neuron**, n. 38, p. 329-37, 2003.

O'DOHERTY, J.; DAYAN, P.; SCHULTZ, J.; DEICHMANN, R.; FRISTON, K.; DOLAN, R. J. Dissociable roles of ventral and dorsal striatum in instrumental conditioning. **Science**, n. 30, p. 452-4, 2004.

O'REILLY, R. C.; FRANK, M. J. Making working memory work: a computational model of learning in the prefrontal cortex and basal ganglia. **Neural Comput**, v. 18, n. 2, p. 283-328, 2006.

O'REILLY, R. C.; FRANK, M. J.; HAZY, T. E.; WATZ, B. PVLV: the primary value and learned value Pavlovian learning algorithm. **Behav Neurosci**, n. 121, p. 31-49, 2007.

PADOA-SCHIOPPA, C.; ASSAD, J. A. Neurons in the orbitofrontal cortex encode economic value. **Nature**, n. 441, p. 223-226, 2006.

_____. The representation of economic value in the orbitofrontal cortex is invariant for changes of menu. **Nat Neurosci**, n. 1, p. 95-102, 2008.

PAULUS, M. P.; LOVERO, K. L.; WITTMANN, M.; LELAND, D. S. Reduced behavioral and neural activation in stimulant users to different error rates during decision making. **Biol Psychiatry**, n. 63, p. 1054-60, 2008.

PAWITAN, Y. **In all likelihood**: statistical modelling and inference using likelihood. Oxford: Oxford University Press, 2001.

PESSIGLIONE, M.; SEYMOUR, B.; FLANDIN, G.; DOLAN, R. J.; FRITH, C. D. Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. **Nature**, n. 442, p. 1042-5, 2006.

PLATT, M. L.; GLIMCHER, P. W. Neural correlates of decision variables in parietal cortex. **Nature**, n. 400, p. 233-238, 1999.

PRESCOTT, T. J.; MONTES GONZÁLEZ, F. M.; GURNEY, K.; HUMPHRIES, M. D.; REDGRAVE, P. A robot model of the basal ganglia: behavior and intrinsic processing. **Neural Netw**, v. 19, n. 1, p. 31-61, 2006.

QUERVAIN, D. J.; FISCHBACHER, U.; TREYER, V.; SCHELLHAMMER, M.; SCHNYDER, U.; BUCK, A.; FEHR, E. The Neural Basis of Altruistic. **Science**, n. 305, p. 1254-1258, 2004.

R DEVELOPMENT CORE TEAM. **R: A language and environment for statistical computing**. R Foundation for Statistical Computing, Vienna, Austria, 2011. ISBN 3-900051-07-0, URL: <<http://www.R-project.org/>>.

RAPOPORT, A.; BUDESCU, D. V. Randomization in individual choice behavior. **Psychol Rev**, n. 104, p. 603-617, 1997.

REDGRAVE, P.; GURNEY, K.; REYNOLDS, J. What is reinforced by phasic dopamine signals? **Brain Res Rev**, n. 58, p. 322-39, 2008.

RESCORLA, R. A.; WAGNER, A. R. A theory of Pavlovian conditioning: variations in the effectiveness of reinforcement and nonreinforcement. In: BLACK, A. H.; PROKASY, W. F. (Eds.). **Classical Conditioning II: Current Research and Theory**. New York: Appleton-Century-Crofts, 1972. p. 64-99.

RILLING, J. K.; GUTMAN, D. A.; ZEH, T. R.; PAGNONI, G.; BERNS, G. S.; KILTS, C. D. Neural Basis for Social Cooperation. **Neuron**, n. 35, p. 395-405, 2002.

ROUGIER, N. P.; NOELLE, D. C.; BRAVER, T. S.; COHEN, J. D.; O'REILLY, R. C. Prefrontal cortex and flexible cognitive control: rules without symbols. **Proc Natl Acad Sci U S A.**, n. 102, p. 7338-43, 2005.

SAMEJIMA, K.; UEDA, Y.; DOYA, K.; KIMURA, M. Representation of action-specific reward values in the striatum. **Science**, n. 310, p. 1337-1340, 2005.

SANABRIA, F.; THRAILKILL, E. Pigeons (*Columba livia*) approach Nash equilibrium in experimental Matching Pennies competitions. **J Exp Anal Behav**, v. 91, n. 2, p. 169-83, 2009.

SANFEY, A. G. Social decision-making: insights from game theory and neuroscience. **Science**, n. 318, p. 598-602, 2007.

SANFEY, A. G.; RILLING, J. K.; ARONSON, J. A.; NYSTROM, L. E.; COHEN, J. D. The neural basis of economic decision making in the Ultimatum game. **Science**, n. 300, p. 1755-1758, 2003.

SCHÖNBERG, T.; DAW, N. D.; JOEL, D.; O'DOHERTY, J. P. Reinforcement learning signals in the human striatum distinguish learners from nonlearners during reward-based decision making. **J Neurosci**, n. 27, p. 12860-7, 2007.

SCHULTZ, W. Dopamine signals for reward value and risk: basic and recent data. **Behav Brain Funct**, n. 23, p. 6-24, 2010.

SCHULTZ, W.; DAYAN, P.; MONTAGUE, P. R. A neural substrate of prediction and reward. **Science**, n. 275, p. 1593-9, 1997.

SEO, H.; BARRACLOUGH, D. J.; LEE, D. Dynamic signals related to choices and outcomes in the dorsolateral prefrontal cortex. **Cerebral Cortex**, n. 17, p. i110-i117, 2007.

