VALDINEI FREIRE DA SILVA

EXTRAÇÃO DE PREFERÊNCIAS POR MEIO DE AVALIAÇÕES DE COMPORTAMENTOS OBSERVADOS

Tese apresentada à Escola Politécnica da Universidade de São Paulo para obtenção do título de Doutor em Engenharia Elétrica.

São Paulo
2009

VALDINEI FREIRE DA SILVA

EXTRAÇÃO DE PREFERÊNCIAS POR MEIO DE AVALIAÇÕES DE COMPORTAMENTOS OBSERVADOS

Tese apresentada à Escola Politécnica da Universidade de São Paulo para obtenção do título de Doutor em Engenharia Elétrica.

Orientador:
Profa. Livre-Docente Anna Helena Reali Costa

Co-orientador:
Prof. Associado Pedro U. Lima

São Paulo
2009

FICHA CATALOGRÁFICA

Silva, Valdinei Freire da

Extração de preferências por meio de avaliações de comportamentos observados / V.F. da Silva.– São Paulo, 2009. 140 p.

Tese (Doutorado) — Escola Politécnica da Universidade de São Paulo. Departamento de Engenharia de Computação e Sistemas Digitais.

1. Inteligência artificial 2. Aprendizado computacional 3. Teoria da Decisão 4. Processos de Markov. I. Universidade de São Paulo. Escola Politécnica. Departamento de Engenharia de Computação e Sistemas Digitais. II. t.

AGRADECIMENTOS

Agradeço a todos que estiveram perto de mim durante a realização desta tese.

Minha família, que sempre apoiou minhas decisões, mesmo não as compreendendo por completo.

Meus amigos de laboratório e agregados, com os quais passei a fazer parte de uma nova família, na qual compartilhamos as preocupações e alegria de cada tese ou dissertação defendida: Alex, Anarosa, André, Antonio, Bianchi, Diana, Esther, Graça, Guillaume, Gustavo, Humberto, Inácio, Jaime, Jomi, Leandro, Lucas, Luciano, Marcelo, Nicolau, Priscilla, Regina, Sara e Valguima. Com eles amadureci o meu trabalho, mas sobre tudo, amadureci como pessoa.

Ao orientador e amigo, Professor Pedro Lima, que me recebeu em terras estrangeiras de forma acolhedora e possibilitou uma experiência especial na minha vida pessoal e como pesquisador.

À minha orientadora, amiga e segunda mãe, Professora Anna. Assim como meus pais me guiaram e me guiam na minha vida pessoal, ela me guiou e, espero, será presença constante na minha vida como pesquisador.

Este trabalho não teria sido possível sem o apoio da FAPESP (processo número 02/13678-0), da CAPES (processo BEX 3388/04-2) e do Programa de Mobilidade Santander.

RESUMO

Recentemente, várias tarefas tem sido delegadas a sistemas computacionais, principalmente quando sistemas computacionais são mais confiáveis ou quando as tarefas não são adequadas para seres humanos. O uso de extração de preferências ajuda a realizar a delegação, permitindo que mesmo pessoas leigas possam programar facilmente um sistema computacional com suas preferências. As preferências de uma pessoa são obtidas por meio de respostas para questões específicas, que são formuladas pelo próprio sistema computacional. A pessoa age como um usuário do sistema computacional, enquanto este é visto como um agente que age no lugar da pessoa. A estrutura e contexto das questões são apontadas como fonte de variações das respostas do usuário, e tais variações podem impossibilitar a factibilidade da extração de preferências. Um forma de evitar tais variações é questionar um usuário sobre a sua preferência entre dois comportamentos observados por ele. A questão de avaliar relativamente comportamentos observados é mais simples e transparente ao usuário, diminuindo as possíveis variações, mas pode não ser fácil para o agente interpretar tais avaliações. Se existem divergências entre as percepções do agente e do usuário, o agente pode ficar impossibilitado de aprender as preferências do usuário. As avaliações são geradas com base nas percepções do usuário, mas tudo que um agente pode fazer é relacionar tais avaliações às suas próprias percepções. Um outro problema é que questões, que são expostas ao usuário por meio de comportamentos demonstrados, são agora restritas pela dinâmica do ambiente e um comportamento não pode ser escolhido arbitrariamente. O comportamento deve ser factível e uma política de ação deve ser executada no ambiente para que um comportamento seja demonstrado. Enquanto o primeiro problema influencia a inferência de como o usuário avalia comportamentos, o segundo problema influencia quão rápido e acurado o processo de aprendizado pode ser feito. Esta tese propõe o problema de Extração de Preferências com base em Comportamentos Observados utilizando o arcabouço de Processos Markovianos de Decisão, desenvolvendo propriedades teóricas em tal arcabouço que viabilizam computacionalmente tal problema. O problema de diferentes percepções é analisado e soluções restritas são desenvolvidas. O problema de demonstração de comportamentos é analisado utilizando formulação de questões com base em políticas estacionárias e replanejamento de políticas, sendo implementados algoritmos com ambas soluções para resolver a extração de preferências em um cenário sob condições restritas.

ABSTRACT

Recently, computer systems have been delegated to accomplish a variety of tasks, when the computer system can be more reliable or when the task is not suitable or not recommended for a human being. The use of preference elicitation in computational systems helps to improve such delegation, enabling lay people to program easily a computer system with their own preference. The preference of a person is elicited through his answers to specific questions, that the computer system formulates by itself. The person acts as an user of the computer system, whereas the computer system can be seen as an agent that acts in place of the person. The structure and context of the questions have been pointed as sources of variance regarding the user's answers, and such variance can jeopardize the feasibility of preference elicitation. An attempt to avoid such variance is asking an user to choose between two behaviours that were observed by himself. Evaluating relatively observed behaviours turn questions more transparent and simpler for the user, decreasing the variance effect, but it might not be easier interpreting such evaluations. If divergences between agent's and user's perceptions occur, the agent may not be able to learn the user's preference. Evaluations are generated regarding user's perception, but all an agent can do is to relate such evaluation to his own perception. Another issue is that questions, which are exposed to the user through behaviours, are now constrained by the environment dynamics and a behaviour cannot be chosen arbitrarily, but the behaviour must be feasible and a policy must be executed in order to achieve a behaviour. Whereas the first issue influences the inference regarding user's evaluation, the second problem influences how fast and accurate the learning process can be made. This thesis proposes the problem of Preference Elicitation under Evaluations over Observed Behaviours using the Markov Decision Process framework and theoretic properties in such framework are developed in order to turn such problem computationally feasible. The problem o different perceptions is analysed and constraint solutions are developed. The problem of demonstrating a behaviour is considered under the formulation of question based on stationary policies and non-stationary policies. Both type of questions was implemented and tested to solve the preference elicitation in a scenario with constraint conditions.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

c.p.1 com probabilidade 1

EP Extração de Preferências

EPCO Extração de Preferências com base em Comportamentos Observados

PMD Processo Markoviano de Decisão

TUE Teoria da Utilidade Esperada

LISTA DE SÍMBOLOS

$\sim$ - indiferente a

$\succ$ - melhor que

$\succeq$ - melhor ou indiferente a

$\langle \mathbf{x}, \mathbf{y} \rangle$ - produto escalar entre os vetores $\mathbf{x}$ e $\mathbf{y}$

$\mathcal{A}$ - um conjunto de ações disponíveis

a - uma ação

a_t - a ação executada pelo agente no tempo t

$\mathrm{co}(X)$ - o envoltório convexo do conjunto X

$\mathcal{D}$ - um conjunto de decisões

d, d', d'' - uma decisão

d^* - a decisão ótima

d^*_u - a decisão ótima segundo a avaliação u

$E^{\pi}_{\boldsymbol{\mu}^{\mathsf{T}}\boldsymbol{\mu}}(s,t)$ - a esperança $\mathrm{E}[\boldsymbol{\mu}(s,t,\pi)^{\mathsf{T}}\boldsymbol{\mu}(s,t,\pi)]$

$\mathrm{E}_{\{x\sim\Pr(x)\}}[f(x)]$ - a esperança da função $f(x)$ quando a variável x é submetida à distribuição de probabilidades $\Pr(x)$

$\mathrm{Info}(q)$ - a informação esperada trazida na resposta a questão q

$\mathcal{L}$ - um conjunto de loterias

$\mathcal{M}$ - o conjunto de todos vetores de atributos esperados que podem ser gerados por políticas estocásticas

$\widehat{\mathcal{M}}$ - uma aproximação para o conjunto $\mathcal{M}$

$\widehat{\mathcal{M}}^{\epsilon}$ - uma aproximação com erro ϵ para o conjunto $\mathcal{M}$

$M(\boldsymbol{\mu})$ - o mapeamento do comportamento real ψ mais provável dada a observação do vetor de atributos $\boldsymbol{\mu}$

$N(\hat{x}, \sigma)$ - a distribuição normal com valor médio $\hat{x}$ e desvio padrão σ

o_t - a observação do agente no tempo t

$\widehat{P}^{r,\pi_q^1,\pi_q^2,u}$ - uma estimativa para a probabilidade $\Pr(r|\pi_q^1,\pi_q^2,u)$.

P_0 - a distribuição de probabilidades para o estado inicial s_0

$p_{\min}^{\{\Psi|\Pi\}}$ - a menor porcentagem de ocorrência para um comportamento arbitrário em Ψ e uma política arbitrária em Π

$\mathcal{Q}$ - um conjunto de questões

q - uma questão

q^* - a questão ótima

$\mathcal{R}_q$ - o conjunto de respostas possíveis para a questão q

r - uma resposta

$\mathrm{Regret}(d,u)$ - o arrependimento de tomar a decisão d frente a função utilidade u

$\mathrm{Replaneja}(t,\Gamma)$ - função que indica se deve haver replanejamento quanto o agente se encontra no tempo t do período Γ

$S_{\mathrm{Us}}^{\boldsymbol{\mu}_{\mathrm{Ag}},\pi,\epsilon}$ - uma hiperesfera que contém, com probabilidade maior que ϵ, os possíveis vetores de atributos observados pelo usuário, quando o agente observa $\boldsymbol{\mu}_{\mathrm{Ag}}$ e a política π é executada

$\mathcal{S}$ - o conjunto de estados do ambiente

s - um estado do ambiente

s_t - o estado do ambiente no tempo t

$T(s'|s,a)$ - a probabilidade de transição do estado s para o estado s' dado que a ação a foi executada

$T_{s,s'}^{\pi,t}$ - a probabilidade de transição do estado s para o estado s' no tempo t, quando a política π é executada

t, t', t'' - indicação de tempo

t^{Γ} - o tempo no qual o período Γ se encontra

t^{G} - o tempo global para o processo de EPCO

$\mathcal{U}$ - um conjunto de funções utilidades candidatas

u_{ψ}, $u(\psi)$ - a utilidade da conseqüência ψ

$u(\mu)$ - a utilidade do valor de atributo escalar μ

$u(\boldsymbol{\mu})$ - a utilidade do vetor de atributos $\boldsymbol{\mu}$

$u_{\mathrm{Ag}}(\boldsymbol{\mu}_{\mathrm{Ag}})$ - a utilidade atribuída pelo agente ao vetor de atributos observado pelo agente

$u_{\mathrm{Us}}(\boldsymbol{\mu}_{\mathrm{Us}})$ - a utilidade atribuída pelo usuário ao vetor de atributos observado pelo usuário

$u_i(\mu_i)$ - a utilidade atribuída ao i-ésimo atributo

V^q - o valor da questão q

V^d - o valor da decisão d

V^π - o valor da política π

V_u^d - o valor da decisão d avaliado pela função utilidade u

V_u^π - o valor da política π avaliada pela função utilidade u

$V'(\tilde{x})$, $V(\tilde{x})$ - valor da loteria $\tilde{x}$

$V(\tilde{x}, \tilde{y})$ - valor que compara as loterias $\tilde{x}$ e $\tilde{y}$

$V^*(s,t)$ - recompensa acumulada esperada a partir de s no tempo t até o final de um período seguindo uma política ótima

$V'(\tilde{x})$, $V(\tilde{x})$ - o valor da loteria $\tilde{x}$

$V(\tilde{x}, \tilde{y})$ - o valor que compara as loterias $\tilde{x}$ e $\tilde{y}$

$v(\psi)$ - o valor da conseqüência ψ

$\mathcal{W}$ - um conjunto de vetores recompensas candidatos

$w(s,a)$ - uma função recompensa

$\mathbf{w}$ - um vetor de pesos ou vetor recompensa

$\mathbf{w}_i^{\max}$ - o vetor recompensa que considera apenas o i-ésimo atributo como relevante e de forma positiva

$\mathbf{w}_i^{\min}$ - o vetor recompensa que considera apenas o i-ésimo atributo como relevante e de forma negativa

w_i - o peso ou recompensa atribuído ao i-ésimo atributo

$\tilde{x}$, $\tilde{x}'$, $\tilde{x}''$ - uma loteria

$\mathbf{x}$ - um estado estendido do processo de EPCO

$\mathbf{x}_t$ - o estado estendido do processo de EPCO no tempo t

α_i - a probabilidade associada a uma conseqüência ou comportamento dentro de

uma loteria

$\alpha_{\tilde{x}_i}$ - a probabilidade associada a uma conseqüência ou comportamento dentro da loteria $\tilde{x}$

Γ_i - um período de demonstração de comportamento

$\boldsymbol{\mu}(\psi)$ - o vetor de atributos que representa a conseqüência (comportamento) ψ

$\boldsymbol{\mu}_{\mathrm{Ag}}(\psi)$ - o vetor de atributos observados pelo agente quando o comportamento ψ é demonstrado

$\boldsymbol{\mu}_{\mathrm{Us}}(\psi)$ - o vetor de atributos observados pelo usuário quando o comportamento ψ é demonstrado

$\bar{\boldsymbol{\mu}}^{\pi}$ - o vetor de atributos esperados para a política π

$\bar{\boldsymbol{\mu}}^{\pi}(t)$ - o vetor de atributos esperados para a política π até o tempo t

$\bar{\boldsymbol{\mu}}^{\pi}(s,t)$ - o vetor de atributos esperados para a política π a partir de s no tempo t até o final do período

$\boldsymbol{\mu}(s,t,\pi)$ - o vetor de atributos acumulados observado a partir de s no tempo t até o final de um período ao executar a política π

$\boldsymbol{\mu}^{*}$ - o melhor vetor de atributos

$\boldsymbol{\mu}^{0}$ - o pior vetor de atributos

$\bar{\boldsymbol{\mu}}_i^{\max}$ - o vetor de atributos esperados para uma política que maximiza a observação do i-ésimo atributo

$\bar{\boldsymbol{\mu}}_i^{\min}$ - o vetor de atributos esperados para uma política que minimiza a observação do i-ésimo atributo

$\mu_i(\psi)$ - a medida do i-ésimo atributo que representa a conseqüência ou o comportamento ψ

μ, μ' - um valor de atributo escalar que representa uma conseqüência ou comportamento

$\mu(\psi)$ - o valor de atributo escalar que representa a conseqüência ou o comportamento ψ

μ_i^{*} - a melhor medida para o atributo i

μ_i^{0} - a pior medida para o atributo i

$\bar{\mu}_{\tilde{x}}$ - o valor de atributo escalar esperado da loteria $\tilde{x}$

$\hat{\mu}_{\tilde{x}}$ - o valor de atributo escalar assegurado da loteria $\tilde{x}$

$\boldsymbol{\phi}(s, a)$ - o vetor de atributos observados no estado s com a execução da ação a

$\boldsymbol{\phi}_{\mathrm{Ag}}(s, a)$ - o vetor de atributos observados pelo agente no estado s com a execução da ação a

$\boldsymbol{\phi}_{\mathrm{Us}}(s)$ - o vetor de atributos observados pelo usuário no estado s com a execução da ação a

$\phi_i(s, a)$ - a medida do i-ésimo atributo observado no estado s com a execução da ação a

Π - um conjunto de políticas

Π_{Det} - o conjunto de políticas deterministas

Π_{Est} - o conjunto de políticas estocásticas

Π_{NonDom} - o conjunto de políticas não dominadas

π - uma política de ação

π^* - a política de ação ótima

$\pi^*_{\mathbf{w}}$ - a política de ação ótima para o vetor recompensa $\mathbf{w}$

π^1_q, π^2_q - políticas de ação a serem executadas que definem a questão q

$\pi(s)$ - a ação a ser executada no estado s

$\pi(s, t)$ - a ação a ser executada no estado s e tempo t

Ψ - um conjunto de conseqüências ou comportamentos

ψ, ψ', ψ'', ψ''', ψ_i - uma conseqüência ou comportamento

ψ^* - a melhor conseqüência ou comportamento possível

ψ_0 - a pior conseqüência ou comportamento possível

ψ_{Ag} - um comportamento observado pelo agente

ψ_{Us} - um comportamento observado pelo usuário

$\psi(d)$ - o comportamento resultante da decisão d

Σ^π - a matriz de covariância do vetor de atributos observados ao executar a política π

$\Sigma^\pi(s, t)$ - a matriz de covariância do vetor de atributos observados acumulados desde o estado s e tempo t ao executar a política π

$\Sigma_{\boldsymbol{\phi}(s,a)}$ - a matriz de covariância da observação $\boldsymbol{\phi}(s, a)$ de ocorrência de atributos

σ - a variância de uma variável aleatória

$\sigma_{x,y}$ - a covariância entre as variáveis x e y

Ξ, Ξ' - um conjunto de atributos

Ξ_{Ag} - o conjunto de atributos considerados pelo agente

Ξ_{Us} - o conjunto de atributos considerados pelo usuário

SUMÁRIO

1 INTRODUÇÃO

Delegação de atribuições e responsabilidade, inteligência e interface orientada a humanos são apresentadas por Wooldridge (2002) como algumas das tendências pelas quais a área de computação é marcada. A delegação implica que o controle de tarefas é dado para sistemas computacionais e isso pode ocorrer por diversos motivos: repetição, o que causaria fadiga e desinteresse em pessoas; perigo, o que colocaria a vida e saúde de pessoas em risco; habilidade, quando pessoas não conseguem executar uma tarefa seja mental ou fisicamente; entre outros. Inteligência e delegação juntas proporcionam a um sistema computacional a capacidade de decidir e agir de forma efetiva no lugar de uma pessoa (SISKOS; SPYRIDAKOS, 1999; ALOYSIUS et al., 2006). Nesse caso, a pessoa responsável, e contemplada pelos efeitos de tais decisões, pode ser considerada um usuário e suas preferências devem ser configuradas no sistema computacional. O uso de interface orientada a humanos tem permitido que um número cada vez maior de usuários possa configurar e programar eficientemente um sistema computacional, mesmo que tais usuários não tenham conhecimentos específicos sobre sistemas computacionais.

A disseminação de sistemas computacionais em suas mais diversas formas trouxe a possibilidade dos mesmos interagirem não somente com pessoas, mas também com outros sistemas, sejam eles computacionais ou não, fazendo com que tais sistemas se percebam inseridos em um ambiente. Nesta tese, sistemas computacionais são vistos como agentes artificiais, ou simplesmente, agentes, que percebem e agem neste ambiente (RUSSELL; NORVIG, 2004).

O objeto desta tese são agentes que facilitam a delegação do controle sobre uma tarefa de interesse do usuário ao próprio agente (DENNIS et al., 2005; VIAPPIANI; FALTINGS; PU, 2006; PU; FALTINGS; TORRENS, 2003). Esses agentes estabelecem uma interação direta com o usuário, eliminando a necessidade de um programador que possua conhecimentos tanto do agente como da tarefa que se deseja controlar e que traduza os objetivos do usuário para uma linguagem do agente. A eliminação do programador é possível, primeiro, ao prover uma arquitetura que permita mapear os objetivos do usuário na programação do agente. Depende dessa arquitetura o

nível de abstração com que o usuário pode programar um agente, dando a este maior ou menor autonomia nas decisões para satisfazer os objetivos daquele (RUSSELL; NORVIG, 2004). Segundo, ao prover um método para facilitar a programação do agente pelo usuário, mesmo que este seja leigo na arquitetura daquele. Com origem na Economia (NEUMANN; MORGENSTERN, 1947), a Extração de Preferências (EP) é um método para descrever formalmente os objetivos do usuário, de modo a permitir a interpretação desses objetivos pelo agente e fazer com que ele possa satisfazê-los. Nesta tese, considerando a literatura de EP tradicional, isto é, de origem na Economia (KEENEY; RAIFFA, 1976; PLOTT, 1996; BRAGA; STARMER, 2005), e a sua aplicação na automatização da delegação de tarefas (BRAZIUNAS, 2006; BOUTILIER et al., 2005; CHAJEWSKA; KOLLER; PARR, 2000), questiona-se a validade da informação obtida pela interação entre agente e usuário ao delegar tarefas a agentes e estabelece-se que é necessário um método de interação entre usuários e agente que torne mais fidedigna a informação obtida junto ao usuário. Isso é necessário pois, quanto maior a confiança no processo de obter informações, melhor será o desempenho do agente para satisfazer o usuário. Tal fidedignidade é alcançada com a proposição de uma interação mais natural entre agente e usuário, aqui denominada por Extração de Preferências com base em Comportamentos Observados (EPCO).

1.1 Arquitetura do Agente

Diversas formas de representação têm sido utilizadas para programar no agente as preferências do usuário, delegando diferentes graus de autonomia ao mesmo. Considere que se deseje construir um agente *Piloto Automático* para guiar o carro de um usuário diariamente de sua casa ao trabalho. Como as preferências entre usuários podem variar, é importante que o agente atenda às expectativas do usuário que ele está conduzindo. Um agente pode ser programado através de uma política, que relaciona cada possível situação a uma ação apropriada. No caso do agente *Piloto Automático*, deve-se definir para cada cruzamento (situação) qual decisão tomar (seguir em frente, à direita, à esquerda). Um agente programado de tal forma não teria nenhuma autonomia para tomar decisões, mas apenas para executá-las, pois o agente é programado com "como satisfazer" um usuário. Neste trabalho é explorada uma segunda forma de programação, que é a de programar o agente com "o que é satisfazer" o usuário e permitir ao agente planejar como satisfazê-lo. Deve-se descrever de forma objetiva como o agente deve avaliar cada uma das possíveis ações a tomar, especificando ao agente restrições que devem ser respeitadas para satisfazer as preferências do usuário, deixando o agente livre para definir e escolher ações que satisfaçam tais restrições.

Atributos ajudam a quantificar o quanto um dado objetivo foi atingido. No exemplo do agente *Piloto Automático*, uma trajetória, ou, genericamente, um comportamento, pode ser avaliada sob a ótica de vários atributos: tempo da trajetória, tamanho da trajetória, pedágio pago ao realizar a trajetória, combustível consumido ao realizar a trajetória, qualidade das vias percorridas na trajetória, etc. Se todos esses atributos são importantes, caracteriza-se um problema de decisão com base em vários atributos, que em geral não podem ser otimizados separadamente, já que a melhoria em um pode significar piora em outro. Portanto, é necessário definir qual o compromisso existente entre cada um desses atributos. Por exemplo, o usuário considera positiva ou negativa a opção de gasto de 20 minutos a mais no tempo da trajetória frente à economia de 1 litro de gasolina?

Além do compromisso entre os atributos, outro aspecto importante é o compromisso com o risco presente quando o ambiente não é determinista. A cada dia o trânsito nas vias de uma cidade é diferente, por exemplo, se ocorre um acidente, o tráfego pode ficar muito mais intenso. Eventos como esse são imprevisíveis e tudo o que se pode fazer *a priori* é considerar probabilidades de ocorrência dos possíveis eventos e as probabilidades de ocorrência dos comportamentos resultantes. Então, deve-se definir como o agente considerará opções entre trajetórias com pouca variação – por exemplo, com tempo da trajetória fixo de 60 minutos, ou seja, com probabilidade 1 (c.p.1) – e trajetórias mais arriscadas – por exemplo, com tempo da trajetória de 40 minutos em dias normais, mas, mediante acidentes, a trajetória pode durar 2 horas.

Funções utilidade são modelos para especificar as preferências de um usuário que contemplam o compromisso entre os atributos e o compromisso com o risco. A Teoria da Utilidade Esperada (TUE) baseia-se no uso de valores escalares para definir a decisão ótima do agente e esses valores representam a utilidade de cada possível comportamento. A decisão ótima é aquela que maximiza a utilidade esperada em relação às probabilidades de ocorrência de cada comportamento (KEENEY; RAIFFA, 1976). Para que funções utilidade possam ser fidedignas ao descrever as preferências de um usuário, este deve apresentar alguns preceitos de racionalidade, que são mapeados em axiomas na TUE. A TUE vem sendo utilizada como modelo para especificar objetivos de usuários, quando os mesmos são delegados a agentes (CHEN; PU, 2004; DENNIS; HEALEY, 2003). Embora seja controverso que a TUE seja capaz de especificar de forma completa as preferências de um usuário, a validade de suas propriedades normativas de racionalidade não são negadas (BRAGA; STARMER, 2005; NEUMANN; MORGENSTERN, 1947).

1.2 Extração de Preferências

O problema de EP de um usuário tem sido estudado por décadas (KEENEY; RAIFFA, 1976; CHANKONG; HAIMES, 1983; BRAGA; STARMER, 2005), onde entrevistas com os usuários são utilizadas para obter informações sobre suas preferências. As questões formuladas nas entrevistas podem envolver diretamente a estrutura da função utilidade, questionando a utilidade de comportamentos arbitrários, atributos relevantes, ou ainda o compromisso entre os diversos atributos envolvidos. No entanto, essas questões exigem que o usuário tenha um conhecimento abstrato sobre as suas preferências, conhecendo, inclusive, a representação de preferências envolvida.

Para ser acessível e confiável a uma maior quantidade de usuários, pode-se utilizar questões cujas respostas gerem o mesmo tipo de informação com relação às preferências dos usuários, mas que exijam dos usuários menos deliberação e menos conhecimento sobre suas próprias preferências. Utilizam-se então questões baseadas em comportamentos hipotéticos, valendo-se de representações de preferências mais comuns a seres humanos e supostamente mais fáceis de serem respondidas do que questões diretas sobre a estrutura de utilidades de comportamentos. Essas questões se baseiam em duas ou mais opções, entre as quais o usuário deve eleger a melhor e, com esse conhecimento, pode-se estabelecer relações de preferências entre as opções envolvidas. A cada resposta dada às questões, restrições podem ser impostas a funções utilidade candidatas. As questões são escolhidas de modo a reduzir a quantidade de funções utilidade candidatas e permitir a tomada de decisão ótima no lugar do usuário.

O uso de comportamentos hipotéticos no lugar de comportamentos reais permite reduzir o problema de confiabilidade na resposta do ser humano, por exemplo, comparando comportamentos que variam apenas um único atributo por vez (KEENEY; RAIFFA, 1976), além de possibilitar a formulação de questões que otimizem a informação disponibilizada em cada interação com o usuário. No entanto, ao efetuar questões e obter respostas hipotéticas de um usuário, considera-se que tanto o usuário como quem está efetuando a questão possuem a mesma interpretação das questões. No caso em que comportamentos se baseiam em atributos, o usuário deve poder analisar um comportamento apenas pelos atributos apresentados, exigindo que estes atributos sejam completos e comuns para o usuário e para quem efetua a questão. Além disso, o usuário é obrigado a trabalhar com quantificações dos comportamentos, que não necessariamente são naturais a ele. Além do problema na descrição do comportamento, ainda existe o problema na descrição das questões, pois o usuário deve ter uma interpretação precisa do que está sendo questionado. Quando se automatiza a

EP, as questões são formuladas e efetuadas por um agente, exigindo uma linguagem comum entre agente e usuário.

1.3 Validade do Princípio da Preferência Revelada

A EP baseia-se no princípio da preferência revelada, que diz que as escolhas realizadas pelos seres humanos refletem as suas preferências (GRUNE, 2004). No entanto, por vezes não se consegue obter consistência entre as preferências reveladas por seres humanos e a TUE, implicando que ou o princípio da preferência revelada não é válido, ou a TUE não é válida (STARMER, 2000). Isso fez com que, desde o seu aparecimento, a TUE fosse criticada como teoria descritiva, isto é, como teoria para descrever como seres humanos tomam decisões, e novas teorias, melhores do ponto de vista descritivo, tenham sido desenvolvidas (STARMER, 2000). Algumas dessas teorias ainda consideram uma função utilidade, que serve ao propósito de ordenar cada um dos possíveis comportamentos, mas ao avaliar decisões, outras formas além da utilidade esperada são consideradas, por exemplo, ao tratar de forma não linear as probabilidades envolvidas. Em outras teorias, essa mudança é ainda mais radical, onde até mesmo a comparação entre comportamentos não pode ser realizada por uma função utilidade, mas são orientadas ao contexto dos comportamentos envolvidos.

A inconsistência da TUE como teoria descritiva aparece de forma sistemática principalmente em experimentos envolvendo opções com probabilidades. Alguns interpretam essa inconsistência como uma inadequação da TUE para representar as preferências de um usuário, pois consideram a decisão do usuário como correta e representante de suas preferências (STARMER, 2000). Uma outra corrente interpreta essa inconsistência como sendo um erro nas decisões do usuário, que podem não interpretar corretamente as opções disponíveis, ou ainda, utilizando uma interpretação mais comportamental, que os usuários se baseiam em estratégias de decisões que não representam uma decisão normativa, mas que equacionam os conhecimentos do usuário e um compromisso entre o que será alcançado com a decisão e o esforço envolvido para realizar tal decisão.

Plott (1996) argumenta que o usuário possui preferências consistentes com a TUE e que tais preferências se evidenciariam caso o usuário fosse submetido a condições adequadas de questionamento. Segundo Plott, as condições ideais devem envolver: transparência e simplicidade, isto é, o usuário deve perceber sem erro de interpretação o que lhe é questionado e responder a tal questão com pouco esforço cognitivo; incentivo, de modo que o usuário sofra as conseqüências de suas decisões para que lhe obrigue a despender um esforço cognitivo adequado; e oportunidades de aprendizado,

para que o usuário perceba os efeitos de suas decisões, tanto em si mesmo como no resultado delas no ambiente, e aprenda com tais observações.

Nem sempre é possível ter conhecimento completo sobre o sistema de avaliação utilizado pelo usuário, nem mesmo o usuário sempre tem acesso explícito ao seu próprio sistema de avaliação, sendo que muitas vezes conhecimentos tácitos são utilizados na avaliação de opções disponíveis para tomada de decisão. O uso de questões hipotéticas na EP coloca o usuário em uma situação não natural, impossibilitando que esse conhecimento tácito venha à tona e propiciando os erros descritos por Plott (1996). Avaliar comportamentos reais, que são interpretados diretamente pelo próprio usuário, como é feito normalmente no dia-a-dia, ao invés de avaliar uma representação abstrata e com um viés lingüístico, pode tornar mais fidedigna a avaliação pelo usuário de uma determinada opção, pois apresenta uma maior simplicidade para o usuário ao não envolver modelos abstratos de suas preferências.

1.4 Extração de Preferências com base em Comportamentos Observados

Neste trabalho, introduz-se o problema de EPCO que impõe a restrição de que as informações sejam obtidas apenas com base em comportamentos reais, comportamento estes que devem ser exibidos pelo agente no ambiente e observados pelo usuário. Se forem consideradas apenas questões comparando comportamentos, o usuário não precisa de nenhuma forma de abstração sobre a TUE, diminuindo a possibilidade de equívocos na interpretação, sendo assim um tipo de questão totalmente transparente. Tudo que é requisitado ao usuário é que observe dois comportamentos executados pelo agente e que escolha aquele que considere melhor.

Ao avaliar comportamentos reais, o usuário não avalia apenas as decisões do agente, mas também o retorno emotivo que os comportamentos demonstrados podem vir a causar no usuário ao percebê-los no ambiente. Isso permite que o usuário, após ter percebido e sentido dois comportamentos, possa escolher o que mais lhe satisfaz, independentemente das ações executadas pelo agente, de uma linguagem comum entre agente e usuário e do não determinismo envolvido no ambiente. A escolha depende somente de como o usuário reage a tais comportamentos, tornando a escolha mais simples. Ainda, sistemas computacionais que já estejam em funcionamento, ou seja, controlando uma tarefa delegada pelo usuário, podem incitar mais facilmente o usuário a emitir críticas quando não contente com a atuação, o que não pode ocorrer quando se utiliza questões hipotéticas.

Apesar do uso de comportamentos observados apresentar uma maior naturalidade na forma como as questões são colocadas ao usuário, não exigindo uma linguagem comum de descrição do comportamento e nenhum conhecimento explícito do usuário sobre suas próprias preferências, o uso de comportamentos observados também apresenta algumas limitações:

1. a EPCO só pode ser aplicada a tarefas que podem ser repetidas e experimentadas, sendo necessária a avaliação do compromisso entre o custo de experimentação de um comportamento (duração de um comportamento, efeitos no ambiente, efeitos afetivos no usuário) e o retorno que uma delegação mais fidedigna pode trazer ao usuário;

2. as questões que podem ser formuladas, ou seja, os comportamentos que podem ser demonstrados, dependem da dinâmica do ambiente, implicando que se deve primeiro conhecer e analisar o ambiente de modo a permitir que o processo de EP possa ser implementado; isso consiste não só em saber quais comportamentos são factíveis, mas também saber quais as chances de obtê-los e qual política executar para obtê-los;

3. a exibição de um comportamento não pode ser obtida c.p.1 em um ambiente não determinista, implicando que os algoritmos de EP não devem apenas escolher a questão mais informativa, isto é, questões que demonstram comportamentos que permitam melhor discriminar entre vários modelos de preferências do usuário, mas também uma que tenha uma boa chance de ocorrência, devendo-se considerar um compromisso entre esses dois fatores; e

4. as observações do comportamento, feitas tanto pelo agente como pelo usuário, podem ser distintas; no problema de EPCO as avaliações emitidas pelo usuário são feitas a partir de sua própria observação; no entanto essas avaliações são transmitidas ao agente e associadas às observações deste. Será que, mesmo quando as observações envolvidas são diferentes, ainda faz sentido obter uma decisão adequada em favor do usuário? Quais tipos de inconsistência podem resultar dessa condição e como evitá-las?

A construção de um agente que trabalhe com qualquer função utilidade, tanto na tomada de decisão, como na EP, pode ser muito custosa computacionalmente. Ao limitar os tipos das funções utilidade consideradas como candidatas pelo agente, facilita-se a tomada de decisão quando já definida uma instância dentre essas funções utilidade, uma vez que se pode escolher um agente especializado neste tipo de

função utilidade. Já para o processo de EP, o uso de funções utilidade com estruturas permite que informações sobre poucos comportamentos possam ser estendidas a outros comportamentos, possibilitando a obtenção de uma decisão ótima com a formulação de poucas questões. No entanto, ao se limitar a alguns tipos de funções utilidade, pode acontecer que a mesma não se adeque às preferências do usuário, tornando importante o conhecimento das características que tais preferências devem apresentar.

Neste trabalho será considerado um agente que representa as preferências de um usuário em uma função utilidade que, considerando o conceito de atributos, apresenta a propriedade de independência aditiva. Esta propriedade implica que a contribuição de cada atributo para a utilidade esperada de uma decisão depende apenas das distribuições de probabilidades marginais dos atributos referentes aos comportamentos envolvidos. Nesse caso, os atributos não se complementam, podendo a contribuição de cada um ser calculada separadamente, e a função utilidade conjunta nada mais é do que a soma dessas contribuições. Além disso, será considerado que o usuário é neutro a risco, o que, junto com a propriedade de independência aditiva, implica que uma decisão com risco possa ser analisada apenas pela esperança marginal de seus atributos, apresentando linearidade na avaliação dos atributos. Além disso, serão considerados ambientes discretos e não deterministas.

1.5 Objetivos

O objetivo desta tese é propor um método que permita uma nova forma de interação entre agentes e seres humanos, tornando mais natural a avaliação das questões por estes e possibilitando a aqueles uma EP mais fidedigna. O método proposto é a EPCO.

Além disso, visa-se analisar a implicação da EPCO nos algoritmos de EP e propor soluções em tais casos. Essa análise é feita sob três aspectos: 1) interação entre agente e ambiente para determinar características das questões que podem ser formuladas, 2) diferentes observações do agente e do usuário, e 3) formulação de questões baseada em comportamentos demonstrados. Esta análise consiste em interpretar essas limitações frente a alguns algoritmos tradicionais de EP, e propor técnicas para lidar com estes problemas quando as preferências do usuário apresentam propriedades de independência aditiva e neutralidade ao risco, culminando na proposta de novos algoritmos para a EPCO.

O objetivo é alcançado através da:

1. definição de um arcabouço formal para o problema de EPCO;
2. análise das relações entre políticas executadas pelo agente e comportamentos demonstrados no ambiente;
3. análise das relações entre agente e usuário sob o aspecto das observações de ambos;
4. análise das relações entre agente e ambiente sob o aspecto de formulação de questões; e
5. proposta, definição e implementação de um algoritmo para a EPCO capaz de equacionar de forma eficaz todas estas relações.

1.6 Contribuições

A principal contribuição desta tese consiste na apresentação, formalização e levantamento de problemas a respeito de um novo cenário para realizar a EP, isto é, a EPCO. Na EPCO acredita-se que avaliações emitidas por um usuário sejam mais fidedignas às suas preferências.

Uma segunda contribuição é o projeto de um algoritmo que define todos as políticas factíveis não dominadas para um ambiente. Esse algoritmo permite restringir o espaço de tomadas de decisões, mesmo antes que o agente obtenha informações sobre as preferências do usuário, auxiliando na redução do custo computacional ao formular questões.

Uma terceira contribuição é a análise feita a respeito do problema de diferentes observações do agente e do usuário. Essa análise apresenta as condições que tais observações devem respeitar para que a EP seja factível sob diferentes observações.

Uma quarta contribuição são duas possíveis soluções para formular questões na EPCO. A primeira solução baseia-se em políticas estacionárias e a formulação de uma questão consiste em determinar duas políticas estacionárias a serem executadas pelo agente para demonstrar os dois comportamentos a serem comparados pelo usuário. A segunda solução considera o replanejamento de políticas estacionárias após observações parciais do comportamento está sendo demonstrado, permitindo amenizar o problema de não determinismo no ambiente para demonstrar comportamentos mais informativos.

Finalmente, duas implementações para o problema de EPCO, utilizando políticas estacionárias e replanejamento, são consideradas em um ambiente que simula um

robô real.

1.7 Organização do Texto

No capítulo 2 é apresentado o problema de delegação sob o ponto de vista das preferências do usuário e das decisões racionais que devem ser tomadas pelo agente, assim como a EP como método para auxiliar na delegação. No capítulo 3 são apresentadas críticas à forma como a EP é feita atualmente, apresentando em seguida a proposta desta tese.

No capítulo 4 são apresentados: o cenário proposto na tese para minimizar o erro emitido por avaliações do usuário, ou seja, a EPCO, o arcabouço da relação entre agente e ambiente que é utilizado nesta tese e os problemas inerentes ao cenário de EPCO neste arcabouço.

No capítulo 5 é apresentada uma análise do arcabouço utilizado, relacionando políticas do agente e comportamentos demonstrados. Experimentos são realizados para demonstrar que um pré-processamento pode ser realizado para reduzir o custo computacional envolvido na EPCO, e que essa redução produz pouca perda na qualidade das decisões ótimas encontradas.

No capítulo 6 é apresentada uma análise para o problema de observação. Experimentos são realizados comparando a validade da EPCO em usuários simulados com funções utilidade lineares e em usuários simulados com funções sem estruturas, demonstrando que, em alguns casos, o primeiro tipo de função pode garantir decisões ótimas mesmo sob diferentes observações do agente e do usuário.

No capítulo 7 são apresentadas duas soluções para o problema de formulação de questões na EPCO: uma baseada em políticas estacionárias e uma baseada em replanejamento de políticas estacionárias. Experimentos são realizados em um ambiente que simula um robô real comparando as duas soluções.

Finalmente, no capítulo 8 são apresentadas as conclusões e sugestões para continuação do trabalho.

2 DELEGAÇÃO DE TAREFAS: PREFERÊNCIAS, DECISÃO RACIONAL E EXTRAÇÃO DE PREFERÊNCIAS

A delegação ocorre quando um indivíduo, o usuário, atribui autoridade a outro indivíduo, o executor, para executar uma determinada tarefa de responsabilidade do usuário. Embora o executor execute tal tarefa, o usuário continua sendo o beneficiário de seu desfecho. Dessa forma, a autoridade para tomar uma decisão é passada do usuário ao executor, mas as responsabilidades, tanto positivas como negativas, continuam sendo do usuário.

Dois aspectos que influenciam diretamente o desfecho da tarefa a ser realizada é a descrição da tarefa e a capacidade do executor para executar a tarefa. Um caso especial é quando o executor é um agente artificial. Enquanto delegar uma tarefa ao agente artificial apresenta várias vantagens, algumas desvantagens podem ser cruciais. Um agente artificial pode ser criado com especialização em completar uma determinada tarefa, muitas vezes atuando melhor que seres humanos, pois, além de ser especializado, em geral, pode-se programar arbitrariamente aspirações próprias, diferentes da tarefa programada, e temores instintivos, que restrinja as ações do agente, mas de forma que garanta a compatibilidade com a tarefa desejada. Por outro lado, compreender a descrição de uma tarefa feita por um ser humano ainda é uma capacidade não dominada de forma satisfatória por agentes artificiais.

O agente artificial – genericamente agente – só pode atender às expectativas do usuário quando possui conhecimento de tais expectativas. Essas expectativas vêm à tona parcialmente ou de forma completa por intermédio de um diálogo entre usuário e agente, diálogo este que depende de um corpo conceitual comum entre os dois. Este diálogo pode ser preponderante em uma dentre duas situações. Uma situação é quando o usuário já possui sintetizada a tarefa que deseja delegar, ele precisa apenas expressa-lá de forma completa e transparente utilizando o corpo conceitual comum, ao agente cabe apenas interpretar tal tarefa, sintetizando-a de uma forma que o ajude a completá-la. Por outro lado, quando o usuário não possui sintetizada a tarefa que deseja delegar e só pode expressá-la de forma vaga, então cabe ao agente,

nesta situação, questionar o usuário até trazer à tona a tarefa em questão de modo transparente e completo. Essa completude depende, é claro, de uma perspectiva *a priori* que o agente possua dessa tarefa, permitindo uma abstração da tarefa; caso contrário pode-se tornar inviável descrevê-la, pois deve-se tratar de forma extensiva todas as perspectivas possíveis.

A segunda situação discutida é estudada na literatura da área de Economia sob o nome de EP (KEENEY; RAIFFA, 1976): extração no sentido de trazer à tona, e preferências no sentido de representação dos desejos do usuário. Mas que relação tais preferências possuem com as expectativas sobre as decisões que o agente deve tomar? As tomadas de decisões e preferências normalmente são relacionadas sob o conceito de racionalidade, no sentido de agir de forma correta. Neste capítulo serão discutidos a relação entre preferências e decisões, os axiomas de racionalidade que podem mapear essa relação e a implicação desses axiomas para a EP.

2.1 Racionalidade e o Problema de Decisão

Embora o conceito de racionalidade seja de uso corrente, onde pessoas normalmente classificam atos ou indivíduos como racionais ou irracionais, definir formalmente racionalidade não é trivial. O modelo aristotélico de racionalidade (ZILHãO, 2001) coloca que a decisão de um agente é racional quando: 1) o agente A tem um desejo F; e 2) o agente A tem uma crença C cujo conteúdo é o de que tomar a decisão d é o melhor que ele tem a fazer para obter F; então, o agente A faz d.

Nesse modelo, Aristóteles confronta a razão com a paixão, pois se pela razão um indivíduo considera mais adequada uma determinada ação, ele deve ignorar seus sentimentos e executar tal ação, independentemente dos receios produzidos por alarmes que o corpo possa dar em sentido contrário. No entanto, nada é dito sobre o que significa ser "o melhor que ele tem a fazer". Mesmo quando o indivíduo possui decisões deterministas que, ao executá-las, sabe-se exatamente qual vai ser o resultado da decisão, pode acontecer de mais de uma decisão atingir um objetivo específico. Nesse caso, qual decisão escolher? Qualquer decisão é equivalente? Se um usuário quer ser levado para casa, o caminho mais longo é equivalente ao mais curto? Em geral, não. Mas assim, talvez o objetivo esteja mal especificado e mais informações devam ser adicionadas para melhor especificá-lo.

Não é incomum descrever um desejo como um compromisso entre vários objetivos (KEENEY; RAIFFA, 1976): chegar em casa, caminho curto, tempo pequeno, sem acidentes, sem causar danos ao carro, etc. Em muitos casos, alguns desses objetivos

são suprimidos e também é delegado ao executor o bom senso de atingi-los. Quando o executor é um agente, esse bom senso já não pode ser considerado *a priori* e os desejos devem ser explicitados por completo no compromisso entre vários objetivos, uma vez que nem sempre todos podem ser maximizados ao mesmo tempo.

Ainda, outro fator que deve ser levado em conta ao explicitar um desejo e decisões que satisfaçam tal desejo é como um agente deve se comportar mediante ambientes não deterministas. Nesse caso, as ações não necessariamente resultarão nos objetivos desejados. Em termos de crença, o que se pode conhecer são as chances de uma determinada conseqüência ocorrer ao tomar uma determinada decisão. Então, dada uma decisão, pode-se atribuir probabilidades de ocorrência para cada possível conseqüência.

Agora, "o melhor que ele tem a fazer" deve ser um compromisso entre os vários objetivos e as probabilidades de ocorrência para cada conseqüência e, conseqüentemente, o quanto cada objetivo é atingido em cada uma das conseqüências, para cada possível decisão a tomar. Formalizando em um problema de decisão, o agente tem disponível para escolha um conjunto de decisões $\mathcal{D}$ e um conjunto de possíveis conseqüências Ψ. O agente também possui a crença de que ao tomar uma decisão $d \in \mathcal{D}$, o resultado dessa decisão depende de uma distribuição de probabilidades $\Pr(\psi|d)$ que mapeia a probabilidade de ocorrência da conseqüência $\psi \in \Psi$ quando a decisão d é tomada. Mediante essa formalização, deve-se determinar como escolher a melhor decisão, tópico explorado na próxima seção.

2.2 Decisões e Preferências

Há dois conceitos fundamentais de relação de preferência entre duas conseqüências: "melhor que" ($\succ$) e "indiferente a" ($\sim$). Algumas propriedades sobre essa relação são vistas por alguns como propriedades inerentes à racionalidade, como completude e transitividade.

A propriedade de completude garante que uma preferência possa sempre ser emitida quando comparando duas opções, para isso, relações de preferência entre todas as conseqüências são necessárias para que se possa julgar qual entre duas opções é a melhor. Então, a propriedade de completude exige que, para quaisquer duas conseqüências ψ' e ψ'', uma das seguintes proposições seja verdadeira: a) $\psi' \succ \psi''$, b) $\psi'' \succ \psi'$ ou c) $\psi' \sim \psi''$.

A transitividade permite que a relação de preferência entre duas conseqüências possa ser estendida naturalmente para o julgamento da melhor entre mais de duas

conseqüências, pois ela evita ciclo de preferências entre as conseqüências. A propriedade de transitividade põe que se $\psi' \succ \psi''$ e $\psi'' \succ \psi'''$, então $\psi' \succ \psi'''$.

Até agora, falou-se apenas de julgamentos sobre conseqüências. Mas, como esses julgamentos podem auxiliar de fato na tomada de decisões reais, que serão implementadas no mundo real e que o usuário experimentará as conseqüências? Se as preferências do usuário forem transitivas e completas, pode-se recorrer à noção de racionalidade aristotélica e tomar a decisão que apresentará a melhor conseqüência. Então, dado um conjunto de decisões $\mathcal{D}$, onde cada decisão $d \in \mathcal{D}$ apresenta uma conseqüência $\psi(d)$, existirá uma opção d^* que não é dominada por nenhuma outra, isto é, existe d^* tal que para todo $d \in \mathcal{D}$, ou $\psi(d^*) \succ \psi(d)$ ou $\psi(d^*) \sim \psi(d)$.

Se as propriedades de completude e transitividade não forem observadas, descrever uma decisão racional não é tão simples, e pode até mesmo ser impossível. No caso em que não haja completude, deve-se então ser conservador e eliminar das opções de tomada de decisão apenas decisões d_{min} que sejam dominadas por alguma outra, isto é, se existe $d \in \mathcal{D}$ tal que $\psi(d) \succ \psi(d_{min})$, significando que d pode ser escolhida em detrimento de d_{min}. Nesse caso, não se opta pela melhor decisão possível, mas por aquela que ao menos não se conheça alguma outra que seja melhor.

O caso da falta de transitividade é ainda pior, pois, dado um conjunto com mais de três decisões, pode ocorrer que qualquer decisão dentro desse conjunto seja dominada por outra, levando a comportamentos absurdos. Como exemplo, considere três conseqüências ψ', ψ'' e ψ''', onde $\psi' \succ \psi''$, $\psi'' \succ \psi'''$ e $\psi''' \succ \psi'$. Uma vez que $\psi' \succ \psi''$, pode-se imaginar uma compensação monetária $\$$ em ψ'' de modo que, mesmo assim, $\psi' \succ (\psi'', \$)$ [1]. Neste caso, na posse de ψ'', o usuário trocaria ψ'' juntamente com uma quantia $\$$ por ψ'. Esse exercício hipotético pode ser realizado também para os pares $\psi''' \succ (\psi', \$)$ e $\psi'' \succ (\psi''', \$)$. Logo, se o usuário fizesse três trocas sucessivas, estaria novamente em posse de ψ'' mas teria gasto $3\$$. Se o exercício fosse perpetuado, o usuário gastaria quantias infinitas para continuar transitando sempre entre as mesmas conseqüências, exemplificando, assim, o problema da falta de transitividade.

Quando existe um determinismo entre a decisão tomada e o resultado obtido, uma relação de preferências completa e com transitividade permite tomadas de decisão racionais. No entanto, quando se considera o não determinismo para o efeito das decisões, a simples relação de preferências não é o bastante. Formalmente, ao tomar uma decisão d, existe uma probabilidade $\Pr(\psi|d)$ associada a cada possível conseqüência ψ. O que seria uma decisão racional, nesse caso de não determinismo?

[1] A notação $(\psi'', \$)$ representa uma conseqüência na qual ocorrem ψ'' e $\$$.

As preferências devem ser emitidas sobre outro tipo de objeto, as loterias. Loteria é um mecanismo de sorteio $([\alpha_1;\psi_1],[\alpha_2;\psi_2],\ldots,[\alpha_n;\psi_n])$ entre n conseqüências ψ_i, onde $\alpha_i \geq 0$, $\sum_{i=1}^{n}\alpha_i = 1$ e α_i indica a probabilidade da conseqüência ψ_i ocorrer. Então, define-se o conjunto de loterias $\mathcal{L}$ como todas as loterias possíveis, isto é,

$$\mathcal{L} = \{([\alpha_1;\psi_1],[\alpha_2;\psi_2],\ldots,[\alpha_{|\Psi|};\psi_{|\Psi|}]) \mid \alpha_i \geq 0 \wedge \sum_{i=1}^{|\Psi|}\alpha_i = 1\}.$$

Dadas as preferências sobre as conseqüências, que características podem ser consideradas racionais para as preferências sobre as loterias? Uma característica é a dominância estatística. Considere as loterias $\tilde{x}$ e $\tilde{y}$ baseadas apenas em duas conseqüências ψ', ψ'' tal que $\psi' \succ \psi''$, ou seja, $\tilde{x} = ([\alpha_{\tilde{x}};\psi'],[1-\alpha_{\tilde{x}};\psi''])$ e $\tilde{y} = ([\alpha_{\tilde{y}};\psi'],[1-\alpha_{\tilde{y}};\psi''])$, é racional considerar que $\tilde{x} \succ \tilde{y} \Leftrightarrow \alpha_{\tilde{x}} > \alpha_{\tilde{y}}$, já que a loteria vencedora produziria a conseqüência mais desejada com maior probabilidade. Pode-se estender este conceito para quaisquer loterias com n conseqüências. Sejam $\psi_1, \psi_2, \ldots, \psi_n$ conseqüências ordenadas da melhor (ψ_1) para a pior (ψ_n). Pode-se dizer que uma loteria $\tilde{x} = ([\alpha_{\tilde{x}_1};\psi_1],[\alpha_{\tilde{x}_2};\psi_2],\ldots,[\alpha_{\tilde{x}_n};\psi_n])$ domina estatisticamente uma outra loteria $\tilde{y} = ([\alpha_{\tilde{y}_1};\psi_1],[\alpha_{\tilde{y}_2};\psi_2],\ldots,[\alpha_{\tilde{y}_n};\psi_n])$ se para todo $i = 1, 2, \ldots, n$:

$$\sum_{j=1}^{i}\alpha_{\tilde{x}_j} \geq \sum_{j=1}^{i}\alpha_{\tilde{y}_j}$$

com uma inequação estrita para pelo menos um i.

Outra propriedade interessante em uma decisão racional é a seguinte condição de independência. Seja $\tilde{x}, \tilde{y}, \tilde{z}$ loterias arbitrárias, então:

$$\tilde{x} \succ \tilde{y} \Leftrightarrow \tilde{x}' = ([\alpha;\tilde{x}],[1-\alpha;\tilde{z}]) \succ \tilde{y}' = ([\alpha;\tilde{y}],[1-\alpha;\tilde{z}]), \tag{2.1}$$

isto é, quaisquer duas loterias $(\tilde{x}, \tilde{y})$ que possuem uma determinada ordenação entre elas $(\tilde{x} \succ \tilde{y})$ quando combinadas com outra loteria arbitrária $(\tilde{z})$ sob qualquer taxa de combinação (α), a ordenação entre as loterias resultantes se mantém $(\tilde{x}' \succ \tilde{y}')$, independente da loteria combinada e da taxa de combinação.

2.3 Valores e Teoria da Utilidade Esperada

As propriedades de completude e transitividade permitem estabelecer uma relação de ordem fraca ($\succeq$) sobre todas as conseqüências possíveis, isto é, uma relação binária que apresenta: transitividade ($a \succeq b \wedge b \succeq c \rightarrow a \succeq c$), reflexividade ($a \succeq a$) e totalidade ($a \succeq b \vee b \succeq a$). A relação $\sim$ pode ser interpretada como $a \sim b \Leftrightarrow a \succeq b \wedge b \succeq a$ e a relação $\succ$ pode ser interpretada como $a \succ b \Leftrightarrow a \succeq b \wedge \neg(b \succeq a)$.

Esta relação de ordem, por sua vez, permite atribuir valores $v(\psi)$ a cada possível conseqüência, e se $v(\psi') \geq v(\psi'')$ então $\psi' \succeq \psi''$. Essa função valor $v(\cdot)$ permite tomar decisões racionais quando o agente se encontra em um ambiente determinista. No entanto, ela não diz nada sobre como tomar decisões mediante loterias.

Enquanto os princípios de racionalidade permitem prescrever a tomada de decisão entre algumas loterias, eles não se aplicam a todas elas, pois pode não ocorrer dominância estatística por parte de nenhuma das loterias. Considere as loterias $\tilde{x}' = ([0,4;\psi'],[0,6;\psi'''])$ e $\tilde{x}'' = ([0,5;\psi''],[0,5;\psi'''])$, onde $\psi' \succ \psi'' \succ \psi'''$. Nenhuma das loterias domina a outra, nesse caso como tomar decisões? Esse exercício mostra que a informação de ordenação entre as conseqüências não basta para tomar decisões sobre loterias, sendo também necessária uma informação escalar, indicando quão melhor é uma conseqüência em relação a outra. Se o usuário fosse indiferente entre ψ' e ψ'', a melhor opção seria $\tilde{x}''$. Então, intuitivamente, se a diferença entre ψ' e ψ'' for pequena, talvez compense escolher uma loteria onde a melhor conseqüência da loteria é um pouco menor, como $\tilde{x}''$, mas com uma maior chance de obter tal melhor conseqüência. Se a diferença entre ψ'' e ψ''' for pequena, então, se aumentar a chance de ocorrência da melhor conseqüência não melhora muito a qualidade da loteria, já que a diferença entre as conseqüências é muito pequena, e seria melhor optar por $\tilde{x}'$. Uma solução que considera essa intuição é a TUE.

A TUE tem origem em 1738, quando Bernoulli propõe o seguinte problema hipotético (BERNOULLI, 1954): quanto uma pessoa pagaria para entrar em um jogo onde uma moeda é jogada até obter a primeira cara e o prêmio recebido é de 2^n, onde n é o número de vezes que a moeda foi jogada até obter a primeira cara. O valor monetário esperado em tal jogo é infinito, no entanto pessoas pagariam pouco para entrar em tal jogo. Bernoulli propõe então uma teoria onde as pessoas possuam um valor subjetivo para os valores monetários e suas escolhas sejam baseadas na esperança de tal valor subjetivo, o qual é chamado de utilidade, resultando na TUE. Dessa forma, a TUE fornece um suporte teórico quando se quer definir as preferências de um usuário.

Pode-se atribuir um valor V^d para cada decisão $d \in \mathcal{D}$ tal que, se $V^{d'} > V^{d''}$ para quaisquer decisões $d', d'' \in \mathcal{D}$ e $d' \neq d''$, então a decisão d' é melhor do que a decisão d''. O valor da decisão d calcula a utilidade esperada e é definido por:

$$V^d = \sum_{\psi \in \Psi} \Pr(\psi|d) u(\psi), \tag{2.2}$$

onde $u(\psi)$ é a utilidade da conseqüência ψ.

2.3.1 Axiomas da Teoria da Utilidade Esperada

O significado da função valor de uma decisão é explicitado na forma de um sistema de axiomas determinados por Neumann e Morgenstern (1947). Estes axiomas são válidos se e somente se for utilizada a definição de função valor da equação (2.2).

Abusando da notação, pode-se considerar combinações convexas entre duas loterias $([\alpha_i; \tilde{x}], [\alpha_j; \tilde{y}])$, onde α_i e α_j indicam respectivamente as probabilidades das loterias $\tilde{x}, \tilde{y} \in \mathcal{L}$ ocorrerem. Tem-se então o seguinte teorema (CHANKONG; HAIMES, 1983):

Teorema 2.1 (Axiomas da Teoria da Utilidade Esperada)**.** *Seja $\mathcal{L}$ o conjunto de loterias (geradas de Ψ) e $\succeq$ uma ordenação de preferência sobre $\mathcal{L}$. Então $\succeq$ satisfaz o seguinte sistema de axiomas para qualquer $\tilde{w}, \tilde{x}, \tilde{y}, \tilde{z} \in \mathcal{L}$:*

A1. *$\succeq$ é uma relação de ordem fraca sobre $\mathcal{L}$;*

A2. *se $\tilde{x} \succ \tilde{y}$, então $\tilde{x} \succ ([\alpha; \tilde{x}], [1-\alpha; \tilde{y}]) \succ \tilde{y}$ para todo $\alpha \in]0, 1[$;*

A3. *se $\tilde{x} \succ \tilde{y} \succ \tilde{z}$, então existe α_1 e $\alpha_2 \in]0, 1[$ tal que $([\alpha_1; \tilde{x}], [1-\alpha_1; \tilde{z}]) \succ \tilde{y} \succ ([\alpha_2; \tilde{x}], [1-\alpha_2; \tilde{z}])$;*

A4. *$([\alpha; \tilde{x}], [1-\alpha; \tilde{y}]) = ([1-\alpha; \tilde{y}], [\alpha; \tilde{x}])$ para qualquer $\alpha \in [0, 1]$; e*

A5. *se $\tilde{w} = ([\alpha; \tilde{x}], [1-\alpha; \tilde{y}])$, então $([\beta; \tilde{w}], [1-\beta; \tilde{y}]) = ([\alpha\beta; \tilde{x}], [1-\alpha\beta; \tilde{y}])$;*

se e somente se uma função de valor real V definida sobre $\mathcal{L}$ existe tal que para qualquer $\tilde{x}, \tilde{y} \in \mathcal{L}$

$$\tilde{x} \succ \tilde{y} \Leftrightarrow V(\tilde{x}) > V(\tilde{y}) \tag{2.3}$$

e

$$V(\alpha\tilde{x}, (1-\alpha)\tilde{y}) = \alpha V(\tilde{x}) + (1-\alpha)V(\tilde{y}) \text{ para qualquer } \alpha \in]0, 1[. \tag{2.4}$$

Além disso, se V' é uma outra função de valor real sobre $\mathcal{L}$, V' satisfará 2.3 e 2.4 se e somente se

$$V'(\tilde{x}) = \lambda V(\tilde{x}) + k, \quad \lambda, k \in \mathbb{R}, \quad \lambda > 0.$$

Enquanto os axiomas *A4* e *A5* apenas definem como combinar loterias, os outros axiomas não gozam da mesma característica. O axioma *A4* apenas coloca que a ordem em que as conseqüências são consideradas em uma loteria é irrelevante. Já o axioma *A5* coloca que não há nenhuma preferência por utilizar loterias, ou seja, uma

opção entre uma loteria simples e uma loteria composta (loterias de loterias) que resultem nas mesmas probabilidades acumuladas para as conseqüências é irrelevante.

O axioma *A1* é necessário, caso contrário, mesmo em um ambiente determinista não seria possível tomar decisões. Mesmo assim, permite-se que exista igualdade entre opções, possibilitando que a decisão possa ser irrelevante quando isto ocorre. Como foi discutido na seção 2.2, essa propriedade é considerada como propriedade elementar para a tomada de decisão racional e livre de contexto, ou seja, a avaliação de uma conseqüência não depende de outras conseqüências, mas é uma característica intrínseca de cada conseqüência.

O axioma *A2* postula que uma loteria composta por outras duas loterias deve ter um valor intermediário entre os valores dessas duas loterias. Essa característica garante que loterias estatisticamente dominadas são preteridas quando comparadas com loterias que as dominam. Essa característica é denominada de monoticidade.

O axioma *A3* postula a condição de continuidade, na qual, se, em termos de preferências, uma loteria $\tilde{y}$ encontra-se entre outras duas loterias $\tilde{x}$ e $\tilde{z}$, existe uma loteria entre $\tilde{x}$ e $\tilde{z}$ que é equivalente à loteria $\tilde{y}$. Esse axioma não permite representar condições extremas de preferências, onde um comportamento é infinitamente melhor ou pior que outro. Considere como exemplo uma situação onde dois eventos e_1 e e_2 representam o fato de um aluno na quarta série ser aprovado em matemática e português respectivamente e define-se $\psi_0 = \neg e_1 \wedge \neg e_2$ (reprovado nas duas disciplinas), $\psi_1 = \neg e_1 \wedge e_2 \vee e_1 \wedge \neg e_2$ (aprovado apenas em uma disciplina) e $\psi^* = e_1 \wedge e_2$ (aprovado nas duas disciplinas). Ele seguirá para a quinta série apenas se passar nas duas disciplinas; ainda, mesmo que não mude de série, ser aprovado em uma disciplina é preferível a ser reprovado em ambas. Tem-se que $\psi_0 \prec \psi_1 \prec \psi^*$. Porém, se o aluno aceitar qualquer risco para ser aprovado nas duas disciplinas, não existem α_1 e α_2 capazes de representar suas preferências, isto é, não é possível encontrar uma loteria composta $([\alpha; \psi_0], [1-\alpha; \psi^*])$ que seja equivalente a ψ_1.

Os axiomas aqui postulados também implicam na condição de independência quando duas loterias são associadas a uma terceira loteria comum. Nesse caso, a relação de preferência entre as duas loterias originais é transportada para as loterias compostas (equação (2.1)).

2.3.2 Estrutura da Função Utilidade e Atributos

Uma função pode ser representada de várias formas: analítica, curva, tabela, etc. Quando não há conhecimentos *a priori* sobre a estrutura da função utilidade, uma

opção para descrevê-la é na forma de tabela. Assim, deve-se determinar a utilidade $u(\psi)$ de cada conseqüência $\psi \in \Psi$ com $u : \Psi \rightarrow \mathbb{R}$. Quanto maior for a cardinalidade de Ψ (conseqüências possíveis), maior será a complexidade de encontrar a função utilidade. Além disso, pode ocorrer desse conjunto não ser enumerável, o que tornaria inviável tal tarefa. A adoção de uma estrutura paramétrica conhecida para a função utilidade permite reduzir o problema a um número finito e menor de parâmetros a serem determinados.

Um arcabouço para estabelecer estruturas para as funções utilidade é fazer uso da noção de atributos numéricos. Um atributo é uma quantidade mensurável, cujo valor medido reflete o grau em que um objetivo em particular é atingido (CHANKONG; HAIMES, 1983). Usualmente se considera a existência de tais atributos *sine qua non* para poder descrever as preferências de um usuário com relação a um objetivo, sem os quais não seria possível acessar o nível de conquista de tal objetivo[2].

Após escolher um conjunto de atributos Ξ com k atributos, associa-se a cada conseqüência $\psi \in \Psi$ um vetor de atributos $\boldsymbol{\mu}(\psi)$ em um espaço de dimensão k, isto é, $\boldsymbol{\mu} : \Psi \rightarrow \mathbb{R}^k$, e diz-se que $\mu_i(\psi)$ é uma medida do i-ésimo atributo da conseqüência ψ. O vetor de atributos descreve numericamente uma conseqüência, de modo que esse vetor é o bastante para definir a utilidade de tal conseqüência. Na prática, se existirem duas conseqüências com o mesmo vetor de atributos, mesmo que as conseqüências sejam diferentes sob algum aspecto, as utilidades de tais conseqüências serão iguais. A adoção de atributos, além de ser mais concisa na maioria dos casos, permite o trabalho dentro de um espaço métrico, auxiliando na definição de uma estrutura para a função utilidade.

Keeney e Raiffa (1976) apresentam algumas propriedades desejáveis para nortear a escolha de um conjunto de atributos Ξ; o conjunto Ξ deve: ser completo, ser operacional, permitir decomposição, não apresentar redundância e ser mínimo. Ξ é "completo se todos os aspectos pertinentes do problema de decisão estão nele representados", isto é, ele indica o nível de conquista do objetivo em questão. Ξ é "operacional se ele puder ser utilizado de alguma forma significativa na análise" do problema de decisão, tanto na EP, como na tomada de decisão. Ξ permite decomposição "se é possível simplificação no processo de avaliação por desagregação do problema de decisão em partes", isso normalmente se atinge por meio de relações de algum tipo de independência entre os atributos. Ξ não apresenta redundância "se nenhum aspecto do problema de decisão for considerado mais de uma vez pelos atributos". Ξ é "mínimo se não houver nenhum outro conjunto completo de atributos

[2]Em casos extremos, pode-se utilizar atributos binários para cada possível conseqüência indicando a ocorrência ou não da mesma.

representando o mesmo objetivo com um número menor de elementos" (CHANKONG; HAIMES, 1983).

Nem sempre é possível obter um conjunto de atributos completo e operacional. Embora às vezes exista um conjunto natural de atributos Ξ para um dado objetivo, nem sempre tais atributos são mensuráveis. Suponha o objetivo de maximizar a saúde de uma pessoa, embora um atributo com relação ao nível de saúde seja ideal, não há um atributo mensurável que relaciona esse nível de saúde. No entanto, pode-se escolher um segundo conjunto de atributos Ξ' que mede indiretamente este objetivo: nível de colesterol, pressão arterial, temperatura corpórea, etc. Esses atributos são chamados de atributos representantes[3], pois eles refletem o grau que um objetivo associado é atingido, mas não mede diretamente o objetivo.

Mais formalmente, suponha que um objetivo possa ser medido pelo conjunto de atributos Ξ gerando medidas $\boldsymbol{\mu}$ e exista um conjunto de atributos representantes Ξ' gerando medidas $\boldsymbol{\mu}'$, pode-se então calcular uma função utilidade induzida u' sobre o vetor de atributos $\boldsymbol{\mu}'$:

$$u'(\boldsymbol{\mu}') = \sum_{\psi \in \Psi} \Pr(\psi|\boldsymbol{\mu}') u(\boldsymbol{\mu}(\psi)), \tag{2.5}$$

onde $\Pr(\psi|\boldsymbol{\mu}')$ é a probabilidade condicional de ter ocorrido a conseqüência ψ dado que os atributos representantes $\boldsymbol{\mu}'$ foram observados. Então, se não é possível medir os atributos Ξ, mas apenas os atributos Ξ', $u'(\boldsymbol{\mu}')$ pode ser utilizada como estimativa da função utilidade desejada $u(\boldsymbol{\mu})$.

2.3.3 Propriedades Qualitativas de uma Função Utilidade

Como foi dito na seção anterior, o uso de atributos pode facilitar uma descrição paramétrica e menos custosa da função utilidade que representa os objetivos do usuário. No entanto, a função utilidade deve apresentar alguma estrutura sobre estes atributos, caso contrário o problema torna-se tão difícil quanto o de representar a função utilidade no espaço de conseqüências.

Nesta seção serão apresentadas algumas propriedades qualitativas que ajudam a formalizar a estrutura de uma função utilidade. Elas serão apresentadas num espaço unidimensional, mas nada impede que sejam estendidas para espaços multidimensionais. No caso unidimensional, uma conseqüência ψ é representada por um único atributo com valor $\mu(\psi)$.

[3] Do inglês *proxy atributes*.

Se para todo μ', μ''

$$\mu' > \mu'' \iff u(\mu') > u(\mu''),$$

então $u(\cdot)$ é uma função monotônica crescente no valor do atributo μ. Isso é verdade, por exemplo, em problemas que consideram valores monetários: usualmente, quanto maior o lucro (dinheiro) melhor é a situação. No entanto, nem todos problemas possuem essa propriedade; por exemplo, o consumo de açúcar por um ser humano é bom até um certo limite, a partir do qual um acréscimo passa a ser prejudicial. O caso decrescente ($\mu' > \mu'' \iff u(\mu') < u(\mu'')$) sempre pode ser transformado no caso crescente, bastando para isso realizar uma transformação afim negativa no atributo medido.

Outra propriedade interessante é a propensão, aversão ou neutralidade ao risco. Considere uma loteria com probabilidades α_i para os valores de atributo μ^i, para $i = 1, 2, \ldots, n$. Primeiramente três conceitos são definidos: valor de atributo esperado de uma loteria, equivalente assegurado[4] de uma loteria e loteria não degenerada.

O valor de atributo esperado $\bar{\mu}$ para essa loteria é:

$$\bar{\mu} = \sum_{i=1}^{n} \alpha_i \mu^i.$$

O valor de atributo $\hat{\mu}$ é um equivalente assegurado dessa loteria se $\hat{\mu}$ é tal que o usuário é indiferente entre a loteria e o vetor de atributos $\hat{\mu}$ c.p.1. Portanto, o conjunto de equivalentes assegurados é definido por:

$$\{\hat{\mu} \mid u(\hat{\mu}) = \mathrm{E}_{\{\mu \sim \alpha_1, \alpha_2, \ldots, \alpha_n\}}[u(\mu)]\}.$$

Se a função utilidade que representa o usuário é monotônica, então o equivalente assegurado é único e definido por:

$$\hat{\mu} = u^{-1}(\mathrm{E}_{\{\mu \sim \alpha_1, \alpha_2, \ldots, \alpha_n\}}[u(\mu)]).$$

A loteria é degenerada se existe i, tal que $\alpha_i = 1$, implicando $\alpha_j = 0$ para $j \neq i$ e a loteria é não degenerada caso contrário. Têm-se então a seguinte definição e o seguinte teorema (CHANKONG; HAIMES, 1983):

Definição 2.1. *Considere um usuário com uma função utilidade $u(\cdot)$ e loterias $\tilde{x} \in \mathcal{L}$ com os respectivos valores de atributos esperados $\bar{\mu}_{\tilde{x}}$ e equivalentes assegurados $\hat{\mu}_{\tilde{x}}$. Então:*

- *o usuário é averso ao risco se para toda loteria $\tilde{x} \in \mathcal{L}$ não degenerada tem-se*

[4] Do inglês *certainty equivalent*.

que $u(\bar{\mu}_{\tilde{x}}) > u(\hat{\mu}_{\tilde{x}})$*;*

- *o usuário é propenso ao risco se para toda loteria* $\tilde{x} \in \mathcal{L}$ *não degenerada tem-se que* $u(\bar{\mu}_{\tilde{x}}) < u(\hat{\mu}_{\tilde{x}})$*; e*
- *o usuário é neutro ao risco se para toda loteria* $\tilde{x} \in \mathcal{L}$ *não degenerada tem-se que* $u(\bar{\mu}_{\tilde{x}}) = u(\hat{\mu}_{\tilde{x}})$*.*

Teorema 2.2. *Um usuário é averso [propenso, neutro] ao risco se e somente se sua função utilidade é côncava [convexa, linear].*

Por exemplo, considere uma aposta onde se pode perder ou ganhar 100 unidades monetárias com probabilidades iguais de valor 0,5. Então, o valor de atributo esperado é 0. Tem-se que o usuário é averso ao risco se ele preferir não apostar a arriscar perder as 100 unidades monetárias. Já o usuário propenso ao risco prefere apostar e tentar ganhar as 100 unidades monetárias. O usuário neutro ao risco é indiferente com relação a apostar ou não apostar.

2.3.4 Função Valor Aditiva

O uso de atributos pode facilitar a EP, uma vez que ele transforma o espaço de conseqüências em um espaço métrico, possibilitando o uso de técnicas de interpolação, regressão, entre outras para adaptar uma curva ou superfície a dados já obtidos junto ao usuário. Além disso, informações qualitativas considerando tais atributos podem ser obtidas junto ao usuário para limitar os tipos de funções que serão utilizadas no processo de EP.

No entanto, o uso de atributos também pode apresentar alguns problemas. Se os atributos utilizados não permitirem decomposição e o conjunto de atributos for grande, o processo de EP pode ser bastante custoso. Primeiro, o usuário pode ter dificuldade em avaliar questões realizadas num espaço n-dimensional, onde vários atributos são considerados e deve-se observar o compromisso entre eles. Segundo, ao adaptar uma superfície neste espaço n-dimensional, o número de informações necessárias para se obter uma aproximação satisfatória pode tornar-se muito grande.

Nesta seção será apresentada uma propriedade de independência entre os atributos, a qual permite que o problema de obter uma superfície de dimensão $(n+1)$ seja decomposto no problema de obter n curvas bidimensionais e fatores de escala entre tais curvas.

Definição 2.2. *Os atributos* $i \in \Xi$ *possuem independência aditiva se preferências sobre loterias dos atributos* $i \in \Xi$ *dependem apenas das respectivas distribuições de*

probabilidades marginais e não das distribuições de probabilidades conjuntas.

Sejam μ_i^0 e μ_i^* respectivamente o pior e o melhor valor possível para o atributo i. A condição de independência aditiva permite formular o seguinte teorema (CHANKONG; HAIMES, 1983).

Teorema 2.3. *A função utilidade aditiva com n-atributos*

$$u(\boldsymbol{\mu}) = \sum_{i=1}^{n} u(\mu_1^0, \mu_2^0, \ldots, \mu_{i-1}^0, \mu_i, \mu_{i+1}^0, \ldots, \mu_n^0) = \sum_{i=1}^{n} w_i u_i(\mu_i)$$

é apropriada se e somente se a condição de independência aditiva aplica-se aos atributos $i \in \Xi$, onde:

1. *u é normalizada por $u(\mu_1^0, \mu_2^0, \ldots, \mu_n^0) = 0$ e $u(\mu_1^*, \mu_2^*, \ldots, \mu_n^*) = 1$,*
2. *u_i é uma função utilidade condicional do atributo i normalizada por $u_i(\mu_i^0) = 0$ e $u_i(\mu_i^*) = 1$, para todo $i \in \Xi$, e*
3. *$w_i = u(\mu_1^0, \mu_2^0, \ldots, \mu_{i-1}^0, \mu_i^*, \mu_{i+1}^0, \ldots, \mu_n^0)$ para todo $i \in \Xi$.*

Quando a propriedade de independência aditiva é observada, pode-se obter a função utilidade para cada atributo independentemente. Ao obter a função utilidade para o atributo i, escolhe-se valores fixos para os atributos restantes, por exemplo, os piores valores $\mu_1^0, \mu_2^0, \ldots, \mu_{i-1}^0, \mu_{i+1}^0, \ldots, \mu_n^0$, e determina-se $u_i(\mu_i)$ como se fosse uma função utilidade unidimensional.

Embora essa seja uma restrição forte, intuitivamente ela se aplicará a todas situações nas quais o conjunto de atributos utilizados for realmente fundamental (completo e mínimo). No caso extremo, existe apenas um atributo que mede o objetivo (CARENINI; POOLE, 2002). No entanto, nem sempre se tem acesso a tais atributos fundamentais e deve-se recorrer a atributos representantes. Um caso mais restrito de função valor aditiva é quando o usuário, que a função utilidade representa, é neutro ao risco. Nesse caso, pode-se adotar funções $u_i(\cdot)$ lineares, sendo necessário apenas encontrar os fatores de escala w_i. Tem-se então a seguinte definição e corolário.

Definição 2.3. *Os atributos $i \in \Xi$ possuem independência aditiva e são neutros ao risco se preferências sobre loterias dos atributos $i \in \Xi$ dependem apenas dos respectivos valores esperados e não das distribuições de probabilidades.*

Corolário 2.1. *A função utilidade aditiva e afim com n-atributos*

$$u(\boldsymbol{\mu}) = \sum_{i=1}^{n} u(\mu_1^0, \mu_2^0, \ldots, \mu_{i-1}^0, \mu_i, \mu_{i+1}^0, \ldots, \mu_n^0) = \sum_{i=1}^{n} w_i \mu_i + K \qquad (2.6)$$

é apropriada se e somente se a condição de independência aditiva e neutralidade ao risco aplica-se aos atributos $i \in \Xi$, onde:

1. *u é normalizada por $u(\mu_1^0, \mu_2^0, \ldots, \mu_n^0) = 0$ e $u(\mu_1^*, \mu_2^*, \ldots, \mu_n^*) = 1$,*
2. *$w_i = \frac{u(\mu_1^0, \mu_2^0, \ldots, \mu_{i-1}^0, \mu_i^*, \mu_{i+1}^0, \ldots, \mu_n^0)}{\mu_i^* - \mu_i^0}$ para todo $i \in \Xi$, e*
3. *$K = -\sum_{i=1}^{n} u(\mu_1^0, \mu_2^0, \ldots, \mu_{i-1}^0, \mu_i^*, \mu_{i+1}^0, \ldots, \mu_n^0) \frac{\mu_i^0}{\mu_i^* - \mu_i^0}$.*

Demonstração. Considerando que a função utilidade de cada atributo seja afim, isto é, $u_i(\mu_i) = w_i'' \mu_i + k_i$, e ainda a condição de normalização $u_i(\mu_i^0) = 0$ e $u_i(\mu_i^*) = 1$, tem-se que:

$$w_i'' = \frac{1}{\mu_i^* - \mu_i^0} \text{ e } k_i = -\frac{\mu_0}{\mu_i^* - \mu_i^0}.$$

Então, pelo teorema 2.3, tem-se que:

$$u(\boldsymbol{\mu}) = \sum_{i=1}^{n} w_i' u_i(\mu_i) = \sum_{i=1}^{n} w_i' w_i'' \mu_i + w_i' k_i = \sum_{i=1}^{n} w_i' w_i'' \mu_i + \sum_{i=1}^{n} w_i' k_i.$$

Lembrando que $w_i' = u(\mu_1^0, \mu_2^0, \ldots, \mu_{i-1}^0, \mu_i^*, \mu_{i+1}^0, \ldots, \mu_n^0)$ e definindo $w_i = w_i' w_i''$ está provado o corolário. □

Esse tipo de função utilidade permite que qualquer subdivisão de uma conseqüência seja avaliada separadamente, já que o aumento ou diminuição do valor de um atributo sempre ocorre de forma linear e que os atributos são independentes entre si.

2.4 Extração de Preferências

A função utilidade pode ser utilizada para representar os objetivos de um usuário, de modo que um agente possa representá-lo e agir em seu lugar. Uma forma de representação pode ser escolhida de modo a permitir que o usuário prescreva suas preferências, ou ainda, como é considerado neste trabalho, pode-se criar um método que facilite essa descrição das preferências, isto é, a EP.

O processo de EP consiste em formular questões e submetê-las ao usuário. O usuário emite respostas consoante às questões e considera-se que tais respostas reflitam suas preferências. Munido de um conjunto de pares questão-resposta, o agente deve definir uma função utilidade que represente as preferências do usuário. Esta função utilidade é então utilizada para guiar a escolha de decisão ótima para o agente agir atendendo às expectativas do usuário. Esse processo consiste na repetição dos seguintes passos:

- Agente formula uma questão e submete ao usuário.
- O usuário responde à questão.
- Agente infere do par questão-resposta informações parciais sobre as preferências do usuário.

Quando o agente achar suficientes as preferências parciais obtidas, ele pode tomar decisões pelo usuário que atendam às preferências deste. Esta decisão é baseada em uma função utilidade estimada (ver figura 2.1).

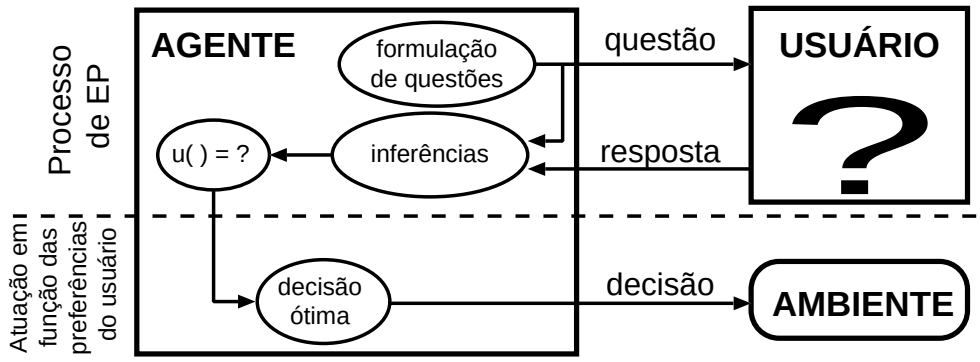

Figura 2.1: Modelo para Extração de Preferências. O agente formula questões, submete as questões ao usuário e o usuário emite respostas. As relações entre questões e respostas são utilizadas para inferir uma função utilidade $u(\cdot)$, que, uma vez aprendida, guiará a decisão ótima do agente para atuar no ambiente em nome do usuário.

Nos últimos anos, trabalhos foram produzidos que formalizaram a EP, especificando a relação entre agente, usuário e decisão ótima (CHAJEWSKA; KOLLER; PARR, 2000; BOUTILIER et al., 2005). A interação entre agente e usuário é modelada por questões e respostas. Define-se então um conjunto de questões possíveis $\mathcal{Q}$ e, para cada questão $q \in \mathcal{Q}$, um respectivo conjunto de respostas possíveis $\mathcal{R}_q$. A semântica envolvida nas perguntas deve ser de comum conhecimento, tanto para o agente como para o usuário. As questões formuladas são usualmente hipotéticas, não envolvendo, no processo de EP, o ambiente onde as decisões serão aplicadas.

Considera-se também um conjunto $\mathcal{U}$ de funções utilidade candidatas. As respostas do usuário às questões são confrontadas com as predições de respostas de cada função utilidade $u \in \mathcal{U}$, testando a aderência das funções utilidade às preferências do usuário. Para isso, é necessário um modelo de respostas do usuário, ou seja, como ele escolhe suas respostas para cada questão dada suas preferências, aqui modeladas como funções utilidade. Este modelo é essencial para inferir, baseado nas respostas do usuário, quais são suas preferências. Define-se então um modelo probabilístico $\Pr(r|q,u)$ que determina a probabilidade da resposta $r \in \mathcal{R}_q$ ser dada à questão q

quando o usuário possui suas preferências modeladas pela função utilidade u. Com tal modelo, pode-se inferir probabilidades $\Pr(u)$ de aderência às informações obtidas para cada função utilidade candidata $u \in \mathcal{U}$.

Dado o modelo probabilístico do usuário, ocorre que as funções utilidade candidatas possuem diferentes níveis de aderência às respostas do usuário, não permitindo a escolha de uma única função utilidade que represente adequadamente o usuário. Se poucos pares questão-resposta foram obtidos, essa condição torna-se ainda mais visível. Mas é desejável que uma decisão possa ser tomada em qualquer parte do processo de EP. Então, deve-se especificar qual é o modelo para tomada de decisão, isto é, dado um conjunto de relações questão-resposta qual é a decisão ótima a ser tomada pelo agente. Define-se também um conjunto $\mathcal{D}$ de possíveis decisões d a serem tomadas e o efeito de cada decisão no ambiente, modelado na distribuição de probabilidade $\Pr(\psi|d)$.

A principal característica do problema de EP não é realizar a inferência, nem mesmo como é feita a tomada de decisão. Esses processos são tratados em outras frentes de pesquisa e na EP são utilizados como ferramentas. O cerne da EP está em formular questões informativas, ou seja, que causarão um maior discernimento entre as funções utilidade candidatas $\mathcal{U}$, ou, como objetivo final, um maior discernimento entre as possíveis tomadas de decisões $\mathcal{D}$. Deve-se então definir um modelo de otimização que guie o processo de EP, isto é, como formular uma entre as questões $\mathcal{Q}$ para obter as preferências do usuário. Ao formular questões, é usual que o agente considere os dados, tanto conhecidos *a priori* como obtidos no processo de EP, que ocorrem na inferência ou tomada de decisão. Dessa forma, um diagrama que representa de forma detalhada o processo de EP é representado na figura 2.2. Nesta seção, será discutido como cada um desses pontos são tratados na EP.

2.4.1 Questões Hipotéticas

Os trabalhos de EP consideram questões hipotéticas, isto é, questões que não precisam de nenhuma interação com o ambiente para serem submetidas pelo agente ao usuário (KEENEY; RAIFFA, 1976). Ao responder às questões, o usuário proporciona informações sobre suas preferências e os métodos devem determinar então quais questões devem ser formuladas, buscando principalmente: 1) realizar uma tomada de decisão ótima segundo as preferências do usuário; e 2) garantir um "baixo" número de interações com o usuário.

Para obter a primeira meta não é necessário ter uma função utilidade com acurácia em todo o espaço de conseqüências, mas apenas o bastante para garantir a

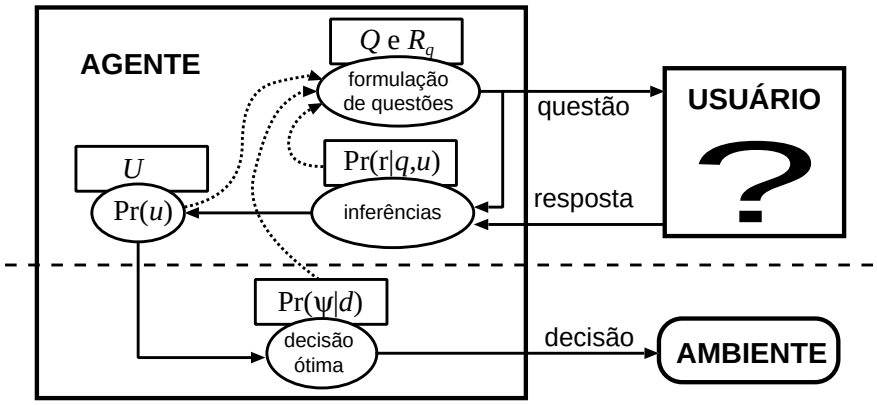

Figura 2.2: Modelo detalhado para Extração de Preferências. Para tomar decisões, o agente considera o efeito de cada decisão no ambiente dado por $\Pr(\psi|d)$ e no usuário $\Pr(u)$. A aderência das funções utilidade são inferidas pelas respostas do usuário às questões e um modelo de respostas para usuários genéricos $\Pr(r|q,u)$. Para formular questões, o agente considera toda informação disponível, tanto dada *a priori*, como obtida no processo de EP.

decisão ótima (LAHAIE; PARKES, 2004). Se a obtenção de uma decisão semi-ótima é considerada, então um compromisso entre o primeiro e o segundo objetivo deve ser observado. Intuitivamente, quanto mais interações forem realizadas com o usuário, maior será a acurácia da função utilidade, proporcionando uma melhor tomada de decisão.

Ao mensurar a segunda meta, algumas dificuldades são encontradas, pois, dependendo do tipo de interação realizada com o usuário, a interação pode proporcionar um menor ou maior desconforto ao mesmo. Dessa forma, deve-se quantificar melhor o quanto cada uma das interações (questão-resposta) exige do usuário. Outro ponto sobre o tipo de interação com o usuário que merece atenção é o tipo de informação resultante da interação, já que se pode considerar a qualidade de tal informação tanto do ponto de vista informativo, isto é, quanto muda a acurácia da função utilidade estimada com o recebimento de tal informação, ou ainda, sob o ponto de vista psicológico, isto é, quão confiável é a resposta disponibilizada pelo usuário.

O uso de questões hipotéticas, ao invés de questões com situações ou conseqüências reais, permite maximizar a informação disponibilizada em cada interação. Essas questões contemplam conseqüências hipotéticas, que não necessariamente ocorrem no problema de decisão. Além disso, também podem ser colocadas opções na tomada de decisão, que não necessariamente são opções reais para a tarefa em questão, resultando em distribuições de probabilidades para as conseqüências que não possuem uma decisão real correspondente. Em geral, essas questões envolvem comparações entre diversas opções, sejam essas opções no nível de conseqüências ou no nível de loterias (KEENEY; RAIFFA, 1976; DENNIS; HEALEY, 2003; CHEN; PU, 2004).

Comparação entre Conseqüências A questão mais simples que se pode considerar é aquela à respeito da comparação entre pares de conseqüências. Nesse tipo de questão é oferecida ao usuário a escolha entre duas conseqüências $\psi', \psi'' \in \Psi$. Essa avaliação relativa é fácil de ser respondida, uma vez que pessoas podem distinguir facilmente entre opções de forma qualitativa. No entanto, se modelos estocásticos forem utilizados para representar a avaliação que um usuário faz sobre as conseqüências (HEY, 1995), no início da EP essas questões são mais facilmente respondidas, onde se elimina as piores conseqüências comparando opções que diferem bastante em seus atributos. Mas, ao fim do processo, quando se deseja refinar as escolhas, responder a essas questões pode ser mais difícil, apresentando inconsistência ou requerendo um maior esforço cognitivo.

Ao responder questões que comparam duas conseqüências, apenas valores de ordenação são obtidos, permitindo, por exemplo, que seja obtida uma ordenação de todas as conseqüências de acordo com as preferências de uma pessoa. No entanto, nenhuma informação absoluta sobre as utilidades das conseqüências pode ser inferida (BASU, 1982). Além disso, se o ambiente for estocástico, a ordenação não é o bastante para definir a tomada de decisões.

Valores Absolutos Valores de utilidade para conseqüências são mais informativos, uma vez que tal informação pode ser atribuída diretamente a tais conseqüências e com tais valores pode-se construir as mesmas informações obtidas utilizando questões sobre comparação entre conseqüências. Nesse caso, o usuário recebe uma conseqüência ψ e deve avaliá-la com um valor $u(\psi)$. Embora esta seja uma informação ideal para a obtenção da utilidade de ψ, não é fácil para um ser humano trabalhar com avaliações quantitativas sobre uma conseqüência, uma vez que o usuário deve ter consciência de todo o problema de decisão para estipular valores que se adeqüem à TUE. Mesmo que esse tipo de questão seja feito utilizando limiares – por exemplo, "a conseqüência ψ tem um valor maior ou menor que 5?" – não é fácil para o usuário estabelecer tais valores. No entanto, comparações envolvendo loterias também podem gerar informações absolutas quando questões são formuladas de forma apropriada, mas são mais fáceis de serem respondidas que questões diretas sobre o valor de uma conseqüência.

Comparação entre Loterias Na sua forma mais simples, denominada de loteria padrão, a questão é formulada sobre a preferência do usuário entre a conseqüência ψ c.p.1 e uma loteria $\tilde{x} = ([\alpha; \psi'], [1 - \alpha; \psi''])$. Essa questão permite obter informação sobre a utilidade da conseqüência ψ e o equivalente assegurado $\hat{x}$ da loteria, estabelecendo relações de preferências sobre os mesmos. Se as conseqüências envolvidas na loteria forem $\psi'' = \psi^0$ e $\psi' = \psi^*$, a pior e a melhor conseqüência respectivamente,

então a utilidade dessas conseqüências podem ser normalizadas por $u(\psi^0) = 0$ e $u(\psi^*) = 1$. Além disso, o valor do equivalente assegurado $\hat{x}$ é conhecido e vale $u(\hat{x}) = \alpha$. Vale também que, para toda conseqüência, existirá alguma loteria que seja equivalente (axioma *A3* do teorema 2.1). Então, em sua forma mais simples, comparação entre loterias permite obter valores limiares do valor da utilidade da conseqüência ψ quando comparada à loteria $\tilde{x} = ([\alpha; \psi^*], [1 - \alpha; \psi^0])$. Se valores sobre uma conseqüência ou sobre uma loteria são desejáveis, sucessivas questões podem ser utilizadas para diminuir tais limiares.

2.4.2 Modelo do Usuário

A primeira pergunta que se pode fazer a respeito do usuário é se ele é consistente, isto é, se para quaisquer perguntas feitas em tempos diferentes $t' \neq t''$ tal que $q_{t'} = q_{t''}$ então as respostas do usuário são iguais $r_{t'} = r_{t''}$. Podem existir perguntas com as quais seja difícil obter consistência, como: perguntas sobre valores, já que podem ocorrer avaliações distintas feitas pelo usuário para uma mesma conseqüência; e perguntas sobre comparação seja entre conseqüências, quando as conseqüências envolvidas possuem vetores de atributos semelhantes, seja entre loterias, quando as loterias envolvidas possuem probabilidades de ocorrências semelhantes.

Se o usuário não for consistente, um segundo fato a considerar é se suas preferências são estacionárias. Uma vez que o usuário não é consistente, pode-se modelar suas respostas com base em uma distribuição de probabilidades nas possíveis respostas $r \in \mathcal{R}_q$, $\Pr(r_t = r | q_t = q, \mathrm{Us})$, onde Us explicita a dependência junto ao usuário. Um usuário possui preferências estacionárias se para quaisquer perguntas feitas em tempos diferentes $t' \neq t''$ tal que $q_{t'} = q_{t''} = q$ então para toda resposta $r \in \mathcal{R}_q$ tem-se que $\Pr(r_{t'} = r | q_{t'} = q, \mathrm{Us}) = \Pr(r_{t''} = r | q_{t''} = q, \mathrm{Us})$.

Considerar um usuário com preferências estacionárias nem sempre é uma representação realista do usuário. Um usuário pode mudar a distribuição de probabilidades por influência do processo. Por exemplo, se ele é questionado pela preferência entre as conseqüências ψ' e ψ'' novamente, é possível que ele mantenha a resposta exibida quando questionado pela primeira vez, principalmente se o espaço de tempo decorrido entre as duas questões for curto. Alguns advogam que as preferências são construídas, através de um processo de deliberação, quando a necessidade de emitir julgamento de preferência entre conseqüências é enfrentada, seja de forma hipotética ou de forma real (PAYNE; BETTMAN; JOHNSON, 1992). No próximo capítulo será apresentada a teoria de Plott (1996), na qual as preferências do usuário são vistas como descobertas pelo próprio, contrariando a teoria construtiva. Por ora considera-se que

as preferências do usuário são estacionárias.

Ainda, outra suposição que pode ser feita é a existência de uma estrutura fixa para as distribuições de probabilidades. O uso de funções utilidade está na base de vários trabalhos de EP. Uma função utilidade representa as preferências de um usuário consistente, cujas avaliações das conseqüências são sempre as mesmas. Pode-se tornar esse modelo mais genérico considerando que o usuário possui uma função utilidade u_{Us} e que a distribuição de probabilidades para a pergunta $q \in \mathcal{Q}$ seja dependente de u_{Us} e que essa dependência seja conhecida pelo agente. Nesse caso o agente tem conhecimento de distribuições de probabilidades $\Pr(r|q, u_{\mathrm{Us}} = u)$ para todo $q \in \mathcal{Q}$, $r \in \mathcal{R}_q$ e $u \in \mathcal{U}$, onde $\mathcal{U}$ é o conjunto de todas funções utilidade candidatas consideradas pelo agente.

Considere, por exemplo, a questão sobre a preferência entre duas loterias; quanto mais próximos forem os valores das loterias, isto é, as suas utilidades esperadas, mais difícil será para o usuário ser consistente. No caso extremo, quando as duas loterias possuem valores iguais, a probabilidade de escolher cada uma delas deveria ser 50%. Um modelo que possui essa propriedade considera que uma decisão entre duas loterias $\tilde{x}$ e $\tilde{y}$ é baseada em uma função $V(\tilde{x}, \tilde{y})$ que representa a real preferência do usuário, isto é, $\tilde{x}$ é preferido sobre $\tilde{y}$ se e somente se $V(\tilde{x}, \tilde{y}) > 0$. No entanto, o usuário utiliza essa regra com algum erro aleatório ϵ. Então, $\tilde{x}$ é considerada a loteria preferida se $V(\tilde{x}, \tilde{y}) + \epsilon > 0$ (HEY, 1995). Considerando a teoria da utilidade, tem-se que $V(\tilde{x}, \tilde{y}) = V(\tilde{x}) - V(\tilde{y})$ e resta apenas definir as propriedades do erro aleatório ϵ.

O erro ϵ é considerado como uma variável aleatória com média nula. Pode-se também considerar que tal variável possua uma distribuição normal com variância σ[5]. Resta ainda definir as propriedades de tal variância. Quando se opta pela parcimônia, pode-se considerar a variância constante para qualquer par de loterias. Considerando opções mais complexas, Hey (1995) observa em experimentos que uma variância que diminui com o tempo gasto pelo usuário para responder uma questão é a que melhor se adequa aos dados. Isso pode implicar que quanto maior o tempo de deliberação, mais seguro de sua resposta está o usuário. Por último, se $V(\tilde{x}, \tilde{y}) > 0$ então $\Pr(\tilde{x} \succ \tilde{y}) > 0,5$ para qualquer valor de σ, pois a distribuição normal é simétrica.

[5]Embora modelos mais complexos de como o usuário responde a questões possam ser considerados (JUNG; HONG; KIM, 2005; JUNG; HONG; KIM, 2002; ZUKERMAN; ALBRECHT, 2001), para os propósitos desta tese, o modelo simples aqui adotado permite explorar as características da proposta da tese.

2.4.3 Regra de Inferência

Dado um modelo de como o usuário responde às questões, pode-se discutir como as resposta do usuário serão utilizadas para inferir as suas preferências. Considere que se tem o conhecimento *a priori* da distribuição de probabilidades $\Pr(u)$ sobre o conjunto de funções utilidade possíveis. Se não há nenhum conhecimento sobre tal distribuição, pode-se considerar uma distribuição uniforme. A cada resposta recebida, este modelo deve ser atualizado para refletir o novo conhecimento obtido (CHAJEWSKA; KOLLER; PARR, 2000). Se o modelo de usuário na seção 2.4.2 for aplicado, esta operação pode-se valer da regra de Bayes:

$$\Pr(u|q,r) \leftarrow \frac{\Pr(r|q,u)\Pr(u)}{\Pr(r|q)} = \frac{\Pr(r|q,u)\Pr(u)}{\sum_{u\in\mathcal{U}}\Pr(r|q,u)\Pr(u)}.$$

Mesmo no caso no qual o usuário é consistente e as respostas às questões representam restrições às possíveis funções utilidade, esse modelo também se aplica, uma vez que, nesse caso, o que ocorre é que existem algumas funções utilidade u' para as quais tem-se $\Pr(r|q,u') = 0$, indicando uma restrição a tais funções utilidade.

Um problema comum à EP é que muitas perguntas podem ser necessárias até que uma decisão possa ser tomada com acurácia. O número de questões envolvidas depende tanto dos atributos envolvidos, que fazem com que um maior número de parâmetros precisem ser definidos, como das decisões possíveis, pois a precisão com que a função utilidade é definida deve ser adequada às diferenças entre as conseqüências resultantes das decisões possíveis.

Quando se considera uma distribuição de probabilidades uniforme sobre todas as possíveis funções utilidade que se adeqüem a uma determinada estrutura de função utilidade escolhida, assume-se completo desconhecimento sobre a distribuição dessas probabilidades na população. Se funções utilidade para um número suficiente de pessoas forem conhecidas, pode-se estimar uma distribuição de probabilidades inicial para as possíveis funções utilidade, caso seja considerado que, em geral, existem tipos de pessoas que possuem preferências parecidas.

Essa distribuição pode auxiliar no processo de EP pois, em posse de funções utilidade mais prováveis, pode-se estabelecer questões com o intuito de distinguir apenas entre tais funções. A EP torna-se então um problema de classificação, no qual deve-se classificar um usuário entre os possíveis estereótipos (CHAJEWSKA et al., 1998; QIN; BUFFETT; FLEMING, 2008). Para definir estes estereótipos como representantes de uma classe de usuários, é necessário ter uma medida de distância entre as preferências dos usuários, sejam essas preferências representadas em uma função utilidade,

ou em um conjunto de pares questão-resposta. Ha e Haddawy (1998) apresentam algumas possibilidades para definição de tal distância.

Em posse de uma função utilidade mais provável entre as funções utilidade conhecidas, pode-se iniciar um processo de refinamento dessas preferências, considerando a possibilidade de pequenas variações dentro desse estereótipo. Mesmo que o refinamento não ocorra, a decisão baseada no estereótipo pode ser adequada em várias tarefas, nas quais as possíveis decisões diferem bastante em suas conseqüências resultantes ou nas probabilidades de ocorrência de tais conseqüências.

2.4.4 Tomando Decisões com Incerteza sobre Funções Utilidade

Nem sempre se pode obter a informação necessária para descrever integralmente a função utilidade. Mesmo com as restrições impostas pelas informações obtidas junto ao usuário e as restrições a tipos específicos de estruturas escolhidas para a função utilidade, várias funções utilidade ainda podem atender a tais restrições. Portanto, deve-se ainda definir como uma decisão deve ser tomada.

Uma opção é escolher uma decisão d^*, de tal forma a minimizar o arrependimento (*regret*) de ter escolhido d^* segundo a distribuição das probabilidades de aderência $\Pr(u_{\mathrm{Us}} = u)$. Se u_{Us} fosse conhecida, poder-se-ia definir o arrependimento de tomar uma decisão d como $\mathrm{Regret}(d, u_{\mathrm{Us}}) = V_{u_{\mathrm{Us}}}^{d^*_{u_{\mathrm{Us}}}} - V_{u_{\mathrm{Us}}}^{d}$, onde $d^*_{u_{\mathrm{Us}}}$ representa a decisão ótima para a função utilidade u_{Us}, e as decisões $d^*_{u_{\mathrm{Us}}}$ e d são avaliadas pela função utilidade u_{Us} (WANG; BOUTILIER, 2003; BOUTILIER et al., 2005). Pode-se definir genericamente o arrependimento de tomar a decisão d frente a uma função utilidade qualquer u, isto é,

$$\mathrm{Regret}(d, u) = V_u^{d^*_u} - V_u^d.$$

Então, define-se d^* como sendo a decisão que causa o menor arrependimento quando considerando o pior caso dentre as funções utilidade candidatas $\mathcal{U}$, ou seja,

$$d^* = \arg\min_{d \in \mathcal{D}} \max_{u \in \mathcal{U}} \mathrm{Regret}(d, u).$$

Essa definição representa uma tomada de decisão conservadora quando não se sabe nada a respeito da distribuição de probabilidade das funções utilidade em $\mathcal{U}$. Se $\mathcal{U}$ possui apenas um elemento, o arrependimento mínimo torna-se 0. Essa medida pode ser utilizada também como uma medida da acurácia da EP até o momento, uma vez que, quanto menor o arrependimento, maior a precisão na tomada de decisão.

Ao comparar arrependimentos $\mathrm{Regret}(d, u)$ é necessário definir uma normaliza-

ção, senão o arrependimento $\mathrm{Regret}(d, u)$, quando diferente de 0, pode ser arbitrário. Uma normalização possível é garantir que os valores das piores e melhores conseqüências sejam iguais, isto é, seja ψ_i^0 e ψ_i^* as piores e melhores conseqüências, respectivamente, dada a função utilidade u_i. Então, para quaisquer funções utilidade $u_i, u_j \in \mathcal{U}$ tem-se que:

$$u_i(\psi_i^0) = u_j(\psi_j^0) = u^\perp \text{ e } u_i(\psi_i^*) = u_j(\psi_j^*) = u^\top.$$

Se uma distribuição de probabilidades sobre o espaço de funções utilidade possíveis é conhecida, então não é necessário ser tão conservador, escolhendo o pior caso de função utilidade. Se o usuário não for consistente, potencialmente todas as funções utilidade candidatas $\mathcal{U}$ podem apresentar $\Pr(u) > 0$, implicando que, mesmo após várias questões, a decisão ótima segundo o critério de arrependimento não se alteraria. Nesse caso, é melhor utilizar a medida de arrependimento, mas calculando a esperança de $\mathrm{Regret}(d, u)$ segundo a distribuição em $\Pr(u)$, isto é,

$$d^* = \arg\min_{d \in \mathcal{D}} \mathrm{E}_{u \sim \Pr(u)}[\mathrm{Regret}(d, u)] = \arg\min_{d \in \mathcal{D}} \sum_{u \in \mathcal{U}} \mathrm{Regret}(d, u) \Pr(u). \tag{2.7}$$

A avaliação de uma decisão d segundo a equação (2.7) pode ser considerada a função valor esperada V_{E}^d, isto é, a esperança dos valores esperados de d baseada na distribuição $\Pr(u)$ e nas funções $u(\cdot)$ (BOUTILIER, 2003). Tem-se então:

$$V_{\mathrm{E}}^d = \sum_{u \in \mathcal{U}} \Pr(u) V_u^d, \tag{2.8}$$

pois,

$$\begin{aligned} d^* &= \arg\min_{d \in \mathcal{D}} \textstyle\sum_{u \in \mathcal{U}} \mathrm{Regret}(d, u) \Pr(u) \\ &= \arg\min_{d \in \mathcal{D}} \textstyle\sum_{u \in \mathcal{U}} [V_u^{d_u^*} - V_u^d] \Pr(u) \\ &= \arg\min_{d \in \mathcal{D}} [\textstyle\sum_{u \in \mathcal{U}} V_u^{d_u^*} \Pr(u) - \sum_{u \in \mathcal{U}} V_u^d \Pr(u)] \\ &= \arg\max_{d \in \mathcal{D}} \textstyle\sum_{u \in \mathcal{U}} V_u^d \Pr(u) \\ &= \arg\max_{d \in \mathcal{D}} V_{\mathrm{E}}^d. \end{aligned} \tag{2.9}$$

2.4.5 Escolhendo Questões

Uma questão pode ser vista sob dois aspectos: 1) o quanto ela melhora a estimativa da função utilidade; e 2) o quanto ela melhora a decisão ótima a ser tomada. No primeiro caso, pode-se calcular a entropia H sobre as probabilidades $\Pr(u)$, isto é, $H = -\Pr(u) \log \Pr(u)$, e, quanto menor for essa entropia, maior acurácia terá a estimativa da função utilidade. No segundo caso, pode-se calcular o arrependimento esperado de uma decisão (equação (2.7)) e, quanto menor esse valor, mais seguro pode-se estar de que a decisão tomada é adequada.

No problema de EP, o que se deseja é conhecer as preferências do usuário de forma adequada a garantir a decisão ótima. O conhecimento das preferências de forma completa garante que a decisão tomada será a melhor possível. Dessa forma, pode-se buscar a acurácia nas probabilidades $\Pr(u)$ como uma forma de obter a decisão ótima. Por outro lado, pode haver casos que, mesmo com baixa acurácia, uma decisão ótima pode ser tomada.

Ao escolher uma questão, pode-se adotar um dos dois aspectos mencionados anteriormente. Uma vez que não é conhecida a resposta que o usuário dará a uma questão, pode-se estimar tal resposta com base no modelo de resposta do usuário $\Pr(r|q,u)$ e na distribuição de probabilidades de aderência $\Pr(u)$ das funções utilidade candidatas, ou seja, $\Pr(r|q) = \sum_{u\in\mathcal{U}} \Pr(r|q,u)\Pr(u)$.

Ao considerar a distribuição de probabilidades $\Pr(r|q)$, uma análise sob o primeiro aspecto da questão, isto é, o quanto ela melhora a estimativa da função utilidade, resultará que qualquer questão sempre produzirá uma melhor estimativa da função utilidade, ou, no pior caso, mantém-se a mesma qualidade de estimativa. No entanto, isso pode não ocorrer após obter as respostas, pois existe a possibilidade de ocorrência de respostas que são contrárias à resposta com maior probabilidade segundo a distribuição $\Pr(r|q)$ estimada até então, diminuindo a acurácia.

Ao considerar o segundo aspecto, o problema é ainda maior, pois pode ocorrer de não haver nenhuma questão que aumente a *esperança* da utilidade esperada, já que elas podem eliminar funções utilidade, tal que seria fácil obter bons valores de utilidade, mesmo havendo um comprometimento com as funções utilidade restantes. Nesse caso, seria mais interessante não fazer pergunta nenhuma, o que seria uma decisão errada, quando o arrependimento não está próximo de 0.

Se o aspecto de arrependimento esperado for utilizado, considere uma questão q e a estimativa da probabilidade de obter a resposta $r \in \mathcal{R}_q$ dada por $\Pr(r|q) = \sum_{u\in\mathcal{U}} \Pr(r|q,u)\Pr(u)$. Pode-se então definir a informação de uma questão baseado na redução média que as respostas causam no arrependimento esperado, ou seja, define-se $\mathrm{Info}(q)$ como:

$$\begin{aligned}
\mathrm{Info}(q) &= \min_{d\in\mathcal{D}} \sum_{u\in\mathcal{U}} \mathrm{Regret}(d,u)\Pr(u) \\
&\quad - \sum_{r\in\mathcal{R}_q} \Pr(r|q) \min_{d\in\mathcal{D}} \sum_{u\in\mathcal{U}} \mathrm{Regret}(d,u)\Pr(u|q,r) \\
&= \min_{d\in\mathcal{D}} \sum_{u\in\mathcal{U}} \mathrm{Regret}(d,u)\Pr(u) \\
&\quad - \sum_{r\in\mathcal{R}_q} \min_{d\in\mathcal{D}} \sum_{u\in\mathcal{U}} \mathrm{Regret}(d,u)\Pr(r|q,u)\Pr(u).
\end{aligned}$$

Dessa forma, pode-se escolher a questão ótima

$$q^* = \arg\max_{q \in \mathcal{Q}} \mathrm{Info}(q) = \arg\min_{q \in \mathcal{Q}} \sum_{r \in \mathcal{R}_q} \min_{d \in \mathcal{D}} \sum_{u \in \mathcal{U}} \mathrm{Regret}(d, u) \Pr(r|q, u) \Pr(u). \quad (2.10)$$

A definição da escolha ótima é apenas com relação à suposição de que há apenas uma questão por fazer. Se forem consideradas questões sucessivas, deve-se considerar o efeito de questões subseqüentes para escolher qual a melhor questão a fazer no momento. Ainda nesse contexto, pode-se pensar em diferentes situações: um número fixo de questões realizadas para então tomar uma decisão; realizar perguntas até se obter um valor mínimo de arrependimento; etc. Boutilier (2002) utiliza a noção de um custo associado a cada questão realizada e um custo associado à tomada de decisão sob incertezas. O agente deve então diminuir o custo total envolvido no processo de EP, isto é, o custo acumulado com as questões e o custo associado a tomada de decisão (DOSHI; ROY, 2008).

Escolher por quaisquer uma dessas opções configura um problema de meta-EP, uma vez que se deve considerar também as preferências do usuário com relação à acurácia com que o processo de EP será programado no agente e o que deve ser alcançado durante o processo de EP. Ainda, muitas vezes esse modelo teórico é abandonado e estratégias *ad hoc* são implementadas que alcançam um melhor desempenho em cenários específicos, muitas vezes considerando um usuário consistente (BUFFETT; FLEMING, 2007; BOUTILIER et al., 2006; PATRASCU et al., 2005)

2.5 Considerações Finais

Neste capítulo foram apresentados alguns conceitos importantes para determinar o significado da delegação, tomando como perspectiva os desejos ou objetivos do usuário e as tomadas de decisões do agente. Essas tomadas de decisões foram relacionadas, sob princípios de racionalidade, às preferências do usuário, e mostrou-se como tais princípios podem ser mapeados em teorias que formalizam como decisões devem ser tomadas.

Com o conhecimento desse arcabouço, demonstrou-se como a delegação pode ser realizada sem que o usuário sintetize seus objetivos em uma linguagem do agente, mas sim, o agente extraindo estes objetivos em um processo de questões e respostas que pode ser aplicado a usuários totalmente leigos na noção de funções utilidade.

A EP torna-se um problema fundamental quando se quer tomar decisões no lugar de um usuário. Embora o modelo normativo para se tomar decisões, a TUE, tenha

sido estabelecido e venha sendo estudado há cinco décadas, só recentemente modelos formais para o processo de EP têm sido consolidados.

No entanto, além do modelo formal para EP, deve-se também especificar com maior formalidade a interação do agente com o usuário. Tradicionalmente se usa questões hipotéticas como meio para obter informações sobre as preferências do usuário. Em geral, problemas de ordem semântica são desconsiderados, ou seja, se o usuário interpreta as questões da mesma forma que elas são implicadas no processo de EP. Isto parte do pressuposto que o usuário consiga avaliar uma situação apenas com os atributos impostos pelo processo de EP.

Por outro lado, muitos trabalhos apresentam incoerências na TUE como modelo descritivo da tomada de decisões de seres humanos, descrevendo casos nos quais a TUE não se conforma sistematicamente às preferências do usuário. Enquanto alguns trabalhos creditam essa não conformação a uma incoerência da TUE, outros trabalhos a consideram como decorrência de uma incoerência na interação entre agente e usuário no processo de EP. No próximo capítulo serão apresentadas essas duas visões das quais vem a motivação para o trabalho apresentado nesta tese.

3 TOMADAS DE DECISÕES PELO USUÁRIO

A EP pode ser necessária por dois motivos diferentes (CARENINI; POOLE, 2002): 1) para prever o que pessoas "similares" farão em circunstâncias "similares", ou seja, para inferir como as pessoas tomam decisões; e 2) para projetar um modelo de preferências de um usuário específico, para ajudá-lo a tomar uma decisão informada e balanceada que seja consistente com seus objetivos.

Embora em ambos os casos as decisões envolvidas sejam ou devam ser influenciadas pelas preferências do usuário, essas decisões não têm necessariamente que convergir a um único modelo (LUCE; WINTERFELDT, 1994). A TUE, como modelo, claramente serve aos objetivos do segundo motivo, não só por seus axiomas intuitivamente apresentarem princípios de racionalidade, mas também por sua simplicidade, tratabilidade e axiomatização. Embora a TUE apresente características normativas – de como as pessoas deveriam agir –, ela não apresenta necessariamente propriedades descritivas – de como as pessoas agem.

Embora ainda hoje a TUE seja a teoria mais utilizada quando aplicada a sistemas que automatizem a EP, na área de Economia ela vem sendo criticada desde os primeiros anos de sua axiomatização, quando exemplos teóricos foram formulados, os quais, intuitivamente, violavam alguns desses axiomas, até anos mais recentes, quando experimentos foram realizados que comprovam sistematicamente tais violações (STARMER, 2000). Alguns desses experimentos testam a condição de independência implicada pelos axiomas da TUE, como os paradoxos de Allais (CUBITT; STARMER; SUGDEN, 2001; TODOROV; GOREN; TROPE, 2007) que propõem experimentos nos quais a condição de independência não se mantêm quando as loterias utilizadas ao formular questões estão próximas de tornarem-se degeneradas.

Outros experimentos demonstram o efeito de dotes[1]. Nesses experimentos, o usuário é confrontado com as seguintes questões sobre dois objetos A e B: trocar A por B, trocar B por A, ou escolher entre A e B. Embora analiticamente todas as questões colocadas versem essencialmente sobre a preferência entre os objetos A e B, as respostas obtidas não são as mesmas (BRAGA; STARMER, 2005). As

[1]Em inglês, endowment effect.

respostas parecem ser influenciadas por uma aversão à perda, isto é, quando os objetos são equivalentes, ou muito parecidos, existe uma preferência por manter o objeto que já se possui. O efeito de dotes contraria duas suposições implícitas na EP: invariância com relação ao procedimento, isto é, preferências sobre loterias são independentes do método utilizado para extraí-las junto a um indivíduo; e invariância com relação à descrição, isto é, preferências sobre loterias são puramente uma função das distribuições de probabilidades de conseqüências implicada pela loteria e não depende de como tais distribuições são descritas.

Por outro lado, experimentos também foram realizados para demonstrar que essas violações da TUE não implicam que as preferências do usuário não podem ser modelas pela TUE, mas que existe uma dificuldade para o usuário interpretar as questões formuladas e também atender com acurácia suas próprias preferências. Schmidt e Neugebauer (2007) observam que tais violações ocorrem mais quando o indivíduo não é consistente em suas respostas, podendo transparecer algum tipo de estocasticidade. Blavatskyy (2007) também apresenta um modelo onde, ao avaliar a utilidade esperada de uma loteria, um indivíduo possa apresentar erros estocásticos, mas que tal erro nunca faz com que a utilidade esperada esteja abaixo ou acima da pior ou da melhor conseqüência respectivamente envolvidas na loteria, fazendo com que o valor esperado seja superestimado ou subestimado para loterias próximas de tornarem-se degeneradas. Bleichrodt, Pinto e Wakker (2001) propõem uma calibração em métodos de EP para refletir erros sistemáticos do usuário conforme a questão utilizada, interpretando as respostas do usuário à luz de conhecimentos sobre o comportamento de seres humanos.

Neste capítulo serão discutidas algumas diretrizes para que as respostas do usuário reflitam melhor as suas preferências, possibilitando efetivamente a EP, e na sua conclusão, apresenta-se a proposta desta tese que consiste em estabelecer uma forma de interação entre agente e usuário que leve em conta tais diretrizes para diminuir esses erros causados pelo usuário.

3.1 Hipótese de Preferências Descobertas

Plott (1996) oferece a Hipótese de Preferências Descobertas "como uma forma de impor algum entendimento sobre um corpo complexo de teoria e dados gerados por economistas e psicólogos". Ele parte da observação de que os exemplos de paradoxos com relação a uma teoria de racionalidade parecem ser de duas classes: tarefas novas, que consiste em situações nas quais os indivíduos que escolhem têm pouca ou nenhuma experiência prévia com a tarefa de escolha/decisão; e outros indivíduos, que

contempla situações onde o comportamento de outros indivíduos é importante para um indivíduo específico.

3.1.1 Definição da Hipótese de Preferências Descobertas

A Hipótese de Preferências Descobertas propõe que escolhas racionais evoluem através de três estágios refletindo experiência e prática (PLOTT, 1996). No primeiro estágio, quando o usuário não possui experiência com as situações de decisões, as escolhas do usuário refletem alguma heurística utilizada para otimização, pois o usuário exibe atenção e conhecimento limitado sobre o ambiente imediato e os efeitos a longo prazo de suas decisões. Dessa forma, as decisões do usuário apresentam uma componente substancial de aleatoriedade, decorrentes da atenção limitada que o usuário apresenta da situação, mas também apresentam aspectos sistemáticos, decorrentes da estratégia de uso de uma heurística de otimização, mesmo que tal estratégia não faça sentido quando analisada da perspectiva de um modelo racional baseado em preferências.

O segundo estágio é obtido quando prática sob incentivo é obtida com repetição, pois provém experiências de decepção e fazem o indivíduo focar-se mais na tarefa. Neste estágio, as escolhas do usuário refletem uma atenção maior ao ambiente e aos efeitos de suas decisões no mesmo, reduzindo a aleatoriedade nas decisões do usuário e apresentando uma conformidade com o modelo racional de tomada de decisão.

No terceiro e último estágio, "escolhas começam a antecipar racionalidade refletida nas escolhas de outros indivíduos. O fato de que outros podem estar agindo racionalmente, e as conseqüências da racionalidade, como ela funciona através da interdependência das instituições sociais, se torna refletida nas escolhas de cada agente."

Então, sob a hipótese de preferências descobertas, violações da TUE seriam erros nas preferências emitidas que podem desaparecer em ambientes que proporcionam certos tipos de aprendizado. A hipótese de preferências descobertas parece isolar a TUE de evidências experimentais não normativas (CUBITT; STARMER; SUGDEN, 2001) e, de um ponto de vista metodológico, parece ser a mais forte defesa da TUE contra evidências experimentais, propondo que cada indivíduo possui preferências coerentes, mas essas preferências não necessariamente são reveladas em decisões. Quando confrontado com uma tarefa de decisão particular, seja dentro ou fora do laboratório, o usuário pode não saber qual das ações oferecidas a ele melhor satisfaz suas preferências. Isso é algo que precisa ser descoberto e tal descoberta pode envolver processos de obtenção de informações, deliberação e aprendizado através de tentativa e erro. Somente quando esse processo estiver completo é que o comportamento do indivíduo revela suas verdadeiras (ou subjacentes) preferências.

Uma característica crucial da hipótese é que preferências subjacentes são independentes do processo particular de descoberta (invariância de procedimento), qualquer processo que provenha oportunidades e incentivo suficientes para deliberação e aprendizado obterão as mesmas preferências subjacentes. Dessa forma, a hipótese de preferências descobertas é diferente da hipótese em que as preferências são construídas no processo de tomar a decisão, e que processos diferentes podem induzir a construção de preferências diferentes (BETTMAN; LUCE; PAYNE, 1998).

3.1.2 Erros ao tomar Decisões

A hipótese de preferências descobertas implica que, para um dado indivíduo e uma dada tarefa experimental, há uma resposta "correta", a resposta que é consistente com as preferências subjacentes. Falhas de dar esta resposta é uma forma de erro, e alguns tipos de erro são (CUBITT; STARMER; SUGDEN, 2001):

1 equívoco sobre o procedimento dos experimentos, isto é, o indivíduo falha ao entender a tarefa que lhe foi colocada. O indivíduo pode não querer revelar que não entendeu, ou pensar que entendeu corretamente a tarefa.

2 inferências lógicas inválidas, isto é, o indivíduo pode entender o procedimento do experimento, mas realizar inferências incorretas sobre como melhor satisfazer suas preferências nesse procedimento.

3 desequilíbrio de crenças em experimentos com vários indivíduos quando o que é melhor para um indivíduo pode depender do que um outro decide. Indivíduos estão em equilíbrio se, quando todos indivíduos agem conforme suas crenças, suas decisões conjuntas confirmam suas crenças.

4 falsas expectativas sobre efeitos, isto é, quando ao escolher entre alternativas de ações, um sujeito pode saber exatamente qual conseqüência objetiva seguirá de cada ação, mas falhar ao prever alguma qualidade afetiva da experiência subjetiva correspondente a aquelas conseqüências, a relação de conseqüência-afeto.

O verdadeiro problema então é o de controle do experimento. Em qualquer pesquisa sobre as características das preferências de seres humanos, se os resultados experimentais são para serem interpretados sob a suposição que um certo tipo de erro não ocorre, então um projeto ideal deve minimizar a possibilidade desse tipo de erro. Alguns critérios para reduzir tais erros são (CUBITT; STARMER; SUGDEN, 2001):

1 transparência e simplicidade: tarefas experimentais devem ser tão simples e transparentes quanto possível, para limitar a quantidade de deliberação e aprendizado que é requerido pelo indivíduo.

2 incentivo: uma vez que o desempenho do indivíduo em uma tarefa é melhorado quando maior esforço mental é despendido em deliberações e aprendizado, indivíduos devem perceber que tal esforço é adequadamente recompensado.

3 oportunidades de aprendizado: indivíduos devem ter oportunidades para aprendizado por tentativa e erro sobre a tarefa experimental e sobre conseqüências de estratégias alternativas que eles podem adotar.

Pode-se dizer que uma tarefa experimental é transparente quando a natureza da tarefa, e o que é esperado do indivíduo na sua execução, são facilmente entendidos pelo indivíduo. Já uma tarefa é simples quando, para um indivíduo que entende o que é dele esperado, sua realização exige pouco esforço cognitivo.

Os mecanismos de extração com base em escolhas são simples e transparentes e, portanto, requerem menos aprendizado do que outras alternativas de EP. Na verdade, dado que tudo que um indivíduo tem que fazer em uma tarefa de escolha é escolher qual de duas opções são preferidas, há muito pouco, se é que há algo, para o indivíduo aprender sobre o procedimento. No entanto, as opções envolvidas podem não ser distinguidas de forma simples, pois, dependendo das opções envolvidas, o usuário pode necessitar de conhecimento a respeito de seus efeitos lógicos no ambiente e de seus efeitos afetivos nele mesmo.

Uma forma pela qual o erro pode ser reduzido é através da disposição do indivíduo para despender esforço cognitivo adequado para executar uma tarefa com acurácia. Indivíduos podem dar mais ou menos atenção às instruções, eles podem deliberar mais ou menos sobre as opções abertas para ele, eles podem ser mais ou menos cuidadosos ao fazer inferências lógicas, e, quando existem oportunidades para aprendizado por tentativa e erro, eles podem dar mais ou menos atenção para tais oportunidades. A principal função do uso de incentivos é aumentar os benefícios do esforço cognitivo demandado, em relação aos seus custos. Os incentivos são cruciais na hipótese de preferências descobertas, pois têm um importante papel nos mecanismos de aprendizado de um indivíduo.

As oportunidades de aprendizado dependem de repetições e do retorno que o indivíduo tem de suas escolhas. Repetição e retorno são importantes por: possibilitarem o alcance de equilíbrio em experimentos interativos, serem um mecanismo para instruir os indivíduos sobre o procedimento experimental e eliminar equívocos, e

permitirem que indivíduos experimentem diferentes respostas e utilizem aprendizado por tentativa e erro como suplemento a inferências lógicas na descoberta do melhor modo de satisfazer suas próprias preferências em uma dada tarefa, inclusive utilizando experiências afetivas para isso.

3.1.3 Experimento Ideal

Plott (1996) recomenda que, para extrair preferências mais confiáveis, deve-se considerar tarefas repetitivas com incentivo e retorno. No entanto, retorno e tarefas repetitivas podem ser muito custosos, seja pela disponibilidade do ambiente, seja pela disponibilidade do indivíduo, ou ainda, pela geração e oferta de incentivos propriamente dito. O problema de custo não é o único, existe também o de factibilidade para repetição das tarefas. Algumas decisões são tomadas apenas uma vez ao longo da vida de um indivíduo – por exemplo, realizar ou não uma operação com risco de vida – e prover o indivíduo com várias oportunidades de aprendizado torna-se impossível.

Em alguns casos, simplesmente não há como extrair dados de preferências confiáveis quando a observação direta da experiência pelo indivíduo não é possível. Nos casos onde as recomendações da hipótese de preferências descobertas não são possíveis, deve-se recorrer a explicações para as possíveis anomalias, explicitando como elas funcionam sistematicamente. Nesse sentido, é importante saber se o usuário está usando alguma estratégia que funcionaria como viés para as decisões tomadas. Dessa forma, seria necessário primeiro determinar qual é a estratégia utilizada e na posse desse conhecimento, calibrar avaliações obtidas em um único experimento para opções que não podem, por sua natureza, ser avaliadas em experimentos repetidos.

3.2 Proposta da Tese

Na área de Economia, a importância da EP é principalmente para prever como pessoas tomam decisões, uma vez que é necessário conhecer como as pessoas se comportam diante de problemas de decisões de modo a construir teorias sobre a reação de uma população junto a uma política social implementada. No caso desta tese, o mais importante é projetar um modelo de preferências para auxiliar uma pessoa em sua tomada de decisão. No caso extremo, uma vez conhecidas as preferências da pessoa e como atuar para satisfazê-las, o próprio agente pode tomar decisões no lugar dessa pessoa, nesse caso, usuário. Porém, ao interpretar as interações entre agente e usuário, é importante conhecer qual informação pode-se efetivamente extrair das respostas

do usuário com relação às suas preferências. A interpretação dessas interações tem sido estudada principalmente na área de Economia e Psicologia.

A área de Economia, ao considerar os experimentos que demonstram violações da TUE nas decisões de seres humanos, desenvolveu outras teorias que possibilitam explicar algumas dessas violações (STARMER, 2000). No entanto, devido ao fato de que nenhuma dessas teorias se sobressaiu como uma substituta definitiva para TUE, a TUE continua sendo a teoria mais utilizada. Mesmo que a aderência da TUE como teoria descritiva tenha sido abalada, como teoria normativa a TUE ainda continua imperando. O que é inegável é que as violações da TUE e algumas explicações para as mesmas fizeram com que uma atenção maior no controle dos experimentos se tornasse necessária(HARRISON; HARSTAD; RUTSTROM, 2004).

Já na área de automatização de tomadas de decisões, a TUE também continua como modelo para as preferências do usuário, mas nenhuma atenção se volta para a condução do processo de EP. Nesse caso, considera-se a máxima de que as preferências do usuário são invariantes com relação ao procedimento e com relação ao contexto, e que qualquer violação da TUE junto aos dados obtidos pode ser explicada por algum modelo estocástico.

3.2.1 Comportamentos Observados

Nesta tese, propõe-se um procedimento para obter informação junto ao usuário, de forma que não ocorram violações da TUE e que torne efetivamente factível a tarefa do agente de obter as preferências do usuário. O procedimento proposto consiste em considerar que o usuário observa uma situação real de decisão e os efeitos de uma determinada decisão no ambiente. Mas, nessa situação real de decisão, a decisão é tomada pelo agente. Dessa forma, o usuário pode observar o resultado das decisões do agente no ambiente, e interpretá-lo do seu próprio ponto de vista.

Na literatura de EP aplicada a sistemas computacionais, usualmente se considera o uso de questões formuladas com conseqüências hipotéticas. Nesta proposta, as conseqüências hipotéticas dão lugar a conseqüências reais, ou seja, os comportamentos demonstrados pelo agente. Ao responder uma questão, o usuário não precisa quantificar de forma consciente os comportamentos, mas mesmo assim pode analisá-los e avaliá-los utilizando conhecimento tácito (IYENGAR; LEE; CAMPBELL, 2001; HUI; BOUTILIER, 2008). Além disso, o único tipo de questão considerado é a comparação entre pares de comportamentos. A EP realizada com esse procedimento denomina-se aqui de EPCO.

3.2.2 Justificativas

Este procedimento apresenta algumas propriedades interessantes para evitar violações da TUE se a hipótese de preferências descobertas for considerada verdadeira. A primeira propriedade é que este procedimento não requisita do usuário a noção de loterias. Ao analisar uma loteria o usuário deve possuir um significado para as probabilidades envolvidas; porém, tal significado é tão subjetivo que, em alguns trabalhos, é apontado como explicação para as violações observadas da TUE (BLAVATSKYY, 2007; TODOROV; GOREN; TROPE, 2007). Então, se loterias forem eliminadas do procedimento de EP, em geral, todas as teorias convencionais seriam equivalentes, já que nelas considera-se uma função utilidade para comportamentos assegurados e varia-se apenas a forma como tal função é combinada mediante uma loteria (STARMER, 2000).

A segunda propriedade é que este procedimento não pressupõe uma linguagem comum entre agente e usuário. O usuário avalia comportamentos com base em sua própria observação, já contando inclusive com o resultado afetivo que tal comportamento possa produzir em si próprio. Nesse caso, o usuário não precisa quantificar o comportamento em uma representação abstrata, nem mesmo realizar inferências lógicas ou afetivas sobre conseqüências hipotéticas, possibilitando ao mesmo tempo transparência na questão formulada ao usuário e simplicidade ao usuário para emitir a resposta à questão.

Por último, o fato de que os comportamentos avaliados são demonstrados, permite que a EP seja feita durante a execução das decisões envolvidas, isto é, quando o agente já deve agir atendendo as preferências do usuário. Dessa forma, o usuário pode perceber de forma direta o resultado de suas avaliações sobre os comportamentos exibidos pelo agente. Nesse caso, o agente deve ter um compromisso entre explotação, que consiste em atuar da melhor forma possível baseado nas atuais crenças do agente, e exploração, que consiste em atuar também de forma não ótima para poder obter mais informações e revisar suas crenças.

Essas propriedades podem ajudar a minimizar ou até diminuir os tipos de erros apontados por Cubitt, Starmer e Sugden (2001). O procedimento de obter avaliações sobre comportamentos observados evita que ocorram equívocos na interpretação das questões e também a necessidade de inferências, tanto pela sua simplicidade, como pela sua transparência. A estrutura da questão é a mais transparente possível: "observe dois comportamentos e escolha o melhor". Nesse caso, não há necessidade de o usuário mudar a perspectiva sobre a qual ele está acostumado acompanhar e avaliar comportamentos, ele não precisa lançar mão de atributos para descrever um comportamento. Além da comparação entre pares de comportamentos ser conside-

rada um tipo de questão muito simples, pois tudo que se exige é a comparação relativa entre apenas dois comportamentos, o fato de ser sobre comportamentos observados e não sobre opções de decisões (loterias) evita a necessidade de inferências lógicas, pois tudo que é pra ser julgado já ocorreu e é de conhecimento do usuário, não sendo necessário realizar nenhuma previsão com relação ao futuro. Além disso, o usuário pode fazer uso de conhecimentos tácitos, pois sua decisão não precisa ser explicada a ninguém, nem mesmo a ele, não havendo a necessidade de quantificações de comportamentos. Tal percepção também ajuda a evitar falsas expectativas com relação ao afeto que uma ação produz sobre o usuário, pois tal afeto já terá ocorrido durante a observação dos comportamentos, eliminando qualquer expectativa e necessidade de aprendizado. O mesmo se pode dizer com relação às crenças sobre conhecimentos e preferências de outros usuários.

3.2.3 Problemas

O cenário de avaliação sobre comportamentos observados, apesar de ser mais realista no que se refere a avaliação de uma situação, pode trazer problemas indesejáveis, os quais esta tese pretende analisar. O primeiro deles consiste no fato do agente e do usuário possuírem diferentes observações das situações analisadas. Pode ocorrer que o usuário considere um atributo que o agente não considera, impossibilitando que o agente avalie e tome decisões em total consonância com o usuário. Por outro lado, também pode ocorrer que o agente possua acesso a atributos que o usuário não possui, mas que este gostaria de considerar na decisão que o agente tome por ele, nesse caso perde-se a capacidade que o agente tem de tomar decisões realmente ótimas.

Outro problema consiste na impossibilidade de usar de forma direta os algoritmos utilizados tradicionalmente na EP. Isso ocorre devido a dois motivos. Primeiro, o fato de apenas situações reais poderem ser analisadas implica que os métodos devem limitar-se a tais situações ao formular uma questão. Segundo, realizar uma questão consiste em demonstrar dois comportamentos, mas, se o ambiente onde este comportamento é exibido for estocástico, uma questão formulada não poderá ser demonstrada ao usuário de forma determinista.

Por último, ainda há o problema da possibilidade de sucessivas experimentações no ambiente. A EPCO só é possível quando a tarefa, que se deseja programar no agente, permita repetição. Isso implica que a tarefa deve ter um início e um fim e que decisões sobre tais tarefas ocorram com freqüência durante a vida de um indivíduo. Além disso, o resultado do comportamento do agente deve produzir o efeito no usuário

antes do mesmo emitir sua avaliação sobre os comportamentos demonstrados.

4 EXTRAÇÃO DE PREFERÊNCIAS COM BASE EM COMPORTAMENTOS OBSERVADOS

O problema de EPCO consiste no processo de EP quando se limita o tipo de interação entre agente e usuário a ocorrer apenas por intermédio de respostas emitidas pelo usuário à seguinte questão: "observe dois comportamentos demonstrados pelo agente no ambiente e responda qual dentre esses dois comportamentos é o melhor?" Na EPCO, as conseqüências usualmente utilizadas em questões no processo de EP dão lugar a comportamentos reais demonstrados no ambiente que, por sua vez, possuem observações subjetivas por parte do usuário e do agente. Dessa forma, o processo de EPCO consiste nos seguintes passos:

- agente formula uma questão;
- agente toma decisões que representam tal questão;
- usuário e agente observam pares de comportamentos $(\psi^1_{\mathrm{Us}}, \psi^2_{\mathrm{Us}})$ e $(\psi^1_{\mathrm{Ag}}, \psi^2_{\mathrm{Ag}})$, respectivamente, como efeito de tais decisões no ambiente;
- usuário responde qual dos dois comportamentos é preferido;
- agente infere, com base na questão formulada, nos comportamentos observados e na resposta do usuário, informações parciais sobre as preferências do usuário; e
- passos anteriores se repetem até que uma condição de término seja atingida.

Um modelo de tal procedimento é exibido na figura 4.1. Comparando com a figura 2.2, destacam-se três principais diferenças. Primeiro, a interação entre agente e usuário é feita por intermédio do ambiente, sendo necessário um módulo específico para demonstrar um comportamento ao usuário. Segundo, o modelo de inferência torna-se mais complexo, pois também devem ser considerados os comportamentos observados pelo agente, mesmo que esses comportamentos estejam relacionados à questão formulada. Terceiro, enquanto no processo de EP a interação entre agente

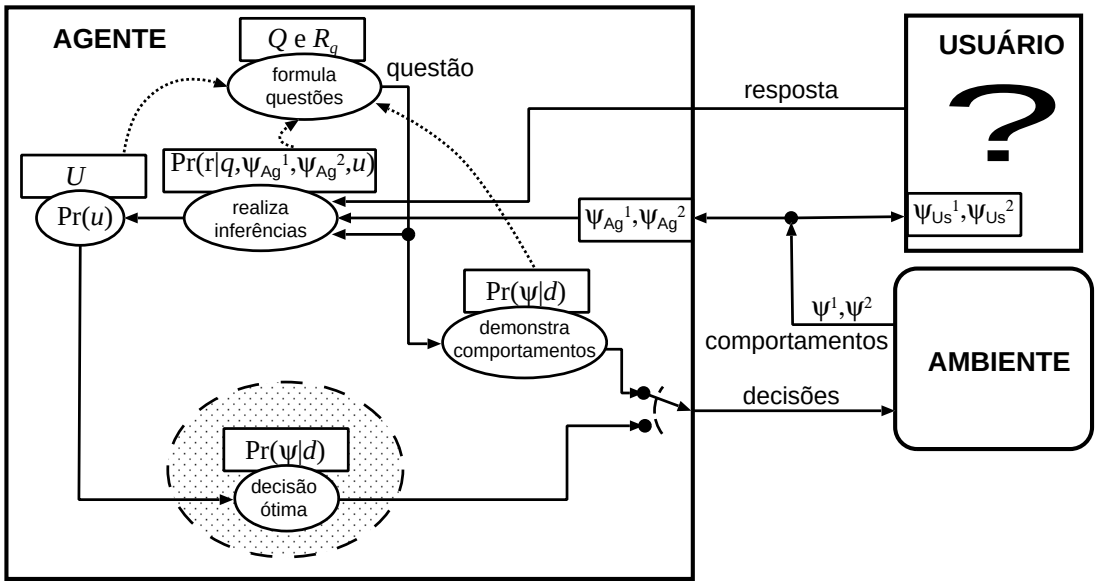

Figura 4.1: Modelo para o EPCO. O agente toma decisões e aplica no ambiente, resultando em dois comportamentos. Em sincronismo, o agente e o usuário observam subjetivamente tais comportamentos. O usuário escolhe qual comportamento é melhor e responde indicando sua escolha ao agente. As relações entre comportamentos observados pelo agente, questão formulada e respostas do usuário são utilizadas para definir probabilidades de aderência para as funções utilidade candidatas. As probabilidades de aderência guiarão a definição da política ótima do agente.

e ambiente só existia após uma função utilidade ser estimada, na EPCO a interação entre agente e ambiente faz parte do processo de EP.

O uso de comportamentos observados no lugar de conseqüências hipotéticas permite tornar o processo de EP mais natural do ponto de vista do usuário, já que ele deverá emitir comparações entre situações, podendo utilizar seu conhecimento tácito para avaliar tais situações, já que não é necessário descrever tais situações de forma abstrata.

Enquanto tal cenário é mais natural ao usuário, ele também apresenta novos desafios para o processo de EP; aqui o problema de equivalência semântica entre as interpretações das questões pelo usuário e pelo agente torna-se um problema de diferentes observações do agente e do usuário. Um segundo problema é a formulação de questões que não deve só considerar os comportamentos possíveis de serem demonstrados, mas também como demonstrá-los. Neste capítulo será apresentada uma formalização para a interação entre agente e ambiente neste novo cenário, a partir da qual serão apresentados os principais problemas originados na EPCO: diferentes observações para agente e usuário, e formulação de questões em ambientes não deterministas.

4.1 Agentes, Ambientes e Desempenho

Um agente, como o próprio nome diz, é projetado para agir, praticando ações em um ambiente. A ação executada pelo agente pode ser considerada sob dois aspectos: 1) o momento em que ela é executada, no qual o ambiente apresenta uma determinada situação; e 2) o efeito que ela causa no ambiente, alterando a situação.

O ambiente aqui considerado apresenta: variáveis, que descrevem a sua situação; e dinâmica, que consiste na relação entre as variáveis e as ações do agente no tempo. A escolha de uma ação por um agente altera diretamente os valores de algumas das variáveis do ambiente. Essa alteração se dá por meio de um atuador, que relaciona as possíveis ações com as variáveis do ambiente.

Até aqui, apenas a relação de causa e efeito entre ações e variáveis do ambiente foi considerada. No entanto, um agente que não tenha percepção do ambiente não pode conhecer o significado de suas ações, isto é, as situações onde uma ação deve ser utilizada. A percepção do agente possui uma relação direta com algumas variáveis do ambiente. Essa relação entre variáveis e percepção se dá por meio de sensores.

Dessa forma, Russell e Norvig (2004) definem um agente como sendo "tudo o que pode ser considerado capaz de perceber seu ambiente por meio de sensores e de agir sobre o ambiente por intermédio de atuadores".

4.1.1 Ambientes Discretos

Embora seja comum considerar variáveis contínuas em ambientes físicos, tanto com relação à situação como com relação ao tempo, a consideração de variáveis discretas muitas vezes facilita o tratamento matemático e, principalmente, o computacional. Mesmo que o agente atue em um ambiente contínuo, a conversão entre variáveis contínuas e variáveis discretas pode ser feita diretamente pelos sensores e atuadores, permitindo que um agente baseado em variáveis discretas possa atuar de forma efetiva em um ambiente contínuo.

A consideração de variáveis discretas do ambiente permite facilmente o mapeamento de cada situação do ambiente a uma variável discreta $s \in \mathcal{S}$, denominada estado do ambiente, ou simplesmente, estado. O conjunto $\mathcal{S}$ possui os estados necessários para representar todas possíveis situações do ambiente. A figura 4.2 mostra um esquema da relação entre agente, tempo e ambiente. Em um tempo t, o agente faz uma observação o_t do estado s_t do ambiente e executa uma ação a_t. O estado s_t descreve o ambiente no tempo t e é tal que os estados futuros $s_{t+1}, s_{t+2}, \ldots$ sejam

independentes dos estados passados $s_0, s_1, \ldots, s_{t-1}$ e ações passadas $a_0, a_1, \ldots, a_{t-1}$, ocorridos antes do tempo t, isto é, apresenta a condição de Markov (ROSS, 1970).

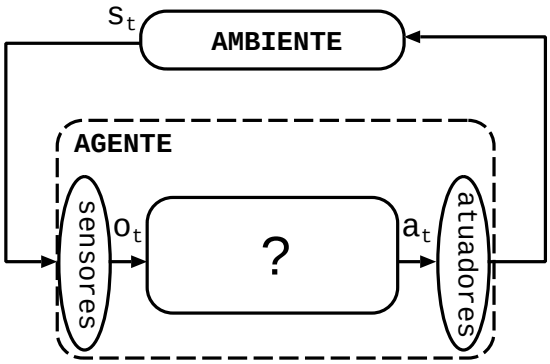

Figura 4.2: Modelo de interação entre ambiente e agente.

Embora seja imposta tal independência entre estados passados e estados futuros de um ambiente, o mesmo não ocorre com as observações. A independência entre as observações passadas e futuras pode não ocorrer por dois motivos: 1) o agente tem um acesso incompleto ao estado do ambiente, isto é, os sensores podem não ter acesso a todas variáveis que caracterizem o estado do ambiente; e 2) o agente tem um acesso ruidoso ao estado do ambiente, isto é, embora haja acesso a todas variáveis, esse acesso não é determinista. Para minimizar este problema, o agente pode considerar observações e ações ocorridas no ambiente antes do tempo t para obter uma melhor estimativa do estado atual; dessa forma, pode-se ter $o_t = f(o_0, o_1, \ldots, o_{t-1}, a_0, a_1, \ldots, a_{t-1}, s_t)$.

Dadas essas formalidades da relação do ambiente com o agente, pode-se determinar algumas características do ambiente relatadas por Russell e Norvig (2004). Um ambiente é determinista se para qualquer estado presente s_t, o estado futuro s_{t+1} é completamente determinado pelo estado s_t e pela ação a_t, isto é, $s_{t+1} = f(s_t, a_t)$. Em caso contrário, o ambiente é estocástico e o próximo estado s_{t+1} torna-se uma variável aleatória com uma função probabilidade de transição $T(s_{t+1} = s' | s_t = s, a_t = a)$ para todo $s' \in \mathcal{S}$, tal que $T(s'|s, a) \geq 0$ e $\sum_{s' \in \mathcal{S}} T(s'|s, a) = 1$. O ambiente determinista pode ser visto como uma especialização do ambiente estocástico onde, dados s e a, existe s' tal que $T(s'|s, a) = 1$ e $T(s''|s, a) = 0$ para todo $s'' \in \mathcal{S}$ e $s'' \neq s'$.

Outro aspecto do ambiente é a observabilidade que o agente possui sobre o estado do ambiente. Para que o agente possa realizar uma tarefa de forma eficaz é importante que o mesmo tenha acesso às variáveis do ambiente que sejam relevantes. Em *stricto sensu*, um ambiente é completamente observável por um agente se $o(s_t) = s_t$, ou seja, a observação possui uma relação de identidade com o estado do ambiente. No entanto, pode-se relaxar essa condição conforme os objetivos do

agente, quando observações diferentes da identidade possam ser tão eficazes quanto a identidade. Além de produzir a mesma eficácia, é importante que a observação apresente propriedades que uma observação identidade apresentaria, por exemplo, a propriedade de Markov, uma vez que tais propriedades podem ser exploradas para resolver o problema de forma ótima. Nesta tese será considerado que as observações do agente possuem a propriedade de Markov.

O horizonte de tempo em que as tarefas são consideradas também é um aspecto importante[1]. Em algumas tarefas, as interações entre agente e ambiente podem ser naturalmente separadas em episódios independentes, onde as decisões tomadas em um episódio não interferem no resultado obtido em outros episódios. Nesse caso, a tarefa é chamada de episódica e o ambiente pode ser considerado como possuidor de estados terminais e absorventes. Estes estados são terminais no sentido de que uma vez que eles são alcançados, o ambiente vai para um estado inicial segundo uma distribuição característica do ambiente e um novo episódio se inicia; e são absorventes no sentido de que a partir de qualquer estado de um episódio, estados terminais serão alcançados em um futuro finito c.p.1. Este é o caso, por exemplo, do agente *Piloto Automático*, pois após conduzir o usuário até o trabalho, mais tarde ele deve levá-lo de volta a casa, e este processo se repete diariamente. Embora o agente possa aprender com a experiência obtida em cada um dos trajetos, as avaliações de cada trajeto são independentes.

4.1.2 Programando Agentes

Um agente atuando em um ambiente deve ser programado, segundo sua arquitetura, para que ele possa atingir os objetivos de um usuário. Um agente reativo é um agente programado com uma função $\pi : \mathcal{S} \rightarrow \mathcal{A}$ que mapeia cada estado $s \in \mathcal{S}$ em uma ação $a \in \mathcal{A}$, ou ainda, uma função $\pi : \mathcal{S} \rightarrow (\mathcal{A} \rightarrow [0, 1])$ que mapeia cada estado $s \in \mathcal{S}$ em uma função probabilidade $\Pr(a|s)$ que indica a probabilidade de a ser executada. Esta função π que determina qual ação a_t será executada dado um estado s_t é chamada de política. No primeiro caso tem-se uma política determinista, enquanto no segundo, uma política estocástica.

Uma forma mais abstrata de programar um agente é através da definição de uma função utilidade, que mapeia um comportamento ψ, aqui definido como uma seqüência de estados e ações $\psi = s_0, a_0, s_1, a_1, \ldots, s_{N-1}, a_{N-1}, s_N \in \Psi$ em um episódio, em um escalar $u(\psi)$, isto é, $u : \Psi \rightarrow \mathbb{R}$. O agente deve encontrar uma

[1] A definição de ambiente episódico colocada por Russell e Norvig (2004) é diferente da definição aqui adotada. Esta nomenclatura foi baseada emSutton e Barto (1998), que diferenciam entre tarefa episódica e tarefa contínua

política ótima π^* que obtenha bons comportamentos (bons valores de utilidade). A política ótima é definida segundo o problema de decisão, onde a decisão a ser tomada é a escolha de uma política.

Considere uma tarefa onde um agente pode escolher entre políticas pertencentes a um conjunto finito de políticas Π. O resultado de aplicar uma política $\pi \in \Pi$ é uma distribuição de probabilidades $\Pr(\psi|\pi)$ que indica a probabilidade do comportamento ψ ser obtido com a política π. Atribui-se um valor V^π para cada política $\pi \in \Pi$ definido por:

$$V^\pi = \sum_{\psi \in \Psi} \Pr(\psi|\pi) u(\psi).$$

Se o agente conhecer *a priori* o conjunto de políticas $\pi \in \Pi$, assim como as respectivas distribuições $\Pr(\psi|\pi)$ e a função utilidade $u(\cdot)$, pode-se escolher a política ótima no sentido de maximizar a utilidade esperada, pelo menos em teoria, ou seja, realizar o planejamento (BONET; GEFFNER, 2001).

4.1.3 Processo Markoviano de Decisão

Um tipo de função utilidade muito comum é a função utilidade baseada em recompensas. As recompensas são avaliações parciais de um comportamento e são associadas aos estados $s \in \mathcal{S}$ e ações $a \in \mathcal{A}$ por intermédio de uma função recompensa $w : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, isto é, ao passar pelo estado s e aplicar a ação a o agente recebe a recompensa $w(s, a)$. Recompensas positivas devem ser almejadas, enquanto recompensas negativas devem ser evitadas. Essa idéia é formalizada em um Processo Markoviano de Decisão (PMD).

PMDs são construídos sob um já estabelecido formalismo matemático, que compensa as condições simplificadas ou aproximações feitas na dinâmica do sistema. Portanto, algumas restrições são adotadas para adequar-se à modelagem por PMD. A principal restrição é que o agente deve possuir observação completa do estado do ambiente.

Um PMD é definido por:

- um conjunto finito de ações possíveis $a \in \mathcal{A}$;
- um conjunto finito de estados do ambiente $s \in \mathcal{S}$;
- um processo markoviano discreto no tempo, modelado por probabilidades de transições $T(s'|s, a)$;
- uma função recompensa limitada $w(s, a) \in \mathbb{R}$; e

- a dinâmica do sistema é tal que, se o processo está no estado s_t no tempo t e a ação a_t é escolhida, então o próximo estado $s_{t+1} = s'$ ocorre de acordo com a probabilidade de transição $T(s'|s_t, a_t)$ e uma recompensa com valor $w(s_t, a_t)$ é recebida.

O agente deve encontrar uma política de ações $a_t = \pi^*(s_t)$ que maximize uma função valor V^π que representa as recompensas recebidas, isto é, $V^{\pi^*} = \max_{\pi \in \Pi} V^\pi$. Uma opção para a função valor é considerar simplesmente a soma de recompensas recebidas:

$$V^\pi = \mathrm{E}_{s_0 \sim P_0}[\sum_{t=0}^{N-1} w(s_t, a_t)|a_t = \pi(s_t)], \tag{4.1}$$

que define a função utilidade:

$$u(\psi) = \sum_{t=0}^{N-1} w_t^\psi,$$

onde $w_t^\psi = w(s_t^\psi, a_t^\psi)$, s_t^ψ é o estado do comportamento ψ no tempo t, a_t^ψ é a ação executada pelo agente no tempo t e P_0 é uma distribuição de probabilidades para o estado inicial.

Ao utilizar a função valor na equação (4.1), a função recompensa $w(s, a)$ considerada em um PMD é suficiente para representar funções recompensas mais complexas. Por exemplo, se no tempo t a recompensa w_t recebida é uma variável aleatória e limitada por $[W_{min}, W_{max}]$ definida pela função densidade de probabilidade $f(w_t|s_t, a_t, s_{t+1})$, basta definir a função recompensa determinista $w(s, a)$ como:

$$w(s, a) = \sum_{s' \in \mathcal{S}} \int_{W_{min}}^{W_{max}} x f(x|s, a, s')dx.$$

As políticas ótimas e suas funções valores serão iguais para a função recompensa estocástica e para a função recompensa determinista.

Quando se considera um horizonte finito N para a demonstração de um comportamento, deve-se levar em conta o tempo t, definindo-se uma política $\pi^*(s, t)$ para cada $t < N$, isto é, $\pi : \mathcal{S} \times \mathbb{N} \to \mathcal{A}$ ou $\pi : \mathcal{S} \times \mathbb{N} \to (\mathcal{A} \to [0, 1])$.

4.1.4 Atributos, PMD e Funções Aditivas Lineares

Embora até o momento a função recompensa tenha sido associada a uma função definida em $\mathcal{S} \times \mathcal{A}$, nem sempre é necessário explicitar a relação das recompensas com os estados, isto é, pode-se descrever a função recompensa de forma mais compacta.

Considere um conjunto de atributos Ξ com elementos $1, 2, \ldots, k$, e um mapea-

mento $\boldsymbol{\phi}$ de cada par (s, a) para um vetor de k valores de atributos observados, isto é, $\boldsymbol{\phi} : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}^k$ (NG; RUSSELL, 2000). Então, qualquer comportamento ψ com $N(\psi)$ estados pode ser transformado em um vetor de atributos $\boldsymbol{\mu}(\psi) = (\mu_i(\psi))_k$, onde $\mu_i(\psi) = \sum_{t=0}^{N(\psi)-1} \phi_i(s_t^\psi, a_t^\psi)$.

Pode-se considerar recompensas associadas linearmente aos valores de cada atributo, isto é, para cada atributo i existe uma recompensa correspondente w_i, gerando o vetor $\mathbf{w} = (w_i)_k$. Se for considerada a função utilidade $u(\psi) = \langle \mathbf{w}, \boldsymbol{\mu}(\psi) \rangle$, essa representação permite minimizar a quantidade de parâmetros a serem definidos para descrever a função recompensa.

A função $\boldsymbol{\phi}(\cdot)$ não precisa necessariamente assumir uma definição em forma de tabela, pois ela pode ter relação direta com variáveis lidas diretamente pelos sensores, sendo a única restrição necessária aquela relacionada com a condição de Markov, isto é, os sensores devem produzir observações dos atributos, mesmo que não deterministas, que tenham uma esperança dependente apenas do estado e ação atuais.

Agora, pode-se associar uma função utilidade baseada em recompensas a funções aditivas lineares. Considere o seguinte vetor de atributos esperados $\bar{\boldsymbol{\mu}}^\pi$ de uma política π:

$$\bar{\boldsymbol{\mu}}^\pi = \sum_{\psi \in \Psi} \Pr(\psi|\pi) \boldsymbol{\mu}(\psi).$$

Então, se a constante de normalização K na equação (2.6) não for considerada, uma vez que ela atribui um valor constante para toda política, pode-se escrever o valor de uma política π como o produto interno entre o vetor de atributos esperados e o vetor recompensas $\mathbf{w} = (w_i)_n$ do seguinte modo:

$$V^\pi = \sum_{\psi \in \Psi} \Pr(\psi|\pi) u(\psi) = \sum_{i=1}^{k} w_i \sum_{\psi \in \Psi} \Pr(\psi|\pi) \mu_i(\psi) = \sum_{i=1}^{k} w_i \bar{\boldsymbol{\mu}}_i^\pi = \langle \mathbf{w}, \bar{\boldsymbol{\mu}}^\pi \rangle.$$

Note que a função recompensa aqui apresentada, por ser uma função markoviana, só pode ser utilizada para representar as preferências do usuário, quando tais preferências podem ser modeladas em funções lineares. Se não for o caso, recompensas mais complexas devem ser utilizadas, por exemplo, recompensas não markovianas (THIEBAUX et al., 2006)

4.2 O problema de EPCO utilizando o Arcabouço de PMD

No início deste capítulo foi apresentado o problema de EPCO, no qual comportamentos são vistos como conseqüências reais observadas pelo agente. Nesta seção será apresentada uma formalização para a interação entre agente, usuário e ambiente utilizando o arcabouço de PMD.

4.2.1 Comportamentos e Estados do Ambiente

Embora apenas os comportamentos observados pelo agente e usuário interfiram no processo de EPCO, é útil também definir um comportamento completo ocorrido no ambiente, sob o qual os primeiros podem ser definidos. Então, define-se um comportamento completo ψ^i como a seqüência de estados ocorridos e ações executadas em um período $\Gamma_i = 0, 1, 2, 3, \ldots, N$, isto é, $\psi^i = s_0 a_0 s_1 a_1 s_2 a_2 \ldots s_{N-1} a_{N-1} s_N$.

A partir de ψ^i pode-se definir os comportamentos observados pelo agente e pelo usuário no período Γ_i. Uma conseqüência usualmente é descrita por meio de vetores de atributos, pode-se então especificar o mesmo tipo de codificação para os comportamentos observados. Então, deve-se definir vetores de atributos $\boldsymbol{\mu}_{\mathrm{Ag}}(\psi^i)$ e $\boldsymbol{\mu}_{\mathrm{Us}}(\psi^i)$ para o agente e o usuário, respectivamente. Consideram-se funções de observações de atributos $\boldsymbol{\phi}_{\mathrm{Ag}}(s, a)$ e $\boldsymbol{\phi}_{\mathrm{Us}}(s)$ para o agente e usuário, respectivamente. Note que a observação de atributos realizada pelo usuário não contempla as ações executadas pelo agente, mas apenas os efeitos dessas ações no estado do ambiente são considerados. Logo, tem-se que $\boldsymbol{\mu}_{\mathrm{Ag}}(\psi^i) = \sum_{i=0}^{N} \boldsymbol{\phi}_{\mathrm{Ag}}(s_i, a_i)$ e $\boldsymbol{\mu}_{\mathrm{Us}}(\psi^i) = \sum_{i=0}^{N} \boldsymbol{\phi}_{\mathrm{Us}}(s_i)$.

Outro aspecto importante que deve ser formalizado é o espaço de decisão $\mathcal{D}$. Em um PMD, a decisão a ser tomada é a escolha de uma política $\pi \in \Pi$ e pode ser interpretada como várias decisões seqüenciais, isto é, a escolha de ação em cada tempo discreto. Ao considerar um problema de decisão seqüencial, o número de políticas possíveis é exponencial na duração dos períodos demonstrados e nas cardinalidades dos conjuntos de ações e observações do agente. Além disso, se políticas estocásticas são consideradas, o espaço de decisões torna-se contínuo.

4.2.2 Comparação entre Comportamentos

Apesar de uma formalização do comportamento do usuário ser apresentada, o usuário não precisa ter conhecimento sobre tal formalização. O usuário deve apenas comparar dois comportamentos e responder qual deles é melhor. No entanto, essa formalização

é útil ao realizar inferências sobre preferências do usuário. Embora nenhuma abstração em forma de atributos ou quantificações seja necessária, algumas convenções semânticas devem ser feitas para que suas respostas possam ter significados para o agente.

Primeiro, agente e usuário devem possuir a mesma semântica sobre o significado da resposta emitida pelo usuário, que é a comparação entre comportamentos. Após observar dois períodos Γ_i e Γ_j resultando nos comportamentos $\boldsymbol{\mu}^i_{\mathrm{Us}}$ e $\boldsymbol{\mu}^j_{\mathrm{Us}}$ respectivamente, o usuário emite uma resposta entre duas opções: `melhor` ou `pior`. Se a resposta é `melhor`, interpreta-se $u_{\mathrm{Us}}(\boldsymbol{\mu}^i_{\mathrm{Us}}) > u_{\mathrm{Us}}(\boldsymbol{\mu}^j_{\mathrm{Us}})$. Se a resposta é `pior`, interpreta-se $u_{\mathrm{Us}}(\boldsymbol{\mu}^i_{\mathrm{Us}}) < u_{\mathrm{Us}}(\boldsymbol{\mu}^j_{\mathrm{Us}})$.

Segundo, agente e usuário devem possuir um sincronismo ao interpretar as observações realizadas em um mesmo tempo t como fazendo parte de um mesmo período. Se o início e o final de um comportamento não forem evidentes pela tarefa a ser executada, deve-se estabelecer uma comunicação entre agente e usuário para que a observação de ambos dependam do mesmo comportamento real.

4.2.3 Funções Utilidade Candidatas, Questões e Inferências

Em geral, os algoritmos de EP consideram conhecidos *a priori* um conjunto de funções utilidade candidatas, um conjunto de questões possíveis de serem formuladas e uma regra de inferência entre as respostas do usuário e suas preferências (BOUTILIER, 2002; BOUTILIER et al., 2005; CHAJEWSKA; KOLLER; PARR, 2000).

Ao considerar PMDs, usuários com preferências neutras ao risco e com independência aditiva e a noção de atributos, uma função utilidade pode ser descrita simplesmente por um vetor recompensa $\mathbf{w}$, possibilitando a definição de um espaço de vetores recompensas candidatos $\mathcal{W}$. O espaço de vetores recompensas pode ainda ser reduzido ao considerar apenas vetores recompensas normalizados, pois assim desconsidera vetores recompensas que representem as mesmas preferências, isto é, vetores recompensas que são transformações lineares de outros vetores. Ainda, deve-se considerar combinações entre vetores recompensas $\mathbf{w}$ e desvios padrões σ associados à utilidade produzida por $\mathbf{w}$, modelando um usuário não consistente.

As questões que podem ser formuladas pelo agente, devem ser uma descrição de como o agente demonstrará o comportamento, isto é, políticas. Dessa forma, a formulação de uma questão q consiste em definir duas políticas π^1_q e π^2_q que serão utilizadas pelo agente para demonstrar os comportamentos. Deve-se então escolher o conjunto de políticas Π que será utilizado para formular questões, podendo considerar

vários tipos de políticas: estacionárias deterministas, estacionárias estocásticas ou não-estacionárias[2]. Em todos os casos, as respostas do usuário são: `melhor` ou `pior`. Note que as decisões seqüenciais de uma política dependem dos estados do ambiente como observados pelo agente.

Dadas as funções utilidade candidatas e possíveis questões, resta definir explicitamente a regra de inferência da primeira sobre a segunda, que depende da distribuição de probabilidade $\Pr(r|q, \psi^1_{\text{Ag}}, \psi^2_{\text{Ag}}, u)$. Segundo o arcabouço de PMD, essa distribuição de probabilidades é escrita de forma explicita por $\Pr(r|\pi_1, \pi_2, \boldsymbol{\mu}^1_{\text{Ag}}, \boldsymbol{\mu}^2_{\text{Ag}}, \mathbf{w}, \sigma)$. Como ao executar uma política, o comportamento resultante ainda é desconhecido, para formular uma questão o agente precisará também da distribuição de probabilidades $\Pr(r|\pi_1, \pi_2, \mathbf{w}, \sigma)$, que, quando a distribuição de probabilidades $\Pr(\boldsymbol{\mu}_{\text{Ag}}|\pi)$ é conhecida, pode ser calculada por:

$$\Pr(r|\pi_1, \pi_2, \mathbf{w}, \sigma) = \sum_{\boldsymbol{\mu}^1_{\text{Ag}}} \sum_{\boldsymbol{\mu}^2_{\text{Ag}}} \Pr(r|\pi_1, \pi_2, \boldsymbol{\mu}^1_{\text{Ag}}, \boldsymbol{\mu}^2_{\text{Ag}}, \mathbf{w}, \sigma) \Pr(\boldsymbol{\mu}^1_{\text{Ag}}|\pi^1) \Pr(\boldsymbol{\mu}^2_{\text{Ag}}|\pi^2).$$

Na figura 4.3 pode-se ver o modelo de EPCO utilizando o arcabouço de PMD, enquanto o algoritmo 1 é um pseudo-algoritmo para a EPCO no mesmo arcabouço. Nós próximos capítulos serão analisados especificamente como: obter as distribuições de probabilidades utilizadas na formulação de questões e na inferência das aderências de cada função utilidade candidata, assim como definir tal conjunto de funções utilidade candidatas e um conjunto de políticas para serem utilizadas na formulação de questões.

4.3 Problemas na EPCO

Dada a formalização de relação entre agente, usuário e ambiente, pode-se especificar os problemas que aparecem no cenário utilizado na EPCO. O primeiro problema, mesmo que não inviabilize a EPCO, pode trazer junto um alto custo computacional. Na seção 2.4 foi apresentado o modelo de arrependimento esperado para tomada de decisão sob incerteza nas preferências do usuário e métodos baseados nesse modelo para formulação de questões ótimas para melhorar as estimativas das preferências do usuário. No entanto, o custo computacional deste modelo pode ser proibitivo, pois nele todas as possíveis decisões são confrontadas para determinar o pior caso e to-

[2]Nesta tese, políticas não estacionárias não estão relacionadas à dependência no tempo t de um episódio, já que toda política deve considerar tal tempo para atingir otimalidade em tarefas episódicas, mas às observações de atributos ocorridos antes do tempo t, já que, como será visto no capítulo 7, o fato de ser não estacionária pode auxiliar uma política a demonstrar um comportamento próximo de um comportamento arbitrário no espaço de vetores de atributos.

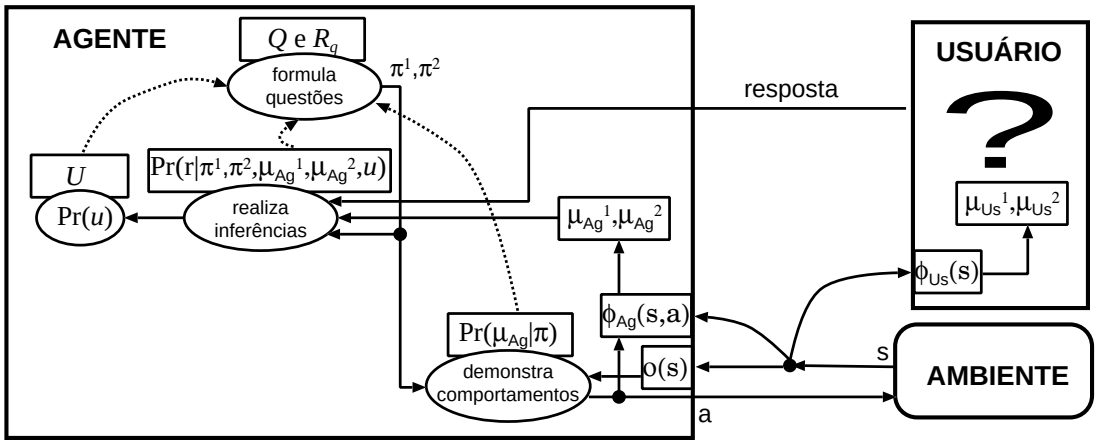

Figura 4.3: Modelo para o EPCO no arcabouço de PMDs. O agente formula uma questão q escolhendo duas políticas π_q^1 e π_q^2 que são executadas pelo agente no ambiente, resultando nos comportamentos $\boldsymbol{\mu}_{\mathrm{Ag}}^1$ e $\boldsymbol{\mu}_{\mathrm{Ag}}^2$ observados pelo agente por meio de $\boldsymbol{\phi}_{\mathrm{Ag}}(s,a)$ e comportamentos $\boldsymbol{\mu}_{\mathrm{Us}}^1$ e $\boldsymbol{\mu}_{\mathrm{Us}}^2$ observados pelo usuário por meio de $\boldsymbol{\phi}_{\mathrm{Us}}(s)$. O usuário escolhe qual comportamento é melhor e responde indicando sua escolha ao agente. As relações entre comportamentos observados pelo agente, questão formulada e respostas do usuário são utilizadas para definir probabilidades de aderência para as funções utilidade candidatas.

mar uma decisão conservadora. Ao utilizar PMD, a cardinalidade $|\Pi|$ do conjunto de políticas disponíveis pode ser muito elevada, tornando ainda maior o custo computacional do uso do arrependimento esperado. Dessa forma, são necessárias técnicas que ajudem a minimizar este problema, tornando factível o uso desse modelo mesmo em problemas de decisões seqüenciais.

Na EP, uma questão é formulada com base em um espaço de conseqüências, considerando que a resposta obtida versará sobre as conseqüências envolvidas. Na EPCO o mesmo procedimento não pode ser utilizado: mesmo que se conheça os comportamentos possíveis, deve-se saber como demonstrar esses comportamentos. Em PMDs, muitas vezes a probabilidade de um determinado comportamento ser obtido é muito baixa, devido ao alto número de comportamentos possíveis e às possibilidades estocásticas em cada momento de decisão. Deve-se então determinar políticas de atuação no ambiente e não comportamentos desejáveis no ambiente. A questão que se coloca é como formular questões em um cenário tão diferente dos utilizados em algoritmos de EP tradicionais?

Outro problema é que, ao responder uma questão, o usuário baseia-se apenas nos comportamentos $\boldsymbol{\mu}_{\mathrm{Us}}^1$ e $\boldsymbol{\mu}_{\mathrm{Us}}^2$ que ele observou e nas suas preferências. No entanto, o agente não tem acesso a tais comportamentos, mas sim aos comportamentos $\boldsymbol{\mu}_{\mathrm{Ag}}^1$ e $\boldsymbol{\mu}_{\mathrm{Ag}}^2$ que ele próprio observou para a questão formulada. Mas, se os comportamentos observados pelo agente diferem dos comportamentos observados pelo usuário como é possível realizar inferências sobre as preferências do usuário, isto é, construir o modelo

Algoritmo 1: Pseudo-algoritmo para EPCO utilizando o arcabouço de PMD.

Data: Π, $\mathcal{W}$, Σ, N, $\Pr(r|\pi^1, \pi^2, \mathbf{w}, \sigma)$, $\Pr(r|\pi^1, \pi^2, \boldsymbol{\mu}^1_{\text{Ag}}, \boldsymbol{\mu}^2_{\text{Ag}}, \mathbf{w}, \sigma)$
define $\mathcal{U} = \{(\mathbf{w}, \sigma)|\mathbf{w} \in \mathcal{W} \wedge \sigma \in \Sigma\}$ e $\Pr(\mathbf{w}, \sigma) = \frac{1}{|\mathcal{U}|}$ para todo $(\mathbf{w}, \sigma) \in \mathcal{U}$;
define $Q = \{(\pi^1, \pi^2)|\pi^1, \pi^2 \in \Pi\}$;
repeat
  forall $q = (\pi^1_q, \pi^2_q) \in Q$ **do**

$$\text{Info}(q) \leftarrow \min_{\pi \in \Pi} \sum_{(\mathbf{w},\sigma) \in \mathcal{U}} \text{Regret}(\pi, \mathbf{w}) \Pr(\mathbf{w}, \sigma) - \sum_{r \in \mathcal{R}_q} \min_{\pi \in \Pi} \sum_{(\mathbf{w},\sigma) \in \mathcal{U}} \text{Regret}(\pi, \mathbf{w}) \Pr(r|\pi^1_q, \pi^2_q, \mathbf{w}, \sigma) \Pr(\mathbf{w}, \sigma)$$

  end
  escolhe $q^* = \max_{q \in Q} \text{Info}(q)$;
  inicializa o primeiro comportamento fazendo $\boldsymbol{\mu}^1_{\text{Ag}} = \mathbf{0}$;
  for $t = 0$ *até* $N - 1$ **do**
    observa o estado do ambiente s_t e executa a ação $a_t = \pi^1_{q^*}(s_t, t)$;
    observa os atributos $\phi_{\text{Ag}}(s_t, a_t)$;
    atualiza $\boldsymbol{\mu}^1_{\text{Ag}} \leftarrow \boldsymbol{\mu}^1_{\text{Ag}} + \phi_{\text{Ag}}(s_t, a_t)$;
  end
  inicializa o segundo comportamento fazendo $\boldsymbol{\mu}^2_{\text{Ag}} = \mathbf{0}$;
  for $t = 0$ *até* $N - 1$ **do**
    observa o estado do ambiente s_t e executa a ação $a_t = \pi^2_{q^*}(s_t, t)$;
    observa os atributos $\phi_{\text{Ag}}(s_t, a_t)$;
    atualiza $\boldsymbol{\mu}^2_{\text{Ag}} \leftarrow \boldsymbol{\mu}^2_{\text{Ag}} + \phi_{\text{Ag}}(s_t, a_t)$;
  end
  usuário escolhe qual comportamento foi melhor emitindo a resposta r;
  forall $(\mathbf{w}, \sigma) \in \mathcal{U}$ **do**

$$\Pr(\mathbf{w}, \sigma|\pi^1_{q^*}, \pi^2_{q^*}, r) \leftarrow \frac{\Pr(r|\pi^1_{q^*}, \pi^2_{q^*}, \mathbf{w}, \sigma) \Pr(\mathbf{w}, \sigma)}{\sum_{(\mathbf{w},\sigma) \in \mathcal{U}} \Pr(r|\pi^1_{q^*}, \pi^2_{q^*}, \mathbf{w}, \sigma) \Pr(\mathbf{w}, \sigma)}$$

  end
until *extração de preferências satisfatória* ;
agente retorna o vetor recompensa esperado $\mathbf{w}_{\text{E}} = \sum_{(\mathbf{w},\sigma) \in \mathcal{U}} \Pr(\mathbf{w}, \sigma)\mathbf{w}$

de inferência $\Pr(r|\pi^1, \pi^2, \boldsymbol{\mu}^1_{\text{Ag}}, \boldsymbol{\mu}^2_{\text{Ag}}, u)$? E se tal modelo de inferência não puder ser construído, existe a possibilidade de extrair as preferências do usuário?

O problema da alta cardinalidade de decisões possíveis impõe principalmente um problema de desempenho computacional, pois influencia diretamente na formulação de questões. O problema de diferentes observações pode impor até mesmo sérias limitações na possibilidade de extrair-se as preferências de um indivíduo. O problema de demonstrar comportamentos em um ambiente não determinista impõe um problema de desempenho da EP junto ao usuário, exigindo uma maior quantidade de interações para poder tomar decisões de forma adequada no lugar do usuário. Esses três problemas serão analisados mais profundamente nos próximos capítulos, onde serão discutidas as condições em que se pode efetivamente realizar a EPCO e como

a EPCO pode ser guiada.

5 ESPAÇO DE POLÍTICAS POSSÍVEIS

Na seção 2.4, a noção de arrependimento foi apresentada como uma técnica tanto para tomar decisões sob incertezas, como para formular questões informativas ao usuário. O cálculo da decisão a ser tomada na equação (2.7), aqui reproduzida para maior clareza:

$$d^* = \arg\min_{d \in \mathcal{D}} \sum_{u \in \mathcal{U}} \mathrm{Regret}(d, u) \Pr(u).$$

e o cálculo da informação que uma questão produz na equação (2.10), também aqui reproduzida:

$$q^* = \arg\min_{q \in \mathcal{Q}} \sum_{r \in \mathcal{R}_q} \min_{d \in \mathcal{D}} \sum_{u \in \mathcal{U}} \mathrm{Regret}(d, u) \Pr(r|q, u) \Pr(u).$$

realizam soma dos arrependimentos, para cada função utilidade candidata, ponderados pelas probabilidades de aderência $\Pr(u)$ das funções utilidade candidatas e uma otimização nas possíveis decisões a serem tomadas. Além disso, para se obter a função $\mathrm{Regret}(d, u)$, deve-se também encontrar a política ótima da função utilidade u.

Quando o arcabouço de PMD é utilizado, tem-se que as possíveis decisões $\mathcal{D}$ a serem tomadas equivalem-se às políticas Π de possível execução pelo agente. No entanto, o conjunto de políticas possíveis Π pode ser muito grande. No caso de políticas deterministas, isto é, $\pi : \mathcal{S} \times \mathbb{N} \rightarrow \mathcal{A}$, a cardinalidade de Π equivale a $N|\mathcal{S} \times \mathcal{A}|$, onde N é a duração máxima do período de demonstração de um comportamento. No caso de políticas estocásticas, isto é, $\pi : \mathcal{S} \times \mathbb{N} \rightarrow (\mathcal{A} \rightarrow [0, 1])$, o problema é ainda maior, pois $|\Pi| = \aleph_0$[1], isto é, a cardinalidade é infinita.

Além da alta cardinalidade, o espaço de possíveis políticas não possui a noção de vizinhança, sendo que uma pequena variação na política – por exemplo, alterando a ação a ser executada em apenas um estado – pode causar uma piora ou melhora significativa no valor de uma política segundo uma função utilidade dada. As faltas de estrutura no espaço de políticas com relação aos comportamentos demonstrados

[1] $\aleph_0$ representa a cardinalidade infinita de primeira ordem, por exemplo, a cardinalidade do conjunto dos reais $\mathbb{R}$.

e, conseqüentemente, no valor atribuído às políticas não permitem que simplificações possam ser feitas de modo a reduzir o custo computacional da EPCO, uma vez que todas políticas devem ser analisadas ao formular uma questão.

Apesar do vetor de atributos esperados de uma política representar apenas parcialmente as propriedades da política, o espaço de vetores de atributos esperados é estruturado e com noção de vizinhança quando atributos numéricos são considerados. Nas próximas seções, utiliza-se o espaço de vetores de atributos esperados para caracterizar o espaço de decisões, permitindo levantar considerações que reduzam o conjunto de políticas a ser considerado ao formular questões, e a seguir um algoritmo que define tal conjunto reduzido de políticas é definido.

5.1 Políticas e Vetores de Atributos Esperados

Um resultado importante na teoria de PMDs é que, dado um PMD $[\mathcal{S}, \mathcal{A}, T(s'|s,a), w(s,a)]$, existe uma política ótima não determinista que o soluciona. Isso permite que o planejamento possa ser limitado apenas a este conjunto restrito de política. Além disso, a política ótima pode ser definida recursivamente baseada na função valor $V^*(s,t)$ que mapeia a recompensa acumulada esperada a partir de s no tempo t até o tempo limite N de um período quando a política ótima é executada, isto é, $V^*(s,t) = \max_{\pi \in \Pi} \mathrm{E}[\sum_{i=t}^{N-1} w(s_i, \pi(s_i)) | s_t = s]$. Define-se então:

$$V^*(s,t) = \begin{cases} 0, & \text{se } t \geq N \\ \max_{a \in \mathcal{A}}[w(s,a) + \sum_{s' \in \mathcal{S}} T(s'|s,a) V^*(s', t+1)], & \text{se } t < N \end{cases},$$

e

$$\pi^*(s,t) \in \{a \mid w(s,a) + \textstyle\sum_{s' \in \mathcal{S}} T(s'|s,a) V^*(s', t+1) \\ \geq \max_{a' \in \mathcal{A}}[w(s,a') + \textstyle\sum_{s' \in \mathcal{S}} T(s'|s,a') V^*(s', t+1)]\}.$$

As características acima também se aplicam ao tomar uma decisão baseada no critério de arrependimento quando as funções utilidade candidatas são lineares, isto é, $u(\psi) = \langle \mathbf{w}_u, \boldsymbol{\mu}(\psi) \rangle$, e a recompensa esperada é

$$w_{\mathrm{E}}(s,a) = \sum_{u \in \mathcal{U}} \Pr(u) \langle \mathbf{w}_u, \boldsymbol{\phi}(s,a) \rangle.$$

Embora um PMD possua características analíticas interessantes para tomar uma decisão baseada no arrependimento, essas mesmas características não se aplicam diretamente no problema de formulação de questões. Na EPCO, as questões devem ser formuladas e demonstradas no ambiente, isso implica que, diferentemente da EP, onde comportamentos podem ser usados diretamente ao formular questões, na EPCO

as questões devem ser formuladas utilizando políticas, sendo os comportamentos utilizados indiretamente.

Na formulação de questões, tudo que se precisa conhecer ao avaliar políticas é a distribuição da probabilidade de ocorrências de cada comportamento $\psi \in \Psi$ para cada política $\pi \in \Pi$, isto é, $\Pr(\psi|\pi)$. Mas tal distribuição não é fácil obter, pois o espaço de comportamentos pode ser muito grande e o cálculo de tal distribuição de probabilidades pode ser proibitivo computacionalmente. Uma opção para descrever tal distribuição de forma compacta é descrever tais conseqüências como vetores de atributos e calcular o vetor de atributos esperados e uma matriz de covariância.

O vetor de atributos esperados $\bar{\boldsymbol{\mu}}^{\pi}$ de uma política determinista π pode ser calculado recursivamente pela equação:

$$\bar{\boldsymbol{\mu}}^{\pi}(s,t) = \begin{cases} [0\ 0\ \cdots\ 0], & \text{se } t \geq N \\ \boldsymbol{\phi}(s,\pi(s)) + \sum_{s' \in \mathcal{S}} T(s'|s,\pi(s))\bar{\boldsymbol{\mu}}^{\pi}(s',t+1), & \text{se } t < N \end{cases},$$

onde $\bar{\boldsymbol{\mu}}^{\pi}(s,t)$ mapeia o vetor de atributos esperados a partir de s no tempo t até o tempo limite N de um período e $\bar{\boldsymbol{\mu}}^{\pi} = \sum_{s \in \mathcal{S}} \Pr(s_0 = s)\bar{\boldsymbol{\mu}}^{\pi}(s,0)$.

A matriz de covariância Σ^{π} também pode ser calculada recursivamente pelas equações:

$$\Sigma^{\pi}(s,t) = \begin{cases} \mathbf{0}, & \text{se } t \geq N \\ \Sigma_{\boldsymbol{\phi}(s,\pi(s))} + \sum_{s' \in \mathcal{S}} T^{\pi,t}_{s,s'} E^{\pi}_{\boldsymbol{\mu}^{\mathsf{T}}\boldsymbol{\mu}}(s',t+1) & \\ \quad -[\sum_{s' \in \mathcal{S}} T^{\pi,t}_{s,s'} \bar{\boldsymbol{\mu}}^{\pi}(s',t+1)]^{\mathsf{T}}[\sum_{s' \in \mathcal{S}} T^{\pi,t}_{s,s'} \bar{\boldsymbol{\mu}}^{\pi}(s',t+1)], & \text{se } t < N \end{cases},$$

onde $T^{\pi,t}_{s,s'} = T(s'|s,\pi(s,t))$, $\Sigma^{\pi}(s,t)$ mapeia a matriz de covariância para o vetor de atributos aleatório $\boldsymbol{\mu}(s,t,\pi) = \sum_{t'=t}^{N-1} \boldsymbol{\phi}(s_{t'}^{(\psi|s_t=s,\pi)}, \pi(s_{t'}^{(\psi|s_t=s,\pi)}, t'))$ [2], $E^{\pi}_{\boldsymbol{\mu}^{\mathsf{T}}\boldsymbol{\mu}}(s,t) = [\Sigma^{\pi}(s,t) + \bar{\boldsymbol{\mu}}^{\pi}(s,t)^{\mathsf{T}}\bar{\boldsymbol{\mu}}^{\pi}(s,t)]$ mapeia a esperança $\mathrm{E}[\boldsymbol{\mu}(s,t,\pi)^{\mathsf{T}}\boldsymbol{\mu}(s,t,\pi)]$ e $\Sigma_{\boldsymbol{\phi}(s,a)}$ é a matriz de covariância da observação $\boldsymbol{\phi}(s,a)$ de ocorrência de atributos. Então, a matriz de covariância Σ^{π} é definida por:

$$\Sigma^{\pi} = \sum_{s \in \mathcal{S}} \Pr(s_0 = s) E^{\pi}_{\boldsymbol{\mu}^{\mathsf{T}}\boldsymbol{\mu}}(s,0) - [\bar{\boldsymbol{\mu}}^{\pi}]^{\mathsf{T}}[\bar{\boldsymbol{\mu}}^{\pi}]. \tag{5.1}$$

Demonstração. Para provar a validade da equação (5.1), primeiro observa-se as seguintes propriedades:

- a covariância $\sigma_{x,y}$ entre duas variáveis x e y pode ser calculada por:

$$\sigma_{x,y} = \mathrm{E}[xy] - \mathrm{E}[x]\mathrm{E}[y]; \text{ e} \tag{5.2}$$

[2] Note que $\boldsymbol{\mu}(s,t,\pi)$ é uma variável (vetor de atributos) aleatória, enquanto $\bar{\boldsymbol{\mu}}^{\pi}(s,t)$ é o valor esperado (vetor de atributos esperados) para tal variável aleatória.

- sejam as variáveis aleatórias x, y, w e z tal que x e y são independentes de w e z, a covariância $\sigma_{x+w,y+z}$ entre as somas $x+w$ e $y+z$ de variáveis aleatórias pode ser calculada por:

$$\sigma_{w+x,y+z} = \sigma_{x,y} + \sigma_{w,z}. \tag{5.3}$$

Então, pela equação (5.2), tem-se que:

$$\begin{aligned}\Sigma^{\pi} &= \textstyle\sum_{s\in\mathcal{S}} \Pr(s_0 = s)\mathrm{E}[\boldsymbol{\mu}(s,0,\pi)^{\mathsf{T}}\boldsymbol{\mu}(s,0,\pi)|s_0 = s] - \bar{\boldsymbol{\mu}}^{\pi\mathsf{T}}\bar{\boldsymbol{\mu}}^{\pi} \\ &= \textstyle\sum_{s\in\mathcal{S}} \Pr(s_0 = s)E^{\pi}_{\boldsymbol{\mu}^{\mathsf{T}}\boldsymbol{\mu}}(s,0) - [\bar{\boldsymbol{\mu}}^{\pi}]^{\mathsf{T}}[\bar{\boldsymbol{\mu}}^{\pi}].\end{aligned}$$

Também é verdade que $E^{\pi}_{\boldsymbol{\mu}^{\mathsf{T}}\boldsymbol{\mu}}(s,t) = [\Sigma^{\pi}(s,t) + \bar{\boldsymbol{\mu}}^{\pi}(s,t)^{\mathsf{T}}\bar{\boldsymbol{\mu}}^{\pi}(s,t)]$, pois, pela equação (5.2), tem-se que:

$$\mathrm{E}[xy] = \sigma_{x,y} + \mathrm{E}[x]\mathrm{E}[y].$$

Por último, pela equação (5.3) e pela equação (5.2), pode-se escrever:

$$\begin{aligned}\Sigma^{\pi}(s,t) &= \Sigma_{\boldsymbol{\phi}(s,\pi(s))} + \{\mathrm{E}[\boldsymbol{\mu}(s',t+1,\pi)^{\mathsf{T}}\boldsymbol{\mu}(s',t+1,\pi)|s_t = s] \\ &\quad -\mathrm{E}[\boldsymbol{\mu}(s',t+1,\pi)|s_t = s]^{\mathsf{T}}\mathrm{E}[\boldsymbol{\mu}(s',t+1,\pi)|s_t = s]\} \\ &= \Sigma_{\boldsymbol{\phi}(s,\pi(s))} + \textstyle\sum_{s'\in\mathcal{S}} T^{\pi}_{s,s'}E^{\pi}_{\boldsymbol{\mu}^{\mathsf{T}}\boldsymbol{\mu}}(s',t+1) \\ &\quad -[\textstyle\sum_{s'\in\mathcal{S}} T^{\pi}_{s,s'}\bar{\boldsymbol{\mu}}^{\pi}(s',t+1)]^{\mathsf{T}}[\textstyle\sum_{s'\in\mathcal{S}} T^{\pi}_{s,s'}\bar{\boldsymbol{\mu}}^{\pi}(s',t+1)]\end{aligned}$$

Para o caso $t \geq N$, como $\boldsymbol{\mu}(s,t,\pi) = \mathbf{0}$ é constante, é trivial que a matriz de covariância é nula. Então, fica provada a equação (5.1). □

5.2 Propriedades do Espaço de Vetores de Atributos Esperados

A vantagem de trabalhar no espaço de vetores de atributos é que se pode definir um espaço métrico com a propriedade de que políticas vizinhas neste espaço possuem valores próximos quando avaliados por $\mathbf{w}$. Uma possível distância nesse espaço é simplesmente a distância euclidiana entre dois vetores de atributos. O espaço de atributos pode ser mais compacto, no sentido de que ele apresenta menor dimensão, isto é, enquanto um comportamento apresenta em sua dimensão a duração do período observado e como possíveis valores o espaço de estados $\mathcal{S}$; o espaço de atributos depende apenas do conjunto de atributos escolhido para estruturar as preferências do usuário.

Pode-se provar outra propriedade importante no espaço de vetores de atributos. Ao utilizar políticas estocásticas, o espaço de atributos permite valores contínuos e é um espaço convexo, onde algoritmos de maximização baseados em gradientes

podem ser implementados. Seja o conjunto de políticas estocásticas Π_{Est}, define-se o conjunto de todos os vetores de atributos esperados $\mathcal{M} = \{\bar{\boldsymbol{\mu}}^{\pi} | \pi \in \Pi_{\text{Est}}\}$. O fato do conjunto $\mathcal{M}$ ser convexo implica que, dado dois vetores $\boldsymbol{\mu}', \boldsymbol{\mu}'' \in \mathcal{M}$, então $\boldsymbol{\mu}^{\alpha} \in \mathcal{M}$, onde $\boldsymbol{\mu}^{\alpha} = \alpha\boldsymbol{\mu}' + (1-\alpha)\boldsymbol{\mu}''$. Considere duas políticas π' e π'' tal que seus respectivos vetores de atributos esperados $\bar{\boldsymbol{\mu}}' = \boldsymbol{\mu}'$ e $\bar{\boldsymbol{\mu}}'' = \boldsymbol{\mu}''$, pode-se definir uma política π^{α} que resulta no vetor de atributos esperados $\bar{\boldsymbol{\mu}}^{\alpha} = \boldsymbol{\mu}^{\alpha}$. A política π^{α} é definida para todo $a \in \mathcal{A}$ e todo $s \in \mathcal{S}$ por

$$\Pr(a|s,t,\pi^{\alpha}) = \frac{\alpha \Pr(s|t,\pi') \Pr(a|s,t,\pi') + (1-\alpha)\Pr(s|t,\pi'')\Pr(a|s,t,\pi'')}{\alpha \Pr(s|t,\pi') + (1-\alpha)\Pr(s|t,\pi'')}. \tag{5.4}$$

Demonstração. Primeiro será provado por indução que a equação (5.4) define $\Pr(s|t,\pi^{\alpha}) = \alpha \Pr(s|t,\pi') + (1-\alpha)\Pr(s|t,\pi'')$. Para $t = 0$ tem-se que:

$$\Pr(s|0,\pi^{\alpha}) = \Pr(s_0 = s) = \alpha \Pr(s|0,\pi') + (1-\alpha)\Pr(s|0,\pi''), \tag{5.5}$$

pois para qualquer política π é verdade que $\Pr(s|0,\pi) = \Pr(s_0 = s)$. Para $t > 0$ tem-se que:

$$\Pr(s|t,\pi^{\alpha}) = \sum_{s' \in \mathcal{S}} \sum_{a \in \mathcal{A}} T(s|s',a) \Pr(a|s',t-1,\pi^{\alpha}) \Pr(s'|t-1,\pi^{\alpha}),$$

aplicando a equação (5.4) e a equação (5.5), obtém-se:

$$\begin{aligned}
\Pr(s|t,\pi^{\alpha}) &= \textstyle\sum_{s' \in \mathcal{S}} \sum_{a \in \mathcal{A}} T(s|s',a)[\alpha \Pr(s|t-1,\pi') \Pr(a|s,t-1,\pi') \\
&\quad +(1-\alpha)\Pr(s|t-1,\pi'')\Pr(a|s,t-1,\pi'')] \\
&= \alpha \textstyle\sum_{s' \in \mathcal{S}} \sum_{a \in \mathcal{A}} T(s|s',a) \Pr(s|t-1,\pi') \Pr(a|s,t-1,\pi') \\
&\quad +(1-\alpha) \textstyle\sum_{s' \in \mathcal{S}} \sum_{a \in \mathcal{A}} T(s|s',a) \Pr(s|t-1,\pi'') \Pr(a|s,t-1,\pi'') \\
&= \alpha \Pr(s|t,\pi') + (1-\alpha)\Pr(s|t,\pi'').
\end{aligned}$$

Considere o vetor de atributos esperados $\bar{\boldsymbol{\mu}}^{\pi}(t)$ que indica o acúmulo de atributos esperados a partir da distribuição $\Pr(s_0 = s)$ até o tempo t executando a política π, isto é,

$$\bar{\boldsymbol{\mu}}^{\pi}(t) = \mathrm{E}_{\{\psi \sim \Pr(s_0=s), T(\cdot)\}} \left[\sum_{i=0}^{t} \boldsymbol{\phi}(s_i^{\psi}, a_i^{\psi}) | \pi \right].$$

O vetor $\bar{\boldsymbol{\mu}}^{\pi}(t)$ pode ser definido recursivamente como:

$$\bar{\boldsymbol{\mu}}^{\pi}(t) = \begin{cases} [0\ 0\ \ldots\ 0], & \text{se } t = 0 \\ \bar{\boldsymbol{\mu}}^{\pi}(t-1) & \\ \quad + \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} \boldsymbol{\phi}(s,a) \Pr(s|t-1,\pi) \Pr(a|s,t-1,\pi), & \text{se } t > 0 \end{cases}.$$

Por indução, pode-se provar que

$$\bar{\boldsymbol{\mu}}(t)^{\pi^\alpha} = \alpha\bar{\boldsymbol{\mu}}(t)' + (1-\alpha)\bar{\boldsymbol{\mu}}(t)'' \tag{5.6}$$

é válido para qualquer $t \geq 0$ quando a equação (5.4) é aplicada. A equação (5.6) é válida para $t = 0$ e para $t > 0$ tem-se que:

$$\begin{aligned}
\bar{\boldsymbol{\mu}}^{\pi^\alpha}(t) &= \bar{\boldsymbol{\mu}}^{\pi^\alpha}(t-1) + \textstyle\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} \boldsymbol{\phi}(s,a)\Pr(s|t-1,\pi^\alpha)\Pr(a|s,t-1,\pi^\alpha) \\
&= \alpha\bar{\boldsymbol{\mu}}(t-1)' + (1-\alpha)\bar{\boldsymbol{\mu}}(t-1)'' \\
&\quad +\boldsymbol{\phi}(s,a)[\alpha\Pr(s|t-1,\pi')\Pr(a|s,t-1,\pi') \\
&\quad +(1-\alpha)\Pr(s|t-1,\pi'')\Pr(a|s,t-1,\pi'')] \\
&= \alpha[\bar{\boldsymbol{\mu}}(t-1)' + \textstyle\sum_{s'\in\mathcal{S}}\sum_{a\in\mathcal{A}} \boldsymbol{\phi}(s,a)\Pr(s|t-1,\pi')\Pr(a|s,t-1,\pi')] \\
&\quad +(1-\alpha)[\bar{\boldsymbol{\mu}}(t-1)'' \\
&\quad + \textstyle\sum_{s'\in\mathcal{S}}\sum_{a\in\mathcal{A}} \boldsymbol{\phi}(s,a)\Pr(s|t-1,\pi'')\Pr(a|s,t-1,\pi'')] \\
&= \alpha\bar{\boldsymbol{\mu}}(t)' + (1-\alpha)\bar{\boldsymbol{\mu}}(t)''.
\end{aligned}$$

□

Então, demonstrou-se que dado um vetor de atributos $\boldsymbol{\mu}^\alpha$ convexo a quaisquer dois vetores de atributos esperados, tais que as políticas que os gerem são conhecidas, é possível construir uma política π^α que gera um vetor de atributos esperados $\bar{\boldsymbol{\mu}}^\alpha = \boldsymbol{\mu}^\alpha$. O caso mais genérico, isto é, a política que gere um vetor de atributos esperados convexo ao conjunto de vetores de atributos esperados das políticas deterministas ($\{\bar{\boldsymbol{\mu}}^\pi \mid \pi \in \Pi_{\text{Det}}\}$), também pode ser definido de modo equivalente. Dessa forma, o envoltório convexo $\mathcal{M} = \text{co}(\{\bar{\boldsymbol{\mu}}^\pi | \pi \in \Pi_{\text{Det}}\})$ determina um poliedro de vetores de atributos esperados que podem ser obtidos através de uma política estocástica apropriada.

O poliedro $\mathcal{M}$ é equivalente ao conjunto de todos os vetores de atributos esperados $\bar{\boldsymbol{\mu}}^\pi$ gerados a partir das políticas estocásticas $\pi \in \Pi_{\text{Est}}$, pois se isso não fosse verdade, existiria uma função recompensa para qual não existiria uma política ótima em Π_{Det}. Na próxima seção será definida uma propriedade que permite a construção de um algoritmo que encontra os vértices do poliedro $\mathcal{M}$.

5.3 Gerando um Conjunto de Políticas não Dominadas

O fato dos vetores de atributos esperados de todas as políticas estacionárias, sejam elas deterministas ou estocásticas, estarem limitados ao envoltório convexo $\mathcal{M} = \text{co}(\{\bar{\boldsymbol{\mu}}^\pi | \pi \in \Pi_{\text{Det}}\})$ das políticas deterministas Π_{Det} permite descrever os possíveis

vetores de atributos esperados de forma compacta. Essa representação compacta é feita pelos vértices que representam tal poliedro e são vetores de atributos esperados associados a políticas não dominadas Π_{NonDom}, isto é, $\text{co}(\{\bar{\boldsymbol{\mu}}^{\pi}|\pi \in \Pi_{\text{NonDom}}\}) = \text{co}(\{\bar{\boldsymbol{\mu}}^{\pi}|\pi \in \Pi_{\text{Det}}\}) = \text{co}(\{\bar{\boldsymbol{\mu}}^{\pi}|\pi \in \Pi_{\text{Est}}\})$. Nesta seção, apresenta-se uma propriedade que permite gerar tais vértices, sem a necessidade de gerar todos vetores de atributos esperados que pertencem internamente ao envoltório convexo, isto é, não necessariamente todas políticas Π_{Det} precisam analisadas.

Os pontos extremos para cada um dos atributos necessariamente são vértices, isto é, os vetores de atributos esperados obtidos quando considera-se PMDs com vetores recompensas baseados em apenas um dos atributos. Para o ponto extremo positivo deve-se maximizar o atributo i utilizando o vetor recompensa $\mathbf{w} = [0\ 0\ \cdots\ w_{i-1} = 0\ w_i = 1\ w_{i+1} = 0\ \cdots\ 0\ 0]$ e obtém-se $\bar{\boldsymbol{\mu}}_i^{\max}$. O mesmo deve ser feito para o atributo i minimizando o seu valor, isto é, com $\mathbf{w} = [0\ 0\ \cdots\ w_{i-1} = 0\ w_i = -1\ w_{i+1} = 0\ \cdots\ 0\ 0]$ e obtém-se $\bar{\boldsymbol{\mu}}_i^{\min}$.

Com os vetores de atributos obtidos, pode-se criar limites mínimos para o conjunto factível de interesse. Se for considerado o envoltório convexo sobre esses pontos, isto é, $\widehat{\mathcal{M}} = \text{co}(\{\bar{\boldsymbol{\mu}}_i^{\max}|i \in \Xi\} \cup \{\bar{\boldsymbol{\mu}}_i^{\min}|i \in \Xi\})$, tem-se que $\widehat{\mathcal{M}} \subseteq \mathcal{M}$. Apesar desse conjunto ser uma aproximação para o conjunto de interesse, nada indica que é a melhor aproximação, outras direções poderiam ter sido escolhidas para definir tal aproximação. No entanto, como ela forma uma base, ela pode ser utilizada para determinar valores máximos e mínimos que uma política pode obter consoante uma função utilidade $u(\cdot)$.

Para determinar o conjunto completo $\mathcal{M} = \text{co}(\{\bar{\boldsymbol{\mu}}_{\pi}|\pi \in \Pi_{\text{Det}}\})$ pode-se inicialmente utilizar o conjunto $\widehat{\mathcal{M}}$ e acrescentar novos vértices ao mesmo até se obter $\mathcal{M}$. A estratégia é testar para todas as faces do poliedro $\widehat{\mathcal{M}}$ se elas também são faces reais do poliedro $\mathcal{M}$, caso não seja uma face real, encontra-se pontos além dela e que pertençam ao poliedro $\mathcal{M}$ para definir novas faces candidatas. Esse procedimento é assegurado pelo seguinte teorema.

Teorema 5.1. *Seja o vetor recompensa* $\mathbf{w}$ *e uma política ótima* $\pi_{\mathbf{w}}^*$ *com seu respectivo vetor de atributos esperados* $\bar{\boldsymbol{\mu}}^{\pi_{\mathbf{w}}^*}$. *Ainda, considere o hiperplano* H *que contém o ponto* $\bar{\boldsymbol{\mu}}^{\pi_{\mathbf{w}}^*}$ *e é perpendicular a* $\mathbf{w}$. *Então, os vetores de atributos esperados que estão além do hiperplano* H, *com respeito a direção* $\mathbf{w}$, *não pertencem ao conjunto* $\text{co}(\{\bar{\boldsymbol{\mu}}^{\pi}|\pi \in \Pi_{Est}\})$.

Demonstração. Considere um vetor de atributos qualquer $\boldsymbol{\mu}$, a operação $\langle \mathbf{w}, \boldsymbol{\mu} \rangle$ avalia o vetor de atributos $\boldsymbol{\mu}$ segundo o vetor recompensa $\mathbf{w}$. Isso implica que $\langle \mathbf{w}, \bar{\boldsymbol{\mu}}^{\pi} \rangle \leq \langle \mathbf{w}, \bar{\boldsymbol{\mu}}^{\pi_{\mathbf{w}}^*} \rangle$ para todo $\pi \in \Pi$, pois $\pi_{\mathbf{w}}^*$ é uma política ótima segundo $\mathbf{w}$.

Por outro lado, o hiperplano H é definido pelo vetor $\mathbf{w}$ e o valor $\langle \mathbf{w}, \bar{\boldsymbol{\mu}}^{\pi^*_{\mathbf{w}}} \rangle$, pois $\mathbf{w}$ é perpendicular a H e $\bar{\boldsymbol{\mu}}^{\pi^*_{\mathbf{w}}}$ pertence ao plano. Logo, $\boldsymbol{\mu} \in H$ se e somente se $\langle \mathbf{w}, \boldsymbol{\mu} \rangle = \langle \mathbf{w}, \bar{\boldsymbol{\mu}}^{\pi^*_{\mathbf{w}}} \rangle$ e $\boldsymbol{\mu}$ está além de H se e somente se $\langle \mathbf{w}, \boldsymbol{\mu} \rangle > \langle \mathbf{w}, \bar{\boldsymbol{\mu}}^{\pi^*_{\mathbf{w}}} \rangle$.

Considere que exista um vetor de atributos esperado $\bar{\boldsymbol{\mu}}'$ além de H e que $\bar{\boldsymbol{\mu}}' \in \mathrm{co}(\{\bar{\boldsymbol{\mu}}^{\pi} | \pi \in \Pi_{\mathrm{Est}}\})$, isto é, o vetor de atributos $\bar{\boldsymbol{\mu}}'$ possui uma política, mesmo que estocástica, que o obtenha como vetor de atributos esperados. Então existe uma política melhor que $\pi^*_{\mathbf{w}}$, resultando em contradição. Logo, esse não pode ser o caso. □

Como resultado, esse teorema permite propor uma regra para incrementar o conjunto auxiliar $\widehat{\mathcal{M}}$ até que o mesmo represente o conjunto de vetores de atributos esperados factíveis $\mathcal{M}$. Inicialmente se considera o poliedro $\widehat{\mathcal{M}} = \mathrm{co}(\{\bar{\boldsymbol{\mu}}_i^{\max} | i \in \Xi\} \cup \{\bar{\boldsymbol{\mu}}_i^{\min} | i \in \Xi\})$ (figura 5.1a). Considere um hiperplano H' que limita o poliedro $\widehat{\mathcal{M}}$ e o respectivo vetor $\mathbf{w}_{H'}$ perpendicular ao plano H' e que aponta para fora do poliedro $\widehat{\mathcal{M}}$ (figura 5.1b e figura 5.1c). Encontre uma política ótima $\pi^*_{\mathbf{w}_{H'}}$ para $\mathbf{w}_{H'}$ e o respectivo vetor de atributos esperados $\bar{\boldsymbol{\mu}}^{\pi^*_{\mathbf{w}_{H'}}}$. Se $\bar{\boldsymbol{\mu}}^{\pi^*_{\mathbf{w}_{H'}}}$ está além de H', então $\bar{\boldsymbol{\mu}}^{\pi^*_{\mathbf{w}_{H'}}}$ pode ser considerado como mais um vértice do poliedro auxiliar $\widehat{\mathcal{M}}$ (figura 5.1b), caso contrário o hiperplano H' limita o conjunto $\mathcal{M}$ (figura 5.1c). Se todos os hiperplanos que limitam $\widehat{\mathcal{M}}$ também limitam o conjunto $\mathcal{M}$, então $\widehat{\mathcal{M}} = \mathcal{M}$. O fim desse processo é garantido, uma vez que a cardinalidade de Π_{Det} é finita.

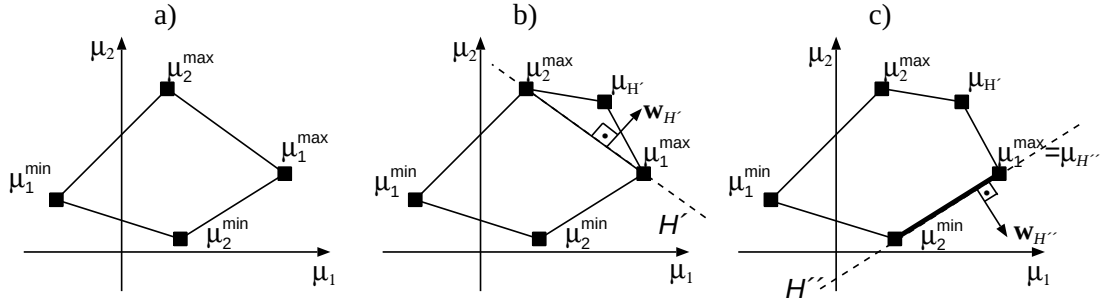

Figura 5.1: Gerando um conjunto de vetores de atributos esperados factíveis. a) Poliedro $\widehat{\mathcal{M}}$ inicial tendo como vértices os vetores que minimizam e maximizam cada um dos atributos isoladamente. b) Exemplo de um plano do poliedro $\widehat{\mathcal{M}}$ que não pertence ao poliedro $\mathcal{M}$. c) Exemplo de um plano do poliedro $\widehat{\mathcal{M}}$ que pertence ao poliedro $\mathcal{M}$.

5.4 Experimentos

Os objetivos dos experimentos realizados neste capítulo é demonstrar o quão compacto a representação do conjunto de políticas Π_{Est} pode ser no espaço de vetores

de atributos esperados. Tal conjunto é representado apenas pelos vértices do poliedro que forma o envoltório convexo dos vetores de atributos de todas políticas estocásticas. O algoritmo para encontrar tal conjunto é testado sob diferentes níveis de precisão, mostrando que a representação pode ser ainda mais compacta, quando quase-otimalidade é considerada.

5.4.1 Tarefa do Agente

A tarefa que se deseja programar no agente é a tarefa de se deslocar de um ponto inicial a um ponto final. Além de completar tal tarefa, o agente deve maximizar ou minimizar a ocorrência de alguns atributos: `distância` (μ_1), `tempo` (μ_2), `curva geral` (μ_3), `curva fechada` (μ_4) e `colisão` (μ_5). No entanto, o compromisso entre tal atributos deve satisfazer as preferências de um usuário.

Como pré-processamento para a EPCO, o agente deve determinar o conjunto de políticas não dominadas, que, como será visto no capítulo 7, é utilizado no processo de formulação de questões da EPCO. Esse conjunto consistirá de políticas dentre as quais certamente encontra-se a política ótima que satisfaz as preferências de qualquer usuário, desde que tal usuário possua preferências lineares e baseie suas preferências nos mesmos atributos observados pelo agente.

Na figura 5.2 é exibido o ambiente considerado nos experimentos. O agente deve partir da posição inicial s_i e chegar na posição final s_f com no máximo 25 ações. Também são exibidas as 105 possíveis posições discretas assumidas pelo agente. Além da posição do agente, também é considerada uma discretização da direção do mesmo, considerando 8 possíveis valores (discretização de 45°). Considerando ainda a passagem do tempo, o robô pode então assumir 21000 estados.

O agente pode escolher, a cada iteração, uma dentre um conjunto de 12 ações: mover-se uma célula em uma das quatro direções cardeais, mover-se duas células em uma das quatro direções cardeais[3], mover-se uma célula em uma das quatro diagonais. Se a próxima posição determinada por uma ação não pode ser ocupada devido a presença de obstáculos, então esta ação não consta como opção ao agente (ver exemplo na figura 5.2, onde, das 12 opções de ações, somente 8 estão disponíveis na situação mostrada). Devido a configuração do ambiente, a quantidade média de ações disponíveis para cada estado é aproximadamente 6,5. O espaço de estados e ações resultam então em um espaço de políticas com $21000^{6,5}$ opções de decisão, ou seja, da ordem de 10^{28} políticas.

[3] Mesmo que exista um obstáculo entre a posição inicial e a posição final, o agente contorna tal obstáculo para chegar à posição desejada.

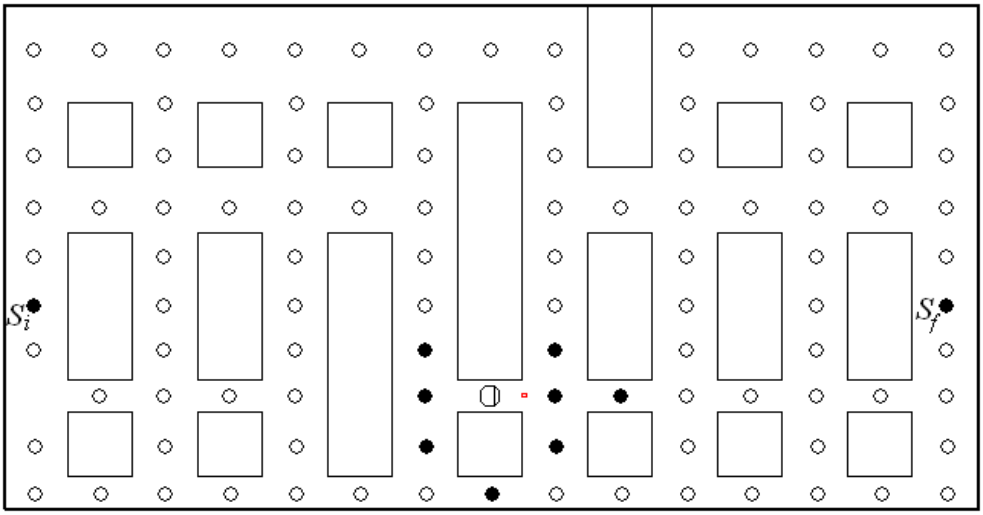

Figura 5.2: Ambiente utilizado nos experimentos para definir conjunto de vetores de atributos factíveis. O tamanho do ambiente é um retângulo com $30m \times 15m$ e obstáculos distribuídos no mesmo de forma estruturada. O robô deve utilizar como posições intermediárias as 105 posições distribuídas no ambiente. O robô inicializa na posição s_i e deve alcançar a posição s_f. Um robô localizado como na figura, em um passo de tempo, pode se mover para qualquer uma das 8 posições marcadas próximas ao robô.

5.4.2 Montagem Experimental

A tarefa de obter uma trajetória entre dois pontos foi simulada por um arcabouço próximo da realidade, o simulador *ARIA* (ACTIVMEDIA ROBOTICS, 2001). Este simulador simula o robô *Pioneer*, que é um robô móvel não holonômico. Na figura 5.3 é exibida uma situação típica do robô, onde são demonstradas as medidas de um dos seus sensores, os sonares, que medem a distância de obstáculos próximos ao robô. Além dos sonares, o robô também possui um sensor de movimento, o hodômetro, que permite estimar a distância percorrida pelo robô, exibida no painel à esquerda.

Junto ao simulador é utilizada a plataforma *SAPHIRA*, que implementa um algoritmo de localização com base em localização markoviana e um algoritmo para controle e planejamento de trajetória baseado em gradientes com relação a posição desejada e os obstáculos no ambiente (ACTIVMEDIA ROBOTICS, 2001). O algoritmo de localização permite minimizar os erros encontrados nas medidas feitas pelos sensores do robô (sonares e hodômetro), e, ao integrar as medidas acumuladas ao longo do tempo, faz com que a postura estimada do robô seja próxima da postura real (a nuvem de pontos vermelhos na figura 5.4 representa possíveis posições onde o robô pode estar). O algoritmo de planejamento permite realizar uma discretização mais grosseira do ambiente, fazendo com que a escolha de ações seja feita sempre em posições discretas a menos de um erro (o robô se posiciona no máximo a $30cm$ do ponto escolhido). Na figura 5.4, o caminho em vermelho é a trajetória planejada até

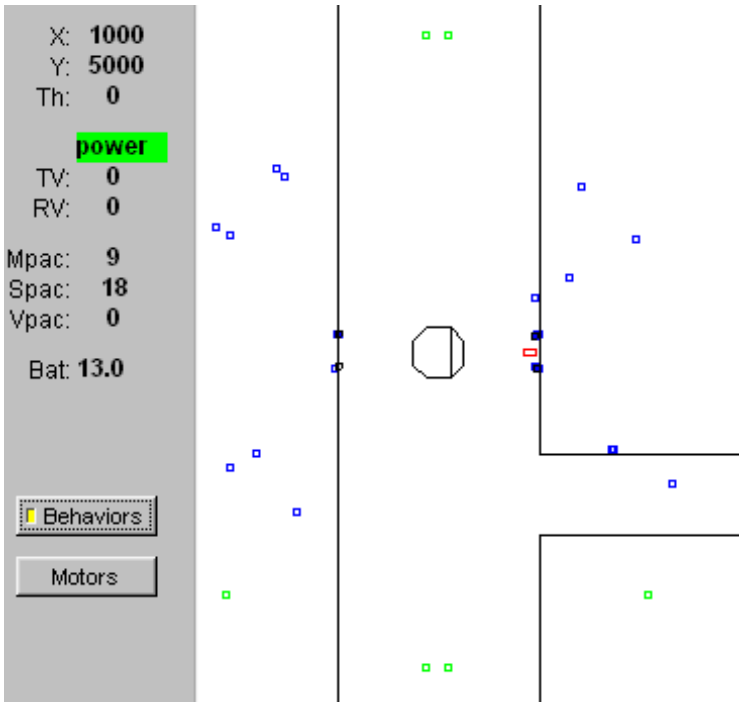

Figura 5.3: Interface do simulador do robô *Pioneer*. Os quadrados verdes representam as medidas atuais dos sonares, enquanto os quadrados azuis representam medidas antigas. O retângulo vermelho representa a direção do robô. A postura do robô estimada pelo acumulo da medida feita pelos hodômetros pode ser vista no painel à esquerda, onde *X* e *Y* representam a posição absoluta a partir de uma origem conhecida e *Th* representa a direção do robô a partir do eixo *X*.

o alvo pelo algoritmo com base em gradientes.

A observação dos 5 atributos considerados pelo agente foi implementada utilizando os sensores do robô. Após um evento discreto de tempo fixo Δt mede-se a variação na posição do robô Δd e a variação da direção do robô $\Delta\theta$. Os atributos `distância` e `tempo` simplesmente medem respectivamente a distância percorrida pelo agente da postura inicial até a posição final ($\sum \Delta d$) e o tempo gasto para realizar este percurso ($\sum \Delta t$). O atributo `curva geral` considera o acúmulo de variações da direção do robô ao longo do percurso ($\sum \Delta\theta$), enquanto o atributo `curva fechada` acumula apenas a variação da direção quando há pouco movimento de translação, isto é, mede curvas mais fechadas. O atributo `curva fechada` foi implementado como $\sum \max\{0, \Delta\theta - 40 * \Delta d\}$ [4], com Δd medido em *metros* e $\Delta\theta$ medido em *graus*. O atributo `colisão` foi implementado analisando o estado dos motores do robô, quando os motores se encontram impedidos de rotacionar, o robô locomove-se em sentido contrário e uma colisão é contabilizada.

Por último, para que algoritmos de planejamento possam ser utilizados, experimentos preparatórios foram realizados para determinar as probabilidades de transições

[4] O coeficiente 40 foi ajustado para que apenas curvas com raio menor que 1,5 metros sejam contabilizadas.

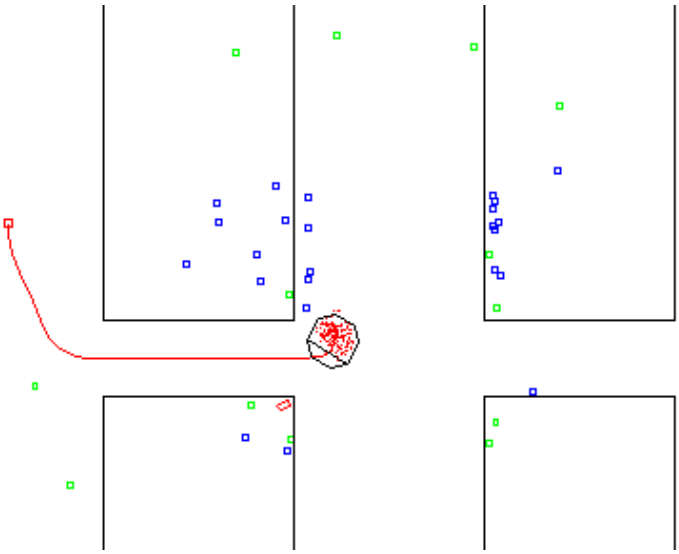

Figura 5.4: Localização e planejamento na plataforma *SAPHIRA*. A nuvem de pontos vermelhos representa possíveis posições do robô, às quais probabilidades de aderência são associadas. O caminho em vermelho representa a trajetória planejada pelo algoritmo baseado em gradiente para atingir a posição alvo.

$T(s'|s,a)$ entre os estados para cada ação e ainda a função de observação $\boldsymbol{\phi}(s,a)$. Essa informação foi levantada durante a repetição sistemática de ao menos 30 vezes da execução de todas as ações em cada uma das possíveis situações.

5.4.3 Resultados

Utilizando as funções $T(s'|s,a)$ e $\boldsymbol{\phi}(s,a)$, é possível calcular a política ótima e o respectivo vetor de atributos esperados para um vetor recompensa $\mathbf{w}$ arbitrário. Calculou-se então os vetores de atributos esperados $\bar{\boldsymbol{\mu}}_i^{\max}$ e $\bar{\boldsymbol{\mu}}_i^{\min}$ referentes às respectivas políticas ótimas para os vetores recompensas $\mathbf{w}_i^{\max} = [0 \ \cdots \ w_{i-1} = 0 \ w_i = 1 \ w_{i+1} = 0 \ \cdots \ 0]$ e $\mathbf{w}_i^{\min} = [0 \ \cdots \ w_{i-1} = 0 \ w_i = -1 \ w_{i+1} = 0 \ \cdots \ 0]$ respectivamente para $i = 1, 2, \ldots, 5$. Os valores dos vetores de atributos esperados podem ser vistos na tabela 5.1, enquanto na tabela 5.2 são exibidos os mesmos valores normalizados entre 0 e 1.

Devido ao grande número de políticas disponíveis, o número de políticas não dominadas ainda pode ser muito alto, no entanto muitas dessas políticas podem trazer pouca melhoria quando confrontadas com outras políticas, mesmo que se utilize uma avaliação arbitrária. Dessa forma, algumas políticas não dominadas podem ser descartadas, sem que ocorra grande perda ao escolher uma política ótima apenas entre as políticas restantes. Então, considerou-se o algoritmo para encontrar os vetores atributos esperados factíveis $\widehat{\mathcal{M}}^{\epsilon}$ sob a limitação de um erro ao testar se um hiperplano pertence ou não pertence ao poliedro que determina $\widehat{\mathcal{M}}^{\epsilon}$. Seja $\bar{\boldsymbol{\mu}} \in H$; $\mathbf{w}_H$, o vetor recompensa perpendicular ao hiperplano H; $\bar{\boldsymbol{\mu}}^{\pi^*_{\mathbf{w}_H}}$, o vetor de atributos esperados da

Tabela 5.1: Vetores de atributos esperados $\bar{\boldsymbol{\mu}}_i^{\max}$ e $\bar{\boldsymbol{\mu}}_i^{\min}$.

atributos	distância	tempo	curva geral	curva fechada	colisão
$\bar{\boldsymbol{\mu}}_1^{\max}$	142,65	4510	8541	5490	6,388
$\bar{\boldsymbol{\mu}}_1^{\min}$	38,63	1241	1111	516	0,000
$\bar{\boldsymbol{\mu}}_2^{\max}$	142,64	4510	8539	5488	6,394
$\bar{\boldsymbol{\mu}}_2^{\min}$	39,66	1150	1049	487	0,508
$\bar{\boldsymbol{\mu}}_3^{\max}$	142,65	4510	8541	5490	6,3881
$\bar{\boldsymbol{\mu}}_3^{\min}$	42,75	1249	874	473	0,000
$\bar{\boldsymbol{\mu}}_4^{\max}$	135,08	4336	8439	5571	5,608
$\bar{\boldsymbol{\mu}}_4^{\min}$	39,11	1176	1016	415	0,390
$\bar{\boldsymbol{\mu}}_5^{\max}$	103,71	3489	4058	2197	17,022
$\bar{\boldsymbol{\mu}}_5^{\min}$	42,29	1204	1068	648	0,000

Tabela 5.2: Vetores de atributos esperados $\bar{\boldsymbol{\mu}}_i^{\max}$ e $\bar{\boldsymbol{\mu}}_i^{\min}$ normalizados.

atributos	distância	tempo	curva geral	curva fechada	colisão
$\bar{\boldsymbol{\mu}}_1^{\max}$	1,0000	1,0000	1,0000	0,9842	0,3753
$\bar{\boldsymbol{\mu}}_1^{\min}$	0,0000	0,0269	0,0308	0,0196	0,0000
$\bar{\boldsymbol{\mu}}_2^{\max}$	1,0000	1,0000	0,9998	0,9838	0,3756
$\bar{\boldsymbol{\mu}}_2^{\min}$	0,0099	0,0000	0,0227	0,0139	0,0298
$\bar{\boldsymbol{\mu}}_3^{\max}$	1,0000	1,0000	1,0000	0,9842	0,3753
$\bar{\boldsymbol{\mu}}_3^{\min}$	0,0396	0,0295	0,0000	0,0113	0,0000
$\bar{\boldsymbol{\mu}}_4^{\max}$	0,9272	0,9484	0,9867	1,0000	0,3294
$\bar{\boldsymbol{\mu}}_4^{\min}$	0,0047	0,0077	0,0185	0,0000	0,0229
$\bar{\boldsymbol{\mu}}_5^{\max}$	0,6257	0,6961	0,4153	0,3456	1,0000
$\bar{\boldsymbol{\mu}}_5^{\min}$	0,0352	0,0160	0,0252	0,0451	0,0000

política ótima sob o vetor recompensa $\mathbf{w}_H$; e ϵ um erro arbitrário; então considera-se $H \in \widehat{\mathcal{M}}^{\epsilon}$ se $\langle \mathbf{w}_H, \bar{\boldsymbol{\mu}} \rangle + \epsilon \geq \langle \mathbf{w}_H, \bar{\boldsymbol{\mu}}^{\pi^*_{\mathbf{w}_H}} \rangle$.

Nos experimentos foram considerados quatro valores de erros ϵ. Na tabela 5.3 são exibidos os resultados obtidos: para cada ϵ é exibida a quantidade de vetores de atributos esperados não dominados utilizada para representar $\widehat{\mathcal{M}}^{\epsilon}$, a quantidade de hiperplanos utilizada para representar $\widehat{\mathcal{M}}^{\epsilon}$, e o hipervolume normalizado que o conjunto $\widehat{\mathcal{M}}^{\epsilon}$ cobre. Quando comparado ao volume do conjunto inicial, mesmo para $\epsilon = 0,1$, o volume obtido é da ordem de dezenas de vezes maior que o volume do conjunto inicial. No entanto, para $\epsilon = 0,0001$ ocorre pouca variação no volume do conjunto de vetores de atributos esperados com relação a $\epsilon = 0,001$, mas a quantidade de vértices e a quantidade de hiperplanos é pelo menos três vezes mais.

Embora o espaço de políticas possíveis possa ser muito grande para um determinado ambiente, a ordem de grandeza do número de políticas que realmente deve ser considerado pelo agente, isto é, as políticas não dominadas, pode ser muito menor. No entanto, o conjunto de políticas não dominadas ainda apresenta uma alta cardinalidade. Mas, quando considerações sobre quase-otimalidade são feitas, o conjunto

Tabela 5.3: Calculando vetores de atributos esperados não dominados.

ϵ	vértices	hiperplanos	volume $(\times 10^{-3})$
conjunto inicial	10	40	0,0015
0,1	15	90	0,1338
0,01	215	4862	0,9979
0,001	1214	30524	1,1753
0,0001	4048	101008	1,1983

de políticas a serem consideradas pode ser ainda mais reduzido, pois pode-se definir um conjunto ainda menor de políticas considerando que uma política é dominada por outra com algum erro, eliminando não apenas políticas dominadas, mas também políticas que possuem uma grande semelhança com políticas já consideradas, as quais não trazem contribuição significativa para o desempenho do agente.

5.5 Considerações Finais

Embora diversos trabalhos discutam sobre a forma de formular questões e as possibilidades de interpretações do usuário, nenhum modelo formal é adotado para estabelecer essas relações. Embora um PMD não represente a riqueza de diversidade com que questões possam ser formuladas por seres humanos, a modelagem do ambiente de interação entre agente e usuário por um PMD permite formalizar de forma simples o principal problema na EP, que é a interpretação das questões. Tanto a interpretação de uma questão como a observação de comportamentos são fenômenos que ocorrem internamente ao usuário e tudo que o agente pode fazer é tentar estimar da forma mais fidedigna a interpretação/observação do usuário. A consideração do arcabouço de PMD será utilizado no próximo capítulo para analisar o problema de diferentes observações entre usuário e agente.

Um ambiente modelado por um PMD permite também que venha à tona problemas particulares da EPCO: o não determinismo na demonstração de comportamentos e a combinação exponencial nas tomadas de decisões das quais os comportamentos são resultantes. Mesmo ao considerar um caso simples de controle estocástico, onde o agente conhece seu estado no ambiente, como em um PMD, guiar um processo de EPCO informativo pode ser difícil. Além das dificuldades que o ambiente pode impor para a realização da EPCO devido ao não determinismo, avaliar a informação resultante das preferências do usuário sobre exemplos de comportamentos obtidos de duas políticas executadas pode ser muito custoso, pois, mesmo uma política determinista pode desdobrar-se em um número muito grande de comportamentos possíveis. O número de políticas possíveis a serem analisadas em métodos mais informativos como

a utilização de arrependimento também pode ser muito custoso. No capítulo 7, as propriedades apresentadas neste capítulo serão utilizadas para reduzir a complexidade da formulação de questões na EPCO.

6 OBSERVABILIDADE E RESTRIÇÕES

Na EPCO, existe a possibilidade de agentes e usuários possuírem percepções diferentes, dificultando a utilização das respostas emitidas pelos usuários sobre os comportamentos observados pelos agentes. Neste capítulo será analisado como um modelo de inferência pode ser definido para que as respostas do usuário possam ser utilizadas corretamente pelo agente. Essa análise consiste em definir condições ideais para que essa utilização não resulte em erros sistemáticos, o que pode levar a uma impossibilidade no aprendizado das preferências do usuário. Nas próximas seções são discutidos possíveis cenários relacionando a percepção do agente, a percepção do usuário e conhecimentos *a priori* dos modelos de ambas percepções, e, dentro desses cenários, análises são feitas análises de quando é possível inferir, ainda que de forma aproximada, as preferências do usuário.

Primeiro, considera-se o caso no qual modelos completos de observações do agente e do usuário são conhecidos, expondo nesse caso ideal como inferências sobre as preferências do usuário podem ser feitas. A seguir, o caso no qual os modelos de observações são desconhecidos é analisado, seja no caso genérico, quando nenhuma suposição a respeito das preferências do usuário são feitas, seja em um caso específico, quando o usuário possui preferências com base em atributos com independência aditiva e neutralidade ao risco.

6.1 Modelos de Observação Conhecidos para o Agente e o Usuário

Suponha que o agente possua o conjunto de atributos Ξ_{Ag} e uma função estocástica de observação de tais atributos $\boldsymbol{\mu}_{\mathrm{Ag}} : \Psi \times \Omega \rightarrow \mathbb{R}^{|\Xi_{\mathrm{Ag}}|}$, onde Ω representa a variação estocástica. Pode-se supor o mesmo para o usuário, ou seja, o conjunto de atributos Ξ_{Us} e a função $\boldsymbol{\mu}_{\mathrm{Us}}$. O caso mais geral é aquele no qual os atributos são diferentes para o agente e o usuário. Nesse caso, nenhuma relação direta pode ser estabelecida com respeito a ψ, $\boldsymbol{\mu}_{\mathrm{Ag}}(\cdot)$ e $\boldsymbol{\mu}_{\mathrm{Us}}(\cdot)$. Alguma relação pode ser estabelecida mediante distribuições fixas na execução dos comportamentos, por exemplo, se

o comportamento ψ é sempre obtido quando executada uma política fixa π.

Na EP é necessário um modelo de inferência que relaciona a probabilidade de uma dada resposta r quando uma questão q é feita ao usuário com função utilidade $u_{\mathrm{Us}} = u$, isto é, $\Pr(r|q,u)$. Na EPCO, uma questão q é formulada pelo agente definindo duas políticas π_q^1 e π_q^2, cujos comportamentos por elas demonstrados e observados pelo usuário, $\boldsymbol{\mu}_{\mathrm{Us}}^1$ e $\boldsymbol{\mu}_{\mathrm{Us}}^2$, deverão ser comparados pelo usuário, que responderá qual é o melhor entre eles. O agente ainda pode utilizar suas próprias observações $\boldsymbol{\mu}_{\mathrm{Ag}}^1$ e $\boldsymbol{\mu}_{\mathrm{Ag}}^2$ dos comportamentos demonstrados, para inferir a relação $\Pr(r|q,u)$, ou seja, na EPCO, a relação de inferência torna-se $\Pr(r|q,\boldsymbol{\mu}_{\mathrm{Ag}}^1,\boldsymbol{\mu}_{\mathrm{Ag}}^2,u)$.

Considerando conhecidas as probabilidades $\Pr(\psi|\pi)$ que definem uma política π, os modelos $\Pr(\boldsymbol{\mu}_{\mathrm{Us}}|\psi)$ de observação do usuário e $\Pr(\boldsymbol{\mu}_{\mathrm{Ag}}|\psi)$ de observação do agente, ambos sobre o espaço de comportamentos $\psi \in \Psi$, tem-se:

$$\begin{aligned}
&\Pr(r|q,\boldsymbol{\mu}_{\mathrm{Ag}}^1,\boldsymbol{\mu}_{\mathrm{Ag}}^2,u) = \\
&\quad = \textstyle\sum_{\boldsymbol{\mu}_{\mathrm{Us}}^1,\boldsymbol{\mu}_{\mathrm{Us}}^2} \Pr(r|\boldsymbol{\mu}_{\mathrm{Us}}^1,\boldsymbol{\mu}_{\mathrm{Us}}^2,u)\Pr(\boldsymbol{\mu}_{\mathrm{Us}}^1,\boldsymbol{\mu}_{\mathrm{Us}}^2|\pi^1,\pi^2,\boldsymbol{\mu}_{\mathrm{Ag}}^1,\boldsymbol{\mu}_{\mathrm{Ag}}^2,u) \\
&\quad = \textstyle\sum_{\boldsymbol{\mu}_{\mathrm{Us}}^1,\boldsymbol{\mu}_{\mathrm{Us}}^2} \Pr(r|\boldsymbol{\mu}_{\mathrm{Us}}^1,\boldsymbol{\mu}_{\mathrm{Us}}^2,u)\Pr(\boldsymbol{\mu}_{\mathrm{Us}}^1,\boldsymbol{\mu}_{\mathrm{Us}}^2|\pi^1,\pi^2,\boldsymbol{\mu}_{\mathrm{Ag}}^1,\boldsymbol{\mu}_{\mathrm{Ag}}^2) \\
&\quad = \textstyle\sum_{\boldsymbol{\mu}_{\mathrm{Us}}^1,\boldsymbol{\mu}_{\mathrm{Us}}^2} \Pr(r|\boldsymbol{\mu}_{\mathrm{Us}}^1,\boldsymbol{\mu}_{\mathrm{Us}}^2,u)\Pr(\boldsymbol{\mu}_{\mathrm{Us}}^1|\pi^1,\boldsymbol{\mu}_{\mathrm{Ag}}^1)\Pr(\boldsymbol{\mu}_{\mathrm{Us}}^2|\pi^2,\boldsymbol{\mu}_{\mathrm{Ag}}^2),
\end{aligned}$$

onde:

$$\Pr(\boldsymbol{\mu}_{\mathrm{Us}}|\boldsymbol{\mu}_{\mathrm{Ag}},\pi) = \frac{\Pr(\boldsymbol{\mu}_{\mathrm{Us}},\boldsymbol{\mu}_{\mathrm{Ag}}|\pi)}{\Pr(\boldsymbol{\mu}_{\mathrm{Ag}}|\pi)} = \frac{\sum_{\psi}\Pr(\boldsymbol{\mu}_{\mathrm{Us}}|\psi)\Pr(\boldsymbol{\mu}_{\mathrm{Ag}}|\psi)\Pr(\psi|\pi)}{\sum_{\psi}\Pr(\boldsymbol{\mu}_{\mathrm{Ag}}|\psi)\Pr(\psi|\pi)}. \quad (6.1)$$

Mesmo quando o modelo de observação do usuário $\Pr(\boldsymbol{\mu}_{\mathrm{Us}}|\psi)$ é conhecido, na EPCO, o modelo de inferência $\Pr(r|q,\boldsymbol{\mu}_{\mathrm{Ag}}^1,\boldsymbol{\mu}_{\mathrm{Ag}}^2,u)$ torna-se mais complexo, sendo necessária uma marginalização nos possíveis vetores de atributos a serem observados pelo usuário ponderada pela expressão na equação (6.1).

Em geral, não são conhecidas informações sobre comportamentos ψ reais, completos e possíveis de serem demonstrados no ambiente, uma vez que uma observação completa do ambiente seria necessária ao agente. Tudo que o agente pode estimar através de experimentação é a distribuição de probabilidades $\Pr(\boldsymbol{\mu}|\pi)$.

Mesmo que a distribuição $\Pr(\boldsymbol{\mu}_{\mathrm{Us}}|\boldsymbol{\mu}_{\mathrm{Ag}},\pi)$ fosse conhecida *a priori*, todas as combinações possíveis de ocorrência dos comportamentos relacionados com as políticas devem ser consideradas. No caso do PMD, onde várias decisões seqüenciais sob estocasticidade são necessárias, o número de comportamentos possíveis cresce exponencialmente, sendo inviável calcular as probabilidades de cada comportamento. Embora, ao planejar uma política ótima no PMD, essas probabilidades são dispensadas devido à consideração de atributos com a propriedade de independência aditiva e neutralidade ao risco, pois só a esperança matemática dos vetores de atributos é

necessária; na formulação de questões na EPCO a distribuição por completo deve ser conhecida para que o critério de arrependimento esperado possa ser utilizado.

6.2 Modelos de Observação Desconhecidos para o Agente e o Usuário

Na seção anterior, um modelo de observação do usuário foi considerado conhecido. No entanto, se tal modelo não for conhecido, quais são as possibilidades de realizar alguma inferência das respostas do usuário? Do ponto de vista do usuário, deseja-se que o agente encontre uma política ótima π^*_{Us} tal que:

$$\sum_{\boldsymbol{\mu}_{\text{Us}}} u_{\text{Us}}(\boldsymbol{\mu}_{\text{Us}}) \Pr(\boldsymbol{\mu}_{\text{Us}}|\pi^*_{\text{Us}}) \geq \sum_{\boldsymbol{\mu}_{\text{Us}}} u_{\text{Us}}(\boldsymbol{\mu}_{\text{Us}}) \Pr(\boldsymbol{\mu}_{\text{Us}}|\pi) \text{ para todo } \pi \in \Pi.$$

Por outro lado, a EP considera um modelo de preferências do usuário $u^*_{\text{Ag}}(\boldsymbol{\mu}_{\text{Ag}})$ que guia o agente para a respectiva política ótima π^*_{Ag}. Deseja-se que $\pi^*_{\text{Us}} = \pi^*_{\text{Ag}}$. Nesta seção, faz-se algumas análises quando este caso pode ocorrer, mesmo não havendo conhecimento sobre os conjuntos de atributos utilizados pelo usuário. Dessa forma, o modelo de inferência deve ser baseado apenas nas observações do agente, resultando em $\Pr(r|\boldsymbol{\mu}^1_{\text{Ag}}, \boldsymbol{\mu}^2_{\text{Ag}}, u)$, e deseja-se especificar quais são as condições para que esse modelo de inferência seja válido para inferir as preferências do usuário.

Fazendo algumas manipulações no valor $V^{\pi^*_{\text{Us}}}_{u_{\text{Us}}}$ da política ótima segundo o usuário, tem-se:

$$\begin{aligned} V^{\pi^*_{\text{Us}}}_{u_{\text{Us}}} &= \textstyle\sum_{\boldsymbol{\mu}_{\text{Us}}} u_{\text{Us}}(\boldsymbol{\mu}_{\text{Us}}) \Pr(\boldsymbol{\mu}_{\text{Us}}|\pi^*_{\text{Us}}) \\ &= \textstyle\sum_{\boldsymbol{\mu}_{\text{Us}}} u_{\text{Us}}(\boldsymbol{\mu}_{\text{Us}}) \sum_{\boldsymbol{\mu}_{\text{Ag}}} \Pr(\boldsymbol{\mu}_{\text{Us}}|\boldsymbol{\mu}_{\text{Ag}}, \pi^*_{\text{Us}}) \Pr(\boldsymbol{\mu}_{\text{Ag}}|\pi^*_{\text{Us}}) \ . \\ &= \textstyle\sum_{\boldsymbol{\mu}_{\text{Ag}}} \Pr(\boldsymbol{\mu}_{\text{Ag}}|\pi^*_{\text{Us}}) \sum_{\boldsymbol{\mu}_{\text{Us}}} u_{\text{Us}}(\boldsymbol{\mu}_{\text{Us}}) \Pr(\boldsymbol{\mu}_{\text{Us}}|\boldsymbol{\mu}_{\text{Ag}}, \pi^*_{\text{Us}}) \end{aligned}$$

Então, uma boa opção é definir $u^*_{\text{Ag}}(\boldsymbol{\mu}_{\text{Ag}}) = \sum_{\boldsymbol{\mu}_{\text{Us}}} u_{\text{Us}}(\boldsymbol{\mu}_{\text{Us}}) \Pr(\boldsymbol{\mu}_{\text{Us}}|\boldsymbol{\mu}_{\text{Ag}}, \pi^*_{\text{Us}})$, já que isso garantiria ao menos à política ótima π^*_{Us} o seu valor correto. No entanto, tal definição de $u^*_{\text{Ag}}(\boldsymbol{\mu}_{\text{Ag}})$ ainda depende da política ótima π^*_{Us}. Outro problema é que ela não garante que:

$$\sum_{\boldsymbol{\mu}_{\text{Ag}}} u^*_{\text{Ag}}(\boldsymbol{\mu}_{\text{Ag}}) \Pr(\boldsymbol{\mu}_{\text{Ag}}|\pi^*_{\text{Us}}) \geq \sum_{\boldsymbol{\mu}_{\text{Ag}}} u^*_{\text{Ag}}(\boldsymbol{\mu}_{\text{Ag}}) \Pr(\boldsymbol{\mu}_{\text{Ag}}|\pi) \text{ para todo } \pi \in \Pi.$$

Considere o ambiente com duas políticas $\pi', \pi'' \in \Pi$, dois comportamentos observados pelo agente $\boldsymbol{\mu}'_{\text{Ag}}$ e $\boldsymbol{\mu}''_{\text{Ag}}$ e dois comportamentos observados pelo usuário $\boldsymbol{\mu}'_{\text{Us}}$ e $\boldsymbol{\mu}''_{\text{Us}}$, onde $u_{\text{Us}}(\boldsymbol{\mu}'_{\text{Us}}) = 1$ e $u_{\text{Us}}(\boldsymbol{\mu}''_{\text{Us}}) = 0$. Considere ainda as seguintes probabilidades $\Pr(\boldsymbol{\mu}'_{\text{Ag}}|\pi') = 0{,}8$, $\Pr(\boldsymbol{\mu}''_{\text{Ag}}|\pi') = 0{,}2$, $\Pr(\boldsymbol{\mu}'_{\text{Ag}}|\pi'') = 1$, $\Pr(\boldsymbol{\mu}''_{\text{Ag}}|\pi'') = 0$,

$\Pr(\boldsymbol{\mu}'_{\mathrm{Us}}|\boldsymbol{\mu}'_{\mathrm{Ag}},\pi')=1$ e $\Pr(\boldsymbol{\mu}'_{\mathrm{Us}}|\boldsymbol{\mu}'_{\mathrm{Ag}},\pi'')=0$. Tem-se então que:

$$\begin{array}{l} V^{\pi'}_{u_{\mathrm{Us}}} = \sum_{\boldsymbol{\mu}_{\mathrm{Us}}} u_{\mathrm{Us}}(\boldsymbol{\mu}_{\mathrm{Us}}) \sum_{\boldsymbol{\mu}_{\mathrm{Ag}}} \Pr(\boldsymbol{\mu}_{\mathrm{Us}}|\boldsymbol{\mu}_{\mathrm{Ag}},\pi') \Pr(\boldsymbol{\mu}_{\mathrm{Ag}}|\pi') \\ \quad = u_{\mathrm{Us}}(\boldsymbol{\mu}'_{\mathrm{Us}}) \Pr(\boldsymbol{\mu}'_{\mathrm{Us}}|\boldsymbol{\mu}'_{\mathrm{Ag}},\pi') \Pr(\boldsymbol{\mu}'_{\mathrm{Ag}}|\pi') = 0{,}8 \end{array}, \text{ e } V^{\pi''}_{u_{\mathrm{Us}}} = 0,$$

logo a política ótima é π' e tem-se $u^*_{\mathrm{Ag}}(\boldsymbol{\mu}'_{\mathrm{Ag}}) = 0{,}8$ e $u^*_{\mathrm{Ag}}(\boldsymbol{\mu}''_{\mathrm{Ag}}) = 0$. No entanto, quando as políticas são avaliadas segundo $u^*_{\mathrm{Ag}}(\boldsymbol{\mu}_{\mathrm{Ag}})$, tem-se o oposto, isto é, que π'' é melhor que π', pois $V^{\pi'}_{u^*_{\mathrm{Ag}}} < V^{\pi''}_{u^*_{\mathrm{Ag}}}$. Neste caso, uma política que não é ótima pode ser elegida como tal quando ela é avaliada pela utilidade construída a custa da política ótima real. Nas próximas seções serão analisadas formas de definir a função utilidade para evitar esse problema.

6.2.1 Dependência entre Avaliação e Política Executada

Considere um cenário onde se pode definir um mapeamento M entre a observação do agente $\boldsymbol{\mu}_{\mathrm{Ag}}$ e o comportamento real ocorrido ψ de forma unívoca, isto é, $M(\boldsymbol{\mu}) = \psi$. Isso implica que, dada uma observação $\boldsymbol{\mu}_{\mathrm{Ag}}$, existe um ψ único, independentemente da política executada, tal que $\Pr(\psi|\boldsymbol{\mu}_{\mathrm{Ag}}) = 1$. Essa condição é suficiente para que $u^*_{\mathrm{Ag}}(\boldsymbol{\mu}_{\mathrm{Ag}})$ independa da política ótima, pois:

$$\Pr(\boldsymbol{\mu}_{\mathrm{Us}}|\boldsymbol{\mu}_{\mathrm{Ag}},\pi^*_{\mathrm{Us}}) = \sum_{\psi} \Pr(\boldsymbol{\mu}_{\mathrm{Us}}|\psi) Pr(\psi|\boldsymbol{\mu}_{\mathrm{Ag}}) = \Pr(\boldsymbol{\mu}_{\mathrm{Us}}|M(\boldsymbol{\mu}_{\mathrm{Ag}}))$$

e

$$u^*_{\mathrm{Ag}}(\boldsymbol{\mu}_{\mathrm{Ag}}) = \sum_{\boldsymbol{\mu}_{\mathrm{Us}}} u_{\mathrm{Us}}(\boldsymbol{\mu}_{\mathrm{Us}}) \Pr(\boldsymbol{\mu}_{\mathrm{Us}}|M(\boldsymbol{\mu}_{\mathrm{Ag}})).$$

Nesse caso, o espaço de vetor de atributos $\boldsymbol{\mu}_{\mathrm{Ag}}$ deve ser maior ou igual ao espaço Ψ e a observação é determinista. Agora se pode elaborar um caso mais relaxado que, embora não mantenha a independência sobre a política executada, pode-se manter as propriedades que tal independência traz.

Considere a distribuição $\Pr(\boldsymbol{\mu}_{\mathrm{Ag}}|\psi)$, então pode-se definir:

$$M(\boldsymbol{\mu}_{\mathrm{Ag}}) = \arg\max_{\psi} \Pr(\boldsymbol{\mu}_{\mathrm{Ag}}|\psi)$$

e a situação aproxima-se da situação ideal quanto mais próximo $\Pr(\boldsymbol{\mu}_{\mathrm{Ag}}|M(\boldsymbol{\mu}_{\mathrm{Ag}}))$ estiver de 1. Considere também o caso onde qualquer política executada gera cada possível comportamento com probabilidade mínima $p^{\{\Psi|\Pi\}}_{\min} = \min_{\pi\in\Pi,\psi\in\Psi} \Pr(\psi|\pi)$. Isso impõe restrições nas políticas que serão consideradas pelo agente, caso um valor limite para $p^{\{\Psi|\Pi\}}_{\min}$ seja desejado.

As propriedades de independência da política executada começam a ser notadas

quando:

$$\Pr(\boldsymbol{\mu}_{\mathrm{Ag}}|M(\boldsymbol{\mu}_{\mathrm{Ag}}))p_{\min}^{\{\Psi|\Pi\}} \gg (1-\Pr(\boldsymbol{\mu}_{\mathrm{Ag}}|M(\boldsymbol{\mu}_{\mathrm{Ag}})))(1-p_{\min}^{\{\Psi|\Pi\}}), \tag{6.2}$$

pois:

$$\Pr(\boldsymbol{\mu}_{\mathrm{Us}}|\boldsymbol{\mu}_{\mathrm{Ag}},\pi) = \frac{\sum_{\psi}\Pr(\boldsymbol{\mu}_{\mathrm{Us}}|\psi)\Pr(\boldsymbol{\mu}_{\mathrm{Ag}}|\psi)\Pr(\psi|\pi)}{\sum_{\psi}\Pr(\boldsymbol{\mu}_{\mathrm{Ag}}|\psi)\Pr(\psi|\pi)},$$

e quando a condição na equação (6.2) é assegurada, ocorre que

$$\begin{aligned}&\Pr(\boldsymbol{\mu}_{\mathrm{Ag}}|M(\boldsymbol{\mu}_{\mathrm{Ag}}))\Pr(M(\boldsymbol{\mu}_{\mathrm{Ag}})|\pi) \geq \Pr(\boldsymbol{\mu}_{\mathrm{Ag}}|M(\boldsymbol{\mu}_{\mathrm{Ag}}))p_{\min}^{\{\Psi|\Pi\}} \gg \\ &\quad \gg (1-\Pr(\boldsymbol{\mu}_{\mathrm{Ag}}|M(\boldsymbol{\mu}_{\mathrm{Ag}})))(1-p_{\min}^{\{\Psi|\Pi\}}) \geq \textstyle\sum_{\psi\neq M(\boldsymbol{\mu}_{\mathrm{Ag}})}\Pr(\boldsymbol{\mu}_{\mathrm{Ag}}|\psi)\Pr(\psi|\pi)\end{aligned},$$

fazendo com que o termo $\Pr(\boldsymbol{\mu}_{\mathrm{Us}}|\psi)$ seja relevante apenas para $\psi = M(\boldsymbol{\mu}_{\mathrm{Ag}})$. Logo:

$$\begin{aligned}\Pr(\boldsymbol{\mu}_{\mathrm{Us}}|\boldsymbol{\mu}_{\mathrm{Ag}},\pi^*_{\mathrm{Us}}) &= \frac{\sum_{\psi\neq M(\boldsymbol{\mu}_{\mathrm{Ag}})}\Pr(\boldsymbol{\mu}_{\mathrm{Us}}|\psi)\Pr(\boldsymbol{\mu}_{\mathrm{Ag}}|\psi)\Pr(\psi|\pi^*_{\mathrm{Us}})}{\Pr(\boldsymbol{\mu}_{\mathrm{Ag}}|M(\boldsymbol{\mu}_{\mathrm{Ag}}))\Pr(M(\boldsymbol{\mu}_{\mathrm{Ag}})|\pi^*_{\mathrm{Us}})+\sum_{\psi\neq M(\boldsymbol{\mu}_{\mathrm{Ag}})}\Pr(\boldsymbol{\mu}_{\mathrm{Ag}}|\psi)\Pr(\psi|\pi^*_{\mathrm{Us}})}+ \\ &\quad \frac{\Pr(\boldsymbol{\mu}_{\mathrm{Us}}|M(\boldsymbol{\mu}_{\mathrm{Ag}}))\Pr(\boldsymbol{\mu}_{\mathrm{Ag}}|M(\boldsymbol{\mu}_{\mathrm{Ag}}))\Pr(M(\boldsymbol{\mu}_{\mathrm{Ag}})|\pi^*_{\mathrm{Us}})}{\Pr(\boldsymbol{\mu}_{\mathrm{Ag}}|M(\boldsymbol{\mu}_{\mathrm{Ag}}))\Pr(M(\boldsymbol{\mu}_{\mathrm{Ag}})|\pi^*_{\mathrm{Us}})+\sum_{\psi\neq M(\boldsymbol{\mu}_{\mathrm{Ag}})}\Pr(\boldsymbol{\mu}_{\mathrm{Ag}}|\psi)\Pr(\psi|\pi^*_{\mathrm{Us}})} \\ &\cong \frac{\Pr(\boldsymbol{\mu}_{\mathrm{Us}}|M(\boldsymbol{\mu}_{\mathrm{Ag}}))\Pr(\boldsymbol{\mu}_{\mathrm{Ag}}|M(\boldsymbol{\mu}_{\mathrm{Ag}}))\Pr(M(\boldsymbol{\mu}_{\mathrm{Ag}})|\pi^*_{\mathrm{Us}})}{\Pr(\boldsymbol{\mu}_{\mathrm{Ag}}|M(\boldsymbol{\mu}_{\mathrm{Ag}}))\Pr(M(\boldsymbol{\mu}_{\mathrm{Ag}})|\pi^*_{\mathrm{Us}})} \\ &\cong \Pr(\boldsymbol{\mu}_{\mathrm{Us}}|M(\boldsymbol{\mu}_{\mathrm{Ag}}))\end{aligned}.$$

O problema de definir uma função utilidade com base na política ótima pode então ser contornado ao considerar apenas políticas com uma certa aleatoriedade. Essa aleatoriedade garante que o valor $u_{\mathrm{Ag}}(\boldsymbol{\mu}_{\mathrm{Ag}})$ inferido para o vetor de atributos $\boldsymbol{\mu}_{\mathrm{Ag}}$ seja similar independente da política utilizada ao questionar o usuário. Mas, as condições colocadas são difíceis de serem encontradas em um problema real de EPCO. Primeiro, demonstrar todos os possíveis comportamentos Ψ com uma probabilidade $p_{\min}^{\{\Psi|\Pi\}}$ relevante nem sempre é possível em um problema de decisões seqüenciais, pois um comportamento específico depende de uma longa cadeia de acasos, fazendo com que alguns comportamentos sejam obtidos apenas com probabilidades muito baixas. Por outro lado, se um comportamento é difícil de ser obtido, o valor de uma política depende pouco do valor real atribuído a tal comportamento. Ainda assim, apenas garantir que todas políticas sejam executadas com uma probabilidade mínima é difícil, já que tal conjunto pode ser muito grande. Além disso, fazer suposições a respeito da qualidade da observação do agente também é irreal, já que nem sempre o agente possui uma observação completa do ambiente, ainda mais quando se trata de conhecer os atributos relevantes para o usuário.

Na próxima seção, suposições mais relaxadas sobre as observações do agente são feitas, mas restrições sobre a avaliação de comportamentos pelo usuário são impostas, garantindo que o usuário possa aprender a avaliar políticas independentemente da política utilizada para definir $u_{\mathrm{Ag}}(\boldsymbol{\mu}_{\mathrm{Ag}})$.

6.2.2 Espaço métrico e Estrutura da Função Utilidade

Até agora, em nenhum momento foi utilizada a suposição de que o espaço de vetores de atributos permite a criação de um espaço métrico para os comportamentos, atribuindo distância entre os vetores de atributos. Junto a essa definição de distância, também se pode fazer a suposição de que dois comportamentos próximos possuem utilidades próximas segundo o usuário. Deste modo, ao calcular $u_{\mathrm{Ag}}(\boldsymbol{\mu}_{\mathrm{Ag}})$, quanto mais acumulados espacialmente estão os vetores de atributos dos comportamentos $\boldsymbol{\mu}_{\mathrm{Us}}$ que são relevantes, dada a ponderação $\Pr(\boldsymbol{\mu}_{\mathrm{Us}}|\boldsymbol{\mu}_{\mathrm{Ag}}, \pi)$, mais próximas as utilidades utilizadas na ponderação podem estar, fazendo com que a aproximação $\Pr(\boldsymbol{\mu}_{\mathrm{Us}}|\boldsymbol{\mu}_{\mathrm{Ag}}, \pi) \cong \Pr(\boldsymbol{\mu}_{\mathrm{Us}}|M(\boldsymbol{\mu}_{\mathrm{Ag}}))$ gere menos erro.

Considere um comportamento real ψ ocorrido no ambiente e probabilidades de observação $\Pr(\boldsymbol{\mu}_{\mathrm{Us}}|\psi)$ e $\Pr(\boldsymbol{\mu}_{\mathrm{Ag}}|\psi)$. Define-se então a menor hiperesfera $S_{\mathrm{Us}}^{\boldsymbol{\mu}_{\mathrm{Ag}},\pi,\epsilon}$ no espaço de vetores de atributos observados pelo usuário tal que

$$\sum_{\boldsymbol{\mu}_{\mathrm{Us}} \in S_{\mathrm{Us}}^{\boldsymbol{\mu}_{\mathrm{Ag}},\pi,\epsilon}} \sum_{\psi \in \Psi} \Pr(\boldsymbol{\mu}_{\mathrm{Us}}|\psi) \Pr(\psi|\boldsymbol{\mu}_{\mathrm{Ag}}, \pi) \geq \epsilon,$$

onde $\epsilon \leq 1$. Então, para uma política π arbitrária, pode-se definir $u_{\mathrm{Ag}}(\boldsymbol{\mu}_{\mathrm{Ag}}|\pi)$ como:

$$\begin{aligned} u_{\mathrm{Ag}}(\boldsymbol{\mu}_{\mathrm{Ag}}|\pi) &= \textstyle\sum_{\boldsymbol{\mu}_{\mathrm{Us}}} u_{\mathrm{Us}}(\boldsymbol{\mu}_{\mathrm{Us}}) \Pr(\boldsymbol{\mu}_{\mathrm{Us}}|\boldsymbol{\mu}_{\mathrm{Ag}}, \pi) \\ &= \textstyle\sum_{\boldsymbol{\mu}_{\mathrm{Us}}} u_{\mathrm{Us}}(\boldsymbol{\mu}_{\mathrm{Us}}) \sum_{\psi \in \Psi} \Pr(\boldsymbol{\mu}_{\mathrm{Us}}|\psi) \Pr(\psi|\boldsymbol{\mu}_{\mathrm{Ag}}\pi,) \\ &= \textstyle\sum_{\boldsymbol{\mu}_{\mathrm{Us}} \in S_{\mathrm{Us}}^{\boldsymbol{\mu}_{\mathrm{Ag}},\pi,\epsilon}} u_{\mathrm{Us}}(\boldsymbol{\mu}_{\mathrm{Us}}) \sum_{\psi \in \Psi} \Pr(\boldsymbol{\mu}_{\mathrm{Us}}|\psi) \Pr(\psi|\boldsymbol{\mu}_{\mathrm{Ag}}, \pi) \\ &\quad + \textstyle\sum_{\boldsymbol{\mu}_{\mathrm{Us}} \notin S_{\mathrm{Us}}^{\boldsymbol{\mu}_{\mathrm{Ag}},\pi,\epsilon}} u_{\mathrm{Us}}(\boldsymbol{\mu}_{\mathrm{Us}}) \sum_{\psi \in \Psi} \Pr(\boldsymbol{\mu}_{\mathrm{Us}}|\psi) \Pr(\psi|\boldsymbol{\mu}_{\mathrm{Ag}}, \pi). \end{aligned}$$

Considerando que as avaliações do usuário são normalizadas, isto é, $\max_{\boldsymbol{\mu}_{\mathrm{Us}}} u_{\mathrm{Us}}(\boldsymbol{\mu}_{\mathrm{Us}}) = 1$ e $\min_{\boldsymbol{\mu}_{\mathrm{Us}}} u_{\mathrm{Us}}(\boldsymbol{\mu}_{\mathrm{Us}}) = 0$, pode-se definir limites para a utilidade $u_{\mathrm{Ag}}(\boldsymbol{\mu}_{\mathrm{Ag}}|\pi)$ como a seguir:

$$\epsilon \min_{\boldsymbol{\mu}_{\mathrm{Us}} \in S_{\mathrm{Us}}^{\boldsymbol{\mu}_{\mathrm{Ag}},\pi,\epsilon}} u_{\mathrm{Us}}(\boldsymbol{\mu}_{\mathrm{Us}}) \leq u_{\mathrm{Ag}}(\boldsymbol{\mu}_{\mathrm{Ag}}|\pi) \leq \epsilon \max_{\boldsymbol{\mu}_{\mathrm{Us}} \in S_{\mathrm{Us}}^{\boldsymbol{\mu}_{\mathrm{Ag}},\pi,\epsilon}} u_{\mathrm{Us}}(\boldsymbol{\mu}_{\mathrm{Us}}) + (1 - \epsilon).$$

Dessa forma, deseja-se que:

$$\epsilon[\max_{\boldsymbol{\mu}_{\mathrm{Us}} \in S_{\mathrm{Us}}^{\boldsymbol{\mu}_{\mathrm{Ag}},\pi,\epsilon}} u_{\mathrm{Us}}(\boldsymbol{\mu}_{\mathrm{Us}}) - \min_{\boldsymbol{\mu}_{\mathrm{Us}} \in S_{\mathrm{Us}}^{\boldsymbol{\mu}_{\mathrm{Ag}},\pi,\epsilon}} u_{\mathrm{Us}}(\boldsymbol{\mu}_{\mathrm{Us}})] + (1 - \epsilon) \gtrapprox 0.$$

Dadas essas definições, pode-se fazer considerações a respeito das observações do agente, observações do usuário e preferências do usuário. No espaço de vetores de atributos observados, tanto pelo agente como pelo usuário, pode-se estabelecer uma medida de distância – por exemplo, a distância euclidiana. Embora o agente não observe o ambiente como o usuário, aquele pode representar este adequadamente se:

1. sejam quaisquer duas observações do usuário $\boldsymbol{\mu}'_{\mathrm{Us}}, \boldsymbol{\mu}''_{\mathrm{Us}}$, então, dada uma função utilidade $u_{\mathrm{Us}}(\cdot)$, existe um fator β tal que $|u_{\mathrm{Us}}(\boldsymbol{\mu}'_{\mathrm{Us}}) - u_{\mathrm{Us}}(\boldsymbol{\mu}''_{\mathrm{Us}})| \leq \beta|\boldsymbol{\mu}'_{\mathrm{Us}} - \boldsymbol{\mu}''_{\mathrm{Us}}|$; e

2. uma observação do agente $\boldsymbol{\mu}_{\mathrm{Ag}}$ mapeia-se em uma pequena região do espaço de observação do usuário, isto é, a região $S_{\mathrm{Us}}^{\boldsymbol{\mu}_{\mathrm{Ag}},\pi,\epsilon}$ é pequena para $\epsilon = 1$ e qualquer $\boldsymbol{\mu}_{\mathrm{Ag}}$ e π.

A primeira condição relaciona a observação do usuário com as suas preferências, segunda as quais observações semelhantes possuem utilidades semelhantes, não permitindo mudanças bruscas na função utilidade $u_{\mathrm{Us}}(\cdot)$. Um exemplo de função utilidade com tal característica é a função utilidade linear $u_{\mathrm{Us}}(\boldsymbol{\mu}_{\mathrm{Us}}) = \langle \mathbf{w}, \boldsymbol{\mu}_{\mathrm{Us}}(\psi) \rangle$, onde $\beta = \frac{1}{\sqrt{|\Xi_{\mathrm{Us}}|}}$ quando a função utilidade u_{Us} é normalizada. A segunda condição relaciona as observações do agente e as observações do usuário. Se, para qualquer observação do agente $\boldsymbol{\mu}_{\mathrm{Ag}}$, sob quaisquer condições, for verdade que as observações possíveis do usuário são limitadas a um espaço pequeno, pela primeira propriedade, a utilidade atribuída a $\boldsymbol{\mu}_{\mathrm{Ag}}$ também é limitada por um intervalo pequeno, isto é, $u_{\mathrm{Ag}}(\boldsymbol{\mu}_{\mathrm{Ag}}) \in [u_{\min}, u_{\max}]$, tal que $[u_{\max} - u_{\min}] \gtrsim 0$.

Enquanto a primeira condição pode ser facilmente obtida, a segunda dificilmente poderá ser mantida para $\epsilon = 1$. No entanto, se a esfera $S_{\mathrm{Us}}^{\boldsymbol{\mu}_{\mathrm{Ag}},\pi,\epsilon}$ for pequena para ϵ próximo a 1, o agente pode representar de forma eficaz o usuário.

Ainda, se a função utilidade que representa as preferências do usuário é linear e as variações da observação de atributos do agente podem ser aproximadas por uma distribuição normal, mesmo que, para obter ϵ próximo a 1, a esfera $S_{\mathrm{Us}}^{\boldsymbol{\mu}_{\mathrm{Ag}},\pi,\epsilon}$ tenha que ser grande, o agente ainda pode tomar decisões adequadas no lugar do usuário. Apesar dos valores atribuídos aos vetores de atributos dependerem da política executada, se a mesma política é utilizada para a obtenção de todos os vetores, uma parte do valor atribuído é o valor esperado da política π utilizada, isto é, $\sum_{\psi \in \Psi} u_{\mathrm{Us}}^{\psi} \Pr(\psi|\pi)$. Essa dependência afeta a avaliação feita pelo agente para qualquer política, possibilitando que ao menos a ordenação das políticas seja mantida segunda as preferências do usuário.

6.3 Experimentos

Os objetivos dos experimentos realizados neste capítulo é demonstrar que mesmo em um ambiente onde o agente possui não determinismo nas suas observações, se as preferências do usuário apresentam independência aditiva e neutralidade ao risco, então esse não determinismo não afeta significativamente a qualidade da decisão que

o agente pode tomar após inferir as preferências do usuário. Por outro lado, quando as preferências do usuário não apresentam nenhuma estrutura, torna-se inviável a EPCO, resultando em tomadas de decisões diferentes da decisão ótima.

A comparação de qualidade da EPCO entre um usuário com preferências estruturadas e um usuário com preferências sem estrutura será feita em um ambiente sintético simulando os dois tipos de usuários. Para o primeiro usuário, sua função utilidade u_{Des} não apresenta nenhuma estrutura, e nesse caso as propriedades métricas dos atributos não trazem nenhum benefício. Para o segundo usuário, a função utilidade u_{Est} apresenta propriedades de independência aditiva e neutralidade ao risco, podendo portanto ser representada por uma função linear, e, nesse caso, as propriedades métricas dos atributos são totalmente desejáveis.

6.3.1 Tarefa do Agente

O agente deve construir a função utilidade que representa as preferências do usuário. Para extrair tais preferências, o agente repete os seguintes passos:

- o agente executa uma política estacionária,
- o agente observa o comportamento resultante segundo sua percepção,
- o usuário observa o comportamento resultante segundo sua percepção,
- o usuário emite a utilidade do comportamento observado, e
- o agente associa a utilidade emitida pelo usuário à sua própria observação.

Note que, devido ao não determinismo, um mesmo comportamento observado pelo agente pode receber diferentes valores de utilidade pelo usuário. Dessa forma, o valor atribuído pelo agente a um vetor de atributos é a média das utilidades associadas a tal vetor.

6.3.2 Montagem Experimental

Modelo do Usuário

Os comportamentos executados pelo agente são observados pelo usuário como apenas dois atributos, onde cada atributo assume 10 valores $(1, 2, \ldots, 10)$, resultando em 100 comportamentos possíveis. A política executada pelo agente é estocástica e é uma combinação estocástica entre 3 políticas. Os comportamentos observados pelo usuário, quando uma dada política é executada, dependem de uma distribuição

de probabilidades. Na figura 6.1 pode-se ver tal distribuição, onde se nota que, embora haja comportamentos comuns decorrentes de diferentes políticas, existem comportamentos que só ocorrem com uma dada política. Como será apresentado a seguir, tais distribuições, associadas à observação não determinista do agente, possibilitam que o agente atribua utilidade mesmo a comportamentos não ocorridos no ambiente, evidenciando os problemas inerentes de observações diferentes entre agente e usuário.

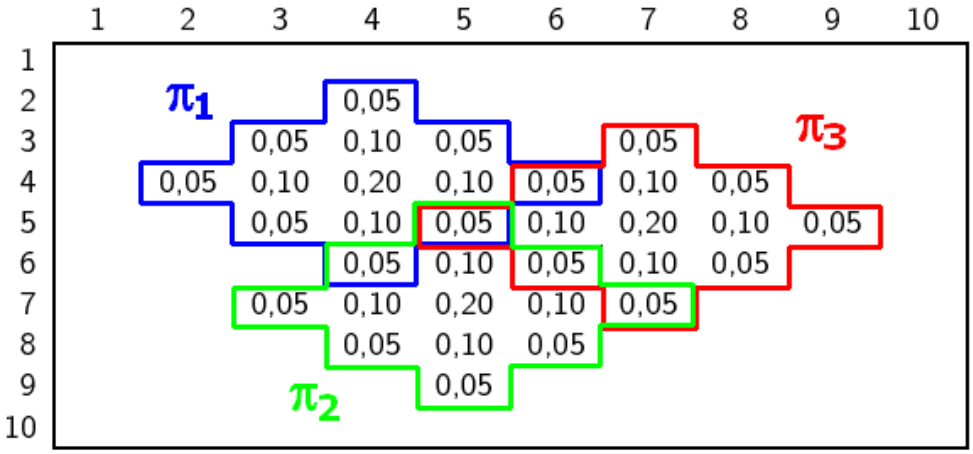

Figura 6.1: Distribuição de probabilidade de ocorrência dos comportamentos para cada uma das três políticas. As linhas representam o valor do atributo 1 e as colunas representam o valor do atributo 2.

A tabela 6.1 e a tabela 6.2 representam as funções utilidade consideradas para o usuário: estruturada u_{Est} e sem estrutura u_{Des}, respectivamente. Note que a função utilidade estruturada pode ser representada analiticamente pela função $u(\boldsymbol{\mu}) = \langle [4\ 6], \boldsymbol{\mu} \rangle$, enquanto a função utilidade sem estrutura não apresenta nenhuma forma analítica simples.

Tabela 6.1: Função utilidade estruturada u_{Est}. As linhas representam o valor do atributo 1 e as colunas representam o valor do atributo 2.

	1	2	3	4	5	6	7	8	9	10
1	10	16	22	28	34	40	46	52	58	64
2	14	20	24	32	38	44	50	56	62	68
3	18	24	30	36	42	48	54	60	66	72
4	22	28	34	40	46	52	58	64	70	76
5	26	32	38	44	50	56	62	68	74	80
6	30	36	42	48	54	60	66	72	78	84
7	34	40	46	52	58	64	70	76	82	88
8	38	44	50	56	62	68	74	80	86	92
9	42	48	54	60	66	72	78	84	90	96
10	46	52	58	64	70	76	82	88	94	100

Tabela 6.2: Função utilidade sem estrutura u_{Des}. As linhas representam o valor do atributo 1 e as colunas representam o valor do atributo 2.

	1	2	3	4	5	6	7	8	9	10
1	58	21	42	21	68	45	61	8	12	23
2	42	38	30	41	21	4	2	45	45	24
3	52	78	13	32	68	3	2	44	72	05
4	33	31	2	96	63	87	19	85	89	8
5	43	46	77	73	27	11	59	95	84	64
6	23	57	97	64	21	38	6	68	25	19
7	58	79	99	74	61	68	37	70	87	84
8	76	6	79	27	63	9	63	73	23	17
9	53	60	44	44	37	4	72	48	80	17
10	64	5	50	93	57	61	69	55	91	99

Política executada pelo Agente

Um dos aspectos analisados nos experimentos realizados foi a probabilidade $p_{\min}^{\{\Psi|\Pi\}}$ imposta às políticas usadas para obter informações sobre as preferências do usuário. A política executada pelo agente durante o processo de EPCO é uma combinação estocástica entre as políticas π_1, π_2 e π_3. Tendo como base a política π_1, ela é executada com uma probabilidade P_1, enquanto as probabilidades de execução para π_2 e π_3 são iguais e possuem valor $P_2 = P_3 = \frac{1-P_1}{2}$.

Ao variar o valor de P_1, pode-se obter diversos valores para $p_{\min}^{\{\Psi|\Pi\}}$. Os valores utilizados nos experimentos para $p_{\min}^{\{\Psi|\Pi\}}$ foram: 0,001; 0,006; 0,011; e 0,016. Mesmo que as políticas fossem executadas com probabilidades iguais $P_1 = P_2 = P_3 = \frac{1}{3}$ e desconsiderando os comportamentos que não são demonstrados por nenhuma política, o maior valor obtido para $p_{\min}^{\{\Psi|\Pi\}}$ é $\frac{0{,}05}{3}$=0,017.

Percepção do Agente

Um segundo aspecto a ser analisado nos experimentos é o efeito de variações das observações do agente tendo como base a observação do usuário. As observações do usuário são consideradas deterministas e completas, enquanto as observações do agente foram consideradas com uma variação em torno da observação do usuário. Essa variação foi modelada por distribuições normais e independentes para cada atributo. Os valores possíveis de observações são novamente discretos e compreendidos entre 1 e 10. As variações da observação do agente foram modeladas por distribuições normais com desvios padrões arbitrários σ e normalizadas pelas restrições de limites e discretização. Na tabela 6.3 é exibida a probabilidade de observação de um valor de

atributo (colunas) dada a observação do usuário de cada valor para o mesmo atributo (linhas) para $\sigma = 2$. Mesmo que nem todos os possíveis comportamentos possam ocorrer no ambiente, segundo as observações do agente esses comportamentos podem vir a ocorrer.

Tabela 6.3: Observação do agente. Probabilidade de observação de um valor de atributo (colunas) dada a observação do usuário de um valor (linhas) para o mesmo atributo e $\sigma = 2$.

	1	2	3	4	5	6	7	8	9	10
1	0,333	0,294	0,202	0,108	0,045	0,015	0,004	0,001	0,000	0,000
2	0,227	0,257	0,227	0,156	0,084	0,035	0,011	0,003	0,001	0,000
3	0,135	0,196	0,223	0,196	0,135	0,072	0,030	0,010	0,003	0,001
4	0,067	0,126	0,183	0,208	0,183	0,126	0,067	0,028	0,009	0,002
5	0,027	0,066	0,122	0,179	0,202	0,179	0,123	0,066	0,027	0,009
6	0,009	0,027	0,066	0,123	0,179	0,202	0,179	0,123	0,066	0,027
7	0,002	0,009	0,028	0,067	0,126	0,183	0,208	0,183	0,126	0,067
8	0,001	0,003	0,010	0,030	0,072	0,135	0,196	0,223	0,196	0,135
9	0,000	0,001	0,003	0,011	0,035	0,084	0,156	0,227	0,257	0,227
10	0,000	0,000	0,001	0,004	0,015	0,045	0,108	0,202	0,294	0,333

Construindo a Função Utilidade

No lugar de simular o processo de EP, onde o agente demonstra comportamentos e recebe avaliações sobre eles, optou-se por calcular a função utilidade do agente analiticamente por $u_{\mathrm{Ag}}(\boldsymbol{\mu}_{\mathrm{Ag}}) = \sum_{\boldsymbol{\mu}_{\mathrm{Us}}} u_{\mathrm{Us}}(\boldsymbol{\mu}_{\mathrm{Us}}) \Pr(\boldsymbol{\mu}_{\mathrm{Us}}|\boldsymbol{\mu}_{\mathrm{Ag}}, \pi)$, onde π é a política executada para obter a informação sobre as preferências do usuário – uma combinação estocástica entre as políticas π_1, π_2 e π_3 –, e ver o efeito de tal função utilidade sobre as três políticas possíveis – as políticas π_1, π_2 e π_3 analisadas isoladamente. Esse cálculo foi feito para diversos valores dos dois aspectos analisados: a probabilidade $p_{\min}^{\{\Psi|\Pi\}}$ e variações sobre a observação do agente. O uso dessa função analítica, no lugar de uma inferida através de interações entre agente e usuário, visa isolar dificuldades do aprendizado, de modo que aqui possam ser demonstrados apenas os efeitos dos dois aspectos analisados.

6.3.3 Resultados

Na montagem experimental aqui utilizada, pode-se calcular os valores $V_u^{\pi_1}$ atribuídos a cada uma das três políticas (π_1,π_2 e π_3) por cada uma das funções utilidade (u_{Est} e u_{Des}). Quando o usuário possui preferências estruturadas, ele avalia as políticas com os seguintes valores $V_{u_{\mathrm{Est}}}^{\pi_1} = 40$, $V_{u_{\mathrm{Est}}}^{\pi_2} = 58$ e $V_{u_{\mathrm{Est}}}^{\pi_3} = 62$, e para tal usuário a política

π_3 é a política ótima. Quando o usuário possui preferências sem estrutura, ele avalia as políticas com os seguintes valores $V^{\pi_1}_{u_{\mathrm{Des}}}$ =56,60, $V^{\pi_2}_{u_{\mathrm{Des}}}$ =51,70 e $V^{\pi_3}_{u_{\mathrm{Des}}}$ =46,30, e para tal usuário a política π_1 é a política ótima.

O objetivo da EPCO é obter informações junto ao usuário de modo a garantir que a decisão tomada pelo agente represente as preferências do usuário. Para verificar o previsto na equação (6.2), alguns experimentos foram realizados variando a observação do agente e a aleatoriedade da política executada para obter informações junto ao usuário. O desvio padrão σ_{Ag} do agente, utilizado para definir as probabilidades de observação do agente conforme a tabela 6.3, foi variado entre 0,1 e 2,0. Considerando a política π_1 como política base e combinando-a com as políticas π_2 e π_3, determinou-se quatro valores para $p^{\{\Psi|\Pi\}}_{\mathrm{min}}$: 0,001; 0,006; 0,011; e 0,016. Os resultados são exibidos na figura 6.2 para os dois tipos de usuários: estruturado (u_{Est}) e sem estrutura (u_{Des}).

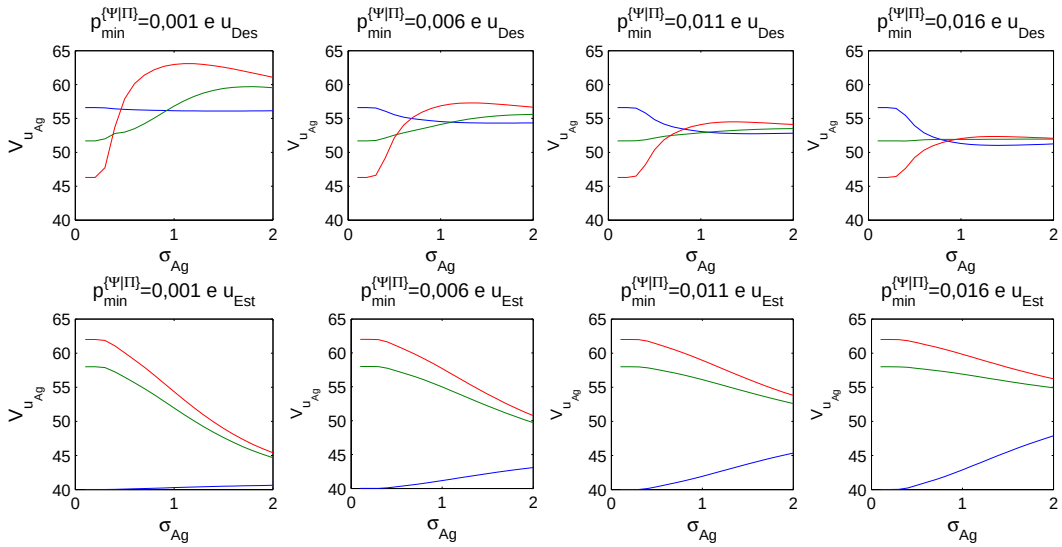

Figura 6.2: Valores encontrados pelo agente para as três possíveis políticas em diversas condições (eixo vertical). Desvio padrão da observação do agente σ_{Ag} variando entre 0,1 e 2,0 (eixo horizontal). Usuário modelado pelas funções utilidade u_{Est} (linha inferior) e u_{Des} (linha superior). Probabilidade mínima de execução de um comportamento factível $p^{\{\Psi|\Pi\}}_{\mathrm{min}}$ com os valores 0,001; 0,006; 0,011; e 0,016 e utilizando π_1 como base. A política π_1 é demonstrada em azul, a política π_2 é demonstrada em verde e a política π_3 é demonstrada em vermelho.

Como previsto pela equação (6.2), quando o desvio padrão da observação do agente é baixo ($\sigma_{\mathrm{Ag}} < 0,2$), os valores das políticas inferidos pelo agente estão próximos dos valores atribuídos pelo usuário. No entanto, conforme aumenta-se o desvio padrão da observação do agente, os valores atribuídos pelo agente e pelo usuário divergem.

Com o desvio padrão da observação do agente alto, pode-se considerar apenas a

ordenação de preferência das políticas no lugar de considerar seus valores absolutos. Sob esse aspecto, as funções utilidade u_{Est} e u_{Des} possuem características bem distintas. No caso da função estruturada u_{Est} (linha inferior), a ordenação das políticas segundo seus valores mantém-se a mesma para agente e usuário ($\pi_3 \succ \pi_2 \succ \pi_1$), ainda que, com um desvio padrão σ_{Ag} maior, os valores segundo as avaliações do agente e do usuário tendem a convergir para um mesmo valor.

No caso da função utilidade sem estrutura u_{Des}, as ordenações das políticas do agente e do usuário mantém-se iguais apenas quando o desvio padrão da observação do agente σ_{Ag} é baixo ($\sigma_{\mathrm{Ag}} < 0,2$). No entanto, essa ordenação pode ser prolongada para valores de $\sigma_{\mathrm{Ag}} > 0,2$ se o valor de $p_{\min}^{\{\Psi|\Pi\}}$ é aumentado. No caso limite, onde a política utilizada é aleatória e $p_{\min}^{\{\Psi|\Pi\}} = 0,016$, obtém-se a mesma ordenação mesmo para $\sigma_{\mathrm{Ag}} = 0,7$.

6.4 Considerações Finais

Os gráficos obtidos com os experimentos confirmam que para garantir uma decisão adequada atendendo as expectativas do usuário deve haver um compromisso entre as variações na observação do agente, que deve ser baixa; e a aleatoriedade na execução da política, que deve ser alta. No entanto, por limitação dos cenários normalmente encontrados (ambiente aleatórios e grande quantidade de comportamentos factíveis), esse segundo aspecto nem sempre pode ser alcançado com sucesso. Primeiro, se apenas políticas com alta aleatoriedade forem consideradas como políticas candidatas, embora seja possível ao agente aprender avaliá-las em consistência com o usuário, políticas realmente boas e desejáveis pelo usuário serão descartadas. Segundo, se a quantidade de comportamentos possíveis for muito grande, mesmo uma distribuição uniforme entre eles já pode ser muito baixa. Terceiro, por limitação da dinâmica de um ambiente aleatório, alguns comportamentos serão muito difíceis de serem obtidos, impossibilitando que uma alta aleatoriedade seja obtida. No entanto, apesar das dificuldades para garantir as propriedades ideais para que a inferência possa ser realizada pelo agente, quando o usuário possui preferências lineares, é possível que o agente tome uma decisão adequada atendendo as expectativas do usuário.

Uma importante discussão na EP nas áreas não relacionadas a computação é a relação entre a intenção da questão proposta por um agente, a interpretação de tal questão pelo usuário e, ainda, a representatividade das preferências do usuário em suas respostas (LUCE; WINTERFELDT, 1994; STARMER, 2000). Ao considerar a EPCO e um modelo formal para o ambiente, ao menos a relação de intenção e interpretação pode ser colocada à vista de uma análise mais formal. Uma solução para este problema

normalmente é considerada sob a forma de calibração das respostas do usuário, isto é, baseada em algum estudo antecipado a respeito de como uma questão é interpretada pelo usuário; considera-se um viés em suas respostas e retira-se tal viés ao fazer inferências sobre as preferências do usuário.

Na EPCO, uma questão formulada não é posta diretamente ao usuário, mas o usuário interpreta (comportamento observado pelo usuário) uma interpretação da questão (política executada) pelo ambiente (comportamento de fato no ambiente). Quando o usuário possui acesso a modelos de tais interpretações (ambiente e usuário, como na seção 6.1), este modelo pode ser utilizado para inferir de forma adequada as preferências do usuário, de certa forma, calibrando as respostas do usuário. No entanto, quando tais modelos são desconhecidos, a normalização das respostas do usuário não é mais possível. No modelo de ambiente utilizado nesta tese, a interpretação da questão feita pelo ambiente pode apenas ser estimada pelo agente por meio de suas observações e, nessas observações, propriedades adequadas, como as descritas neste capítulo, devem aparecer para que a EPCO seja possível.

Quando propriedades adequadas são encontradas no cenário onde deseja-se realizar a EPCO, o fato de que uma função utilidade $u_{\mathrm{Ag}}(\boldsymbol{\mu}_{\mathrm{Ag}})$ pode ser definida adequadamente com base apenas nas avaliações $u_{\mathrm{Us}}(\boldsymbol{\mu}_{\mathrm{Us}})$ dado que o agente observou $\boldsymbol{\mu}_{\mathrm{Ag}}$, independente da política, permite que tal função seja definida por outros tipos de questões, como a comparação entre comportamentos. Essa característica será utilizada no próximo capítulo para tratar o problema de EPCO de forma completa, especificando todas as partes do algoritmo 1.

7 FORMULANDO QUESTÕES E DEMONSTRANDO COMPORTAMENTOS

Para a formulação de uma questão que deve ser demonstrada por meio de comportamentos, resultantes de ações do agente aplicadas ao ambiente, um problema que deve ser analisado é como exibir tais comportamentos. A demonstração de um comportamento depende da política de ação do agente e da dinâmica do ambiente, e, assim, é necessário que a questão seja formulada em termos de políticas de ação para o agente. Então, ao formular questões, o agente deve considerar as probabilidades $\Pr(\psi|\pi)$ de ocorrência de cada comportamento $\psi \in \Psi$ resultantes da política π.

No capítulo 5 foram considerados dois conjuntos de políticas, o conjunto de políticas deterministas Π_{Det} e o conjunto de políticas estocásticas Π_{Est}. Quando se considera os vetores de atributos esperados associados a tais conjuntos de políticas, demonstrou-se que tais vetores formam um conjunto convexo e, além disso, restrito por inequações lineares baseadas em um conjunto de políticas deterministas e não dominadas Π_{NonDom}. Dadas essas propriedades, este conjunto pode ser um bom candidato para formular as questões, isto é, escolher políticas em Π_{NonDom}, que serão utilizadas pelo agente para demonstrar comportamentos.

Devido ao ambiente não ser determinista, ao escolher uma política e executá-la para demonstrar um comportamento, pode acontecer que o comportamento demonstrado não seja informativo sobre as preferências do usuário. Uma política é escolhida com base nos seus comportamentos esperados, mas pode ocorrer que o comportamento resultante esteja longe desse valor esperado e a resposta do usuário sobre tal comportamento não traga nenhuma informação nova sobre suas preferências. Uma opção para tal problema é replanejar a política a ser utilizada pelo agente após cada ação executada no ambiente, isto é, baseando-se no vetor de atributos observados até o momento, pode-se somá-lo ao vetor de atributos esperados daquele momento em diante para uma dada política. Dessa forma, pode-se contornar o acaso da dinâmica do ambiente para obter comportamentos que resultam em informações úteis sobre as preferências do usuário.

Essas duas abordagens, políticas fixas e replanejamento de políticas, serão utili-

zadas nas próximas seções para formular questões informativas[1] e demonstrá-las ao usuário.

7.1 Formulação de Questões

Usualmente, na EP, comportamentos são utilizados como elementos ao formular uma questão. Na EPCO, a formulação de questões consiste em escolher políticas de ações para que o agente possa demonstrar os comportamentos que serão alvo da análise do usuário.

Na EP, considerando o critério do arrependimento esperado, uma questão q deve ser avaliada por

$$V^q = -\sum_{r \in \mathcal{R}_q} \min_{d \in \mathcal{D}} \sum_{u \in \mathcal{U}} \mathrm{Regret}(d, u)\, \Pr(r|q, u)\, \Pr(u).$$

Entretanto, na EPCO modelada como um PMD, as decisões $d \in \mathcal{D}$ devem ser trocadas pelas políticas $\pi \in \Pi$. Além disso, uma questão q é formulada utilizando duas políticas $\pi_q^1, \pi_q^2 \in \Pi$. Quando existe linearidade nas preferências do usuário e as observações do usuário podem ser obtidas de forma aproximada através das observações do agente, então o modelo de inferência é dado por $\Pr(r|\boldsymbol{\mu}_{\mathrm{Ag}}^1, \boldsymbol{\mu}_{\mathrm{Ag}}^2, u)$, onde $\boldsymbol{\mu}_{\mathrm{Ag}}^1$ e $\boldsymbol{\mu}_{\mathrm{Ag}}^2$ são vetores de atributos observados pelo agente ao executar as políticas π_q^1 e π_q^2, respectivamente. Dessa forma, ao avaliar uma questão, deve-se considerar todos os possíveis desdobramento das políticas consideradas, logo, na EPCO, tem-se que:

$$V^q = -\sum_{r \in \mathcal{R}_q} \min_{\pi \in \Pi_{\mathrm{NonDom}}} \sum_{u \in \mathcal{U}} \mathrm{Regret}(\pi, u)\, \Pr(r|\pi_q^1, \pi_q^2, u)\, \Pr(u),$$

onde

$$\Pr(r|\pi_q^1, \pi_q^2, u) = \sum_{\boldsymbol{\mu}_{\mathrm{Ag}}^1, \boldsymbol{\mu}_{\mathrm{Ag}}^2} \Pr(r|\boldsymbol{\mu}_{\mathrm{Ag}}^1, \boldsymbol{\mu}_{\mathrm{Ag}}^2, u)\, \Pr(\boldsymbol{\mu}_{\mathrm{Ag}}^1|\pi_q^1)\, \Pr(\boldsymbol{\mu}_{\mathrm{Ag}}^2|\pi_q^2).$$

7.1.1 Modelo de Respostas do Usuário

Na EPCO, ao formular questões, calcular V^q para todas as questões factíveis é onde reside o maior custo computacional, sendo o cálculo da distribuição de probabilidades $\Pr(r|\pi_q^1, \pi_q^2, u)$ essencial nesse cálculo.

[1] Métodos informativos são aqueles que consideram conhecimentos *a priori* e conhecimentos já obtidos para formular questões que otimizem, sob alguma restrição, a esperança do novo conhecimento após a consideração da resposta do usuário.

Se uma função utilidade u pertence à classe de funções lineares, ela pode ser modelada por um vetor recompensa $\mathbf{w}_u$ e um desvio padrão σ_u que permite modelar as respostas do usuário segundo uma distribuição normal. Dessa forma, pode-se calcular para a resposta `melhor` ($\boldsymbol{\mu}^1_{\mathrm{Ag}} \succ \boldsymbol{\mu}^2_{\mathrm{Ag}}$):

$$\Pr(\texttt{melhor}|\boldsymbol{\mu}^1_{\mathrm{Ag}}, \boldsymbol{\mu}^2_{\mathrm{Ag}}, u) = \Pr(x > 0 | x = N(\langle \mathbf{w}_u, \boldsymbol{\mu}^1_{\mathrm{Ag}} - \boldsymbol{\mu}^2_{\mathrm{Ag}} \rangle, \sigma_u)),$$

onde $N(\hat{x}, \sigma)$ indica uma distribuição normal com valor médio $\hat{x}$ e desvio padrão σ. Para a resposta `pior` ($\boldsymbol{\mu}^1_{\mathrm{Ag}} \prec \boldsymbol{\mu}^2_{\mathrm{Ag}}$) tem-se:

$$\Pr(\texttt{pior}|\boldsymbol{\mu}^1_{\mathrm{Ag}}, \boldsymbol{\mu}^2_{\mathrm{Ag}}, u) = 1 - \Pr(\texttt{melhor}|\boldsymbol{\mu}^1_{\mathrm{Ag}}, \boldsymbol{\mu}^2_{\mathrm{Ag}}, u).$$

Como visto na seção 5.1, uma política π pode ser parcialmente representada por um vetor de atributos esperados $\bar{\boldsymbol{\mu}}^\pi$ e uma matriz de covariância Σ^π. Dado um vetor de recompensas $\mathbf{w}$, como $V^\pi = \langle \mathbf{w}, \bar{\boldsymbol{\mu}}^\pi \rangle$, tem-se que $\sigma_{V^\pi} = \mathbf{w}^\intercal \Sigma^\pi \mathbf{w}$. Dessa forma, considerando variações da função utilidade e de resposta do ambiente como variáveis aleatórias com distribuições normais, pode-se aproximar $\Pr(\texttt{melhor}|\pi^1_q, \pi^2_q, u)$ por:

$$\widehat{P}^{\texttt{melhor},\pi^1_q,\pi^2_q,u} = \Pr(x > 0 | x = N(\langle \mathbf{w}_u, \bar{\boldsymbol{\mu}}^{\pi^1_q}_{\mathrm{Ag}} - \bar{\boldsymbol{\mu}}^{\pi^2_q}_{\mathrm{Ag}} \rangle, \sigma_u + \mathbf{w}_u^\intercal [\Sigma^{\pi^1_q} + \Sigma^{\pi^2_q}] \mathbf{w}_u). \quad (7.1)$$

Uma outra opção é utilizar simulação de Monte Carlo para obter uma aproximação para $\Pr(\texttt{melhor}|\pi^1_q, \pi^2_q, u)$. Na simulação de Monte Carlo, simula-se uma amostragem dos comportamentos demonstrados pelo agente ao executar as políticas π^1_q e π^2_q para estimar $\Pr(r|\pi^1_q, \pi^2_q, u)$. Então, a qualidade de aproximação depende da quantidade de amostras utilizada. Dada uma seqüência de n_1 amostras de vetores de atributos $\boldsymbol{\mu}^{\pi^1_q}_1, \boldsymbol{\mu}^{\pi^1_q}_2, \ldots, \boldsymbol{\mu}^{\pi^1_q}_{n_1}$ obtidos da distribuição $\Pr(\boldsymbol{\mu}_{\mathrm{Ag}}|\pi^1_q)$ e uma seqüência de n_2 amostras de vetores de atributos $\boldsymbol{\mu}^{\pi^2_q}_1, \boldsymbol{\mu}^{\pi^2_q}_2, \ldots, \boldsymbol{\mu}^{\pi^2_q}_{n_2}$ obtidos da distribuição $\Pr(\boldsymbol{\mu}_{\mathrm{Ag}}|\pi^2_q)$, a distribuição de probabilidades $\Pr(\texttt{melhor}|\pi^1_q, \pi^2_q, u)$ pode ser aproximada por:

$$\begin{aligned} \widehat{P}^{\texttt{melhor},\pi^1_q,\pi^2_q,u} &= \tfrac{1}{n_1 n_2} \textstyle\sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \Pr(\texttt{melhor}|\boldsymbol{\mu}^{\pi^1_q}_i, \boldsymbol{\mu}^{\pi^2_q}_j, u) \\ &= \tfrac{1}{n_1 n_2} \textstyle\sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \Pr(x > 0 | x = N(\langle \mathbf{w}_u, \boldsymbol{\mu}^{\pi^1_q}_i - \boldsymbol{\mu}^{\pi^2_q}_j \rangle, \sigma_u)). \end{aligned} \quad (7.2)$$

7.1.2 Conjunto de Questões Restrito

Ao formular uma questão, um dos objetivos a alcançar é diminuir a entropia da distribuição de probabilidades $\Pr(u)$. Mas, se $\Pr(r|\pi^1_q, \pi^2_q, u) \approx 0{,}5$, a redução da entropia será baixa, já que as probabilidades $\Pr(u)$ continuariam com valores parecidos. Essa propriedade de redução de entropia ajuda também na escolha de um conjunto restrito de questões $\mathcal{Q}_{\mathrm{Rest}}$ que devem ser consideradas ao formular uma

questão. Considere duas políticas $\pi', \pi'' \in \Pi_{\text{NonDom}}$ e duas políticas $\pi^\alpha, \pi^\beta \in \Pi_{\text{Est}}$, tal que $\bar{\boldsymbol{\mu}}^{\pi^\alpha} = \alpha\bar{\boldsymbol{\mu}}^{\pi'} + (1-\alpha)\bar{\boldsymbol{\mu}}^{\pi''}$ e $\bar{\boldsymbol{\mu}}^{\pi^\beta} = \beta\bar{\boldsymbol{\mu}}^{\pi'} + (1-\beta)\bar{\boldsymbol{\mu}}^{\pi''}$. Se o ambiente for determinista, tem-se que a entropia resultante de $\Pr(r|\pi', \pi'', u)$ é menor que a entropia resultante de $\Pr(r|\pi^\alpha, \pi^\beta, u)$, pois $\bar{\boldsymbol{\mu}}^{\pi^\alpha}$ e $\bar{\boldsymbol{\mu}}^{\pi^\beta}$ estariam mais próximos um do outro, sendo mais facilmente confundidos pelo usuário.

Como foi visto na seção 5.2, o conjunto com todos os vetores de atributos possíveis pode ser representados por um poliedro. Se o ambiente fosse determinista, as políticas com vetores de atributos esperados internos a tal poliedro poderiam ser descartadas ao formular questões, pois sempre existiriam versões melhoradas de tais questões com políticas nas bordas do poliedro. Vale também dizer que, embora algumas considerações possam ser feitas ao formular as questões, essas considerações não interferem na inferência realizada a partir das respostas do usuário, mas interferem na taxa com que uma função utilidade $u = u_{\text{Us}}$ se destaca das outras funções utilidade candidatas no decorrer da EPCO.

Dessa forma, pode-se restringir a formulação de questões não às políticas com vetores de atributos esperados nas bordas do poliedro, mas apenas aos vértices do mesmo. Então, a EPCO é feita considerando o conjunto de questões $\mathcal{Q}_{\text{Rest}} = \Pi_{\text{NonDom}} \times \Pi_{\text{NonDom}}$ e, dada uma distribuição de probabilidades $\Pr(u)$, escolhe-se a questão q^* segundo:

$$q^* = \arg \min_{q \in \mathcal{Q}_{\text{Rest}}} \sum_{r \in \mathcal{R}_q} \min_{\pi \in \Pi_{\text{NonDom}}} \sum_{u \in \mathcal{U}} \text{Regret}(\pi, u) \widehat{P}^{r, \pi_q^1, \pi_q^2, u} \Pr(u),$$

onde $\widehat{P}^{r, \pi_q^1, \pi_q^2, u}$ é uma estimativa da probabilidade $\Pr(r|\pi_q^1, \pi_q^2, u)$, seja com base em vetores de atributos esperados e matrizes de covariância ou com base em simulação de Monte Carlo.

7.2 Replanejamento de Políticas ao Demonstrar Comportamentos

Note que, ao formular questões, um conjunto de políticas é considerado para determinar as questões possíveis de serem formuladas. Na seção anterior, o conjunto de políticas não dominadas Π_{NonDom} foi utilizado como conjunto base para a formulação de questões. Essas são políticas fixas, pois são utilizadas ao longo de um período completo, independente do ocorrido na demonstração do comportamento. O espaço de estados considerado por uma política fixa leva em conta apenas o estado atual do ambiente e por quanto tempo o comportamento já foi demonstrado no período atual.

Apesar da função utilidade ser um dos elementos de um problema de decisão, o problema de EPCO também pode ser visto como um problema de decisão. A cada estado observado, deve-se tomar a decisão de qual ação executar para exibir um comportamento que retorne informação útil após ser avaliado. Portanto, na EPCO, também pode ser interessante considerar um estado estendido que reflita: o nível de conhecimento corrente do agente e o comportamento exibido até o momento.

O vetor de estado que representa o estado estendido é definido como $\mathbf{x} = (s, t^{\Gamma}, \boldsymbol{\mu}^1, \boldsymbol{\mu}^2, \Pr^{\mathcal{U}}, \Gamma)$, onde s representa o estado do ambiente, t^{Γ} o tempo no período atual sendo demonstrado, $\boldsymbol{\mu}^1$ os atributos acumulados para o primeiro comportamento da questão, $\boldsymbol{\mu}^2$ os atributos acumulados para o segundo comportamento da questão, $\Pr^{\mathcal{U}}$ a crença atual da distribuição de probabilidades no espaço de funções utilidade candidatas $\mathcal{U}$, e $\Gamma \in \{1, 2, \texttt{resp}\}$ indica qual comportamento está sendo executado no momento, o primeiro, o segundo ou se uma resposta do usuário é esperada.

O estado inicial é dado pela distribuição inicial de s_0, que é determinada pelo ambiente, e a distribuição $\Pr_0^{\mathcal{U}}$ é determinada por estágios antecessores do processo de EP. Os vetores de atributos são iniciados com valores nulos $\boldsymbol{\mu}_0^1 = \boldsymbol{\mu}_0^2 = [0\ 0\ \cdots\ 0]$, $t^{\Gamma} = 0$ e $\Gamma_0 = 1$ indica que o primeiro comportamento será executado. Portanto, $\mathbf{x}_0 = (s_0, 0, \mathbf{0}, \mathbf{0}, \Pr_0^{\mathcal{U}_{\text{Ag}}}, 1)$.

Uma recompensa só é recebida após a exposição completa da questão (quando os dois comportamentos foram finalizados) e a aquisição da resposta do usuário. As transições entre os estados estendidos e as recompensas recebidas ocorrem da seguinte maneira, para qualquer tempo global t_G e considerando que um episódio possui N passos:

- se $\Gamma_{t_G} = 1$ (se o primeiro comportamento está sendo demonstrado)
  - se $t_{t_G}^{\Gamma} < N$ (se ainda está demonstrando o primeiro comportamento)
    - $\mathbf{x}_{t_G+1} \leftarrow (s_{t_G+1}, t_{t_G}^{\Gamma} + 1, \boldsymbol{\mu}_{t_G}^1 + \boldsymbol{\phi}(s_{t_G}, a_{t_G}), \boldsymbol{\mu}_{t_G}^2, \Pr_{t_G}^{\mathcal{U}}, \Gamma_{t_G})$
  - se $t_{t_G}^{\Gamma} = N$ (se já finalizou a demonstração do primeiro comportamento)
    - $\mathbf{x}_{t_G+1} \leftarrow (s_0, 0, \boldsymbol{\mu}_{t_G}^1 + \boldsymbol{\phi}(s_{t_G}, a_{t_G}), \boldsymbol{\mu}_{t_G}^2, \Pr_{t_G}^{\mathcal{U}}, 2)$
- se $\Gamma_{t_G} = 2$ (se o segundo comportamento está sendo demonstrado)
  - se $t_{t_G}^{\Gamma} < N$ (se ainda está demonstrando o segundo comportamento)
    - $\mathbf{x}_{t_G+1} \leftarrow (s_{t_G+1}, t_{t_G}^{\Gamma} + 1, \boldsymbol{\mu}_{t_G}^1, \boldsymbol{\mu}_{t_G}^2 + \boldsymbol{\phi}(s_{t_G}, a_{t_G}), \Pr_{t_G}^{\mathcal{U}}, \Gamma_{t_G})$
  - se $t_{t_G}^{\Gamma} = N$ (se já finalizou a demonstração do segundo comportamento)
    - $\mathbf{x}_{t_G+1} \leftarrow (s_{t_G+1}, 0, \boldsymbol{\mu}_{t_G}^1, \boldsymbol{\mu}_{t_G}^2 + \boldsymbol{\phi}(s_{t_G}, a_{t_G}), \Pr_{t_G}^{\mathcal{U}}, \texttt{resp})$

- se $\Gamma_{t_G} = \texttt{resp}$ (se os dois comportamentos foram demonstrados)
  - considerando a questão $q \leftarrow (\boldsymbol{\mu}^1_{t_G} \succ \boldsymbol{\mu}^2_{t_G})$, calcula como recompensa o valor V^q definido por:

$$V^q = -\sum_{r \in \mathcal{R}_q} \min_{\pi \in \Pi_{\text{NonDom}}} \sum_{u \in \mathcal{U}} \text{Regret}(\pi, u) \Pr(r|q, u) \Pr^{\mathcal{U}}_{t_G}(u)$$

Considerar o problema dessa forma pode ser muito custoso computacionalmente. Primeiro, deve-se encontrar uma política estendida para todo o espaço de possíveis distribuições de probabilidades $\Pr(u)$. Mas, apenas algumas dessas distribuições ocorrem ao longo de um processo de EPCO. Dessa forma, seria mais interessante realizar este planejamento apenas para as distribuições de probabilidades $\Pr(u)$ ocorridas no processo, realizando o planejamento para cada questão a ser formulada.

Segundo, outra variável que aumenta exponencialmente o espaço de estados do PMD estendido é o vetor de atributos acumulados. Novamente, apenas alguns dos possíveis valores acumulados ocorrem na demonstração de um comportamento. Mas, diferentemente do que ocorre para as distribuições $\Pr(u)$, todos os valores possíveis de ocorrências no futuro devem ser contemplados.

Uma opção aproximada para contornar este problema é basear as formulações das questões nas políticas não dominadas Π_{NonDom}. Ao formular uma questão pode-se considerar apenas o conjunto de políticas Π_{NonDom}, mas durante a demonstração do comportamento, pode-se reformular a questão escolhendo novas políticas no conjunto Π_{NonDom}. Note que a formulação de questões não considera que, durante a demonstração do comportamento, a política pode ser alterada, mas apenas o comportamento parcial observado até o momento da reformulação.

Dado um estado qualquer s_{t_G}, uma política π pode ser avaliada pelo vetor de atributos esperados $\bar{\boldsymbol{\mu}}^{\pi}_{\text{Ag}}(s_{t_G}, t_{t_G})$ e o vetor de atributos acumulado $\boldsymbol{\mu}_{\text{Ag}}(t_{t_G})$, onde $\bar{\boldsymbol{\mu}}^{\pi}_{\text{Ag}}(s_{t_G}, t_{t_G})$ indica o vetor de atributos esperados a partir do estado s_{t_G} no tempo t_{t_G}, executando a política π, e $\boldsymbol{\mu}_{\text{Ag}}(t_{t_G})$ é