_____. Lateral intraparietal cortex and reinforcement learning during a mixed-strategy game. **J. Neurosci**, n. 29, p. 7278-7289, 2009.

SEO, H.; LEE, D. Behavioral and neural changes after gains and losses of conditioned reinforcers. **Journal of Neuroscience**, n. 29, p. 3627-3641, 2009.

_____. Cortical mechanisms for reinforcement learning in competitive games. **Philos Trans R Soc Lond B Biol Sci**, v. 363, n. 1511, p. 3845-57, 2008.

_____. Temporal filtering of reward signals in the dorsal anterior cingulate cortex during a mixed-strategy game. **Journal of Neurosci**, n. 27, p. 8366-8377, 2007.

SHOBLOCK, J. R.; MAISONNEUVE, I. M.; GLICK, S. D. Differences between d-methamphetamine and d-amphetamine in rats: working memory, tolerance, and extinction. **Psychopharmacology** (Berl), n. 170, p. 150-156, 2003a.

SHOBLOCK, J. R.; SULLIVAN, E. B.; MAISONNEUVE, I. M.; GLICK, S. D. Neurochemical and behavioral differences between d-methamphetamine and d-amphetamine in rats. **Psychopharmacology** (Berl), n. 165, p. 359-369, 2003b.

SOLTANI, A.; LEE, D.; WANG, X. J. Neural mechanism for stochastic behaviour during a competitive game. **Neural Netw**, n. 19, p. 1075-90, 2006.

ST ONGE, J. R.; FLORESCO, S. B. Dopaminergic Modulation of Risk-Based Decision Making. **Neuropsychopharmacol**, n. 34, p. 681-697, 2009.

STUBER, G. D.; HNASKO, T. S.; BRITT, J. P.; EDWARDS, R. H.; BONCI, A. Dopaminergic Terminals in the Nucleus Accumbens But Not the Dorsal Striatum Corelease Glutamate. **Journal of Neuroscience**, n. 30, p. 8229-8233, 2010.

SUTTON, R. S.; BARTO, A. G. **Reinforcement Learning**: An introduction. Boston: MIT Press, 1998.

USHER, M.; COHEN, J. D.; SERVAN-SCHREIBER, D.; RAJKOWSKI, J.; ASTON-JONES, G. The role of locus coeruleus in the regulation of cognitive performance. **Science**, n. 283, p. 549-54, 1999.

USIELLO, A.; BAIK, J.; PICETTI, R.; DIERICH, Â. Distinct functions of the two isoforms of dopamine D2 receptors. **Nature**, n. 408, p. 199-203, 2000.

VIJAYRAGHAVAN, S.; WANG, M.; BIRNBAUM, S. G.; WILLIAMS, G. V.; ARNSTEN, A. F. Inverted-U dopamine D1 receptor actions on prefrontal neurons engaged in working memory. **Nat Neurosci**, n. 10, p. 376-84, 2007.

WANG, Z.; KAI, L.; DAY, M.; RONESI, J.; YIN, H. H.; DING, J.; TKATCH, T.; LOVINGER, D. M.; SURMEIER, D. J. Dopaminergic control of corticostriatal long-term synaptic depression in medium spiny neurons is mediated by cholinergic interneurons. **Neuron**, n. 50, p. 443-452, 2006.

WICKENS, J. R.; BUDD, C. S.; HYLAND, B. I. Striatal Contributions to Reward and Decision Making Making Sense of Regional Variations. **Ann. N.Y. Acad Sci**, n. 1104, p. 192-212, 2007.

WILLIAMS, G. V.; CASTNER, S. A. Under the curve: critical issues for elucidating D1 receptor function in working memory. **Neuroscience**, n. 139, p. 263-76, 2006.

WINSTANLEY, C. A.; THEOBALD, D. E.; DALLEY, J. W.; CARDINAL, R. N.; ROBBINS, T. W. Double Dissociation between Serotonergic and Dopaminergic Modulation of Medial Prefrontal and Orbitofrontal Cortex during a Test of Impulsive Choice. **Cerebral Cortex**, n. 16, p. 106-114, 2006.

WINTER, S.; DIECKMANN, M.; SCHWABE, K. Dopamine in the prefrontal cortex regulates rats behavioral flexibility to changing reward value. **Behavioural Brain Research**, n. 198, p. 206-213, 2009.

WU, Q.; REITH, M. E.; WALKER, Q. D.; KUHN, C. M.; CARROLL, F. I.; GARRIS, P. A. Concurrent Autoreceptor-Mediated Control of Dopamine Release and Uptake during Neurotransmission: An In Vivo Voltammetric Study. **Journal of Neuroscience**, n. 22, p. 6272-6281, 2002.

YIN, H. H.; KNOWLTON, B. J. The role of the basal ganglia in habit formation. **Nature Reviews Neuroscience**, n. 7, p. 464-476, 2006a.

YIN, H. H.; KNOWLTON, B. J.; BALLEINE, B. W. Inactivation of dorsolateral striatum enhances sensitivity to changes in the action-outcome contingency in instrumental conditioning. **Behav Brain Res**, n. 166, p. 189-96, 2006b.

YIN, H. H.; OSTLUND, S. B.; KNOWLTON, B. J.; BALLEINE, B. W. The role of the dorsomedial striatum in instrumental conditioning. **Eur J Neurosci**, n. 22, p. 513-23, 2005.

ZEEB, F. D.; ROBBINS, T. W.; WINSTANLEY, C. A. Serotonergic and Dopaminergic Modulation of Gambling Behavior as Assessed Using a Novel Rat Gambling Task. **Neuropsychopharmacology**, n. 34, p. 2329-2343, 2009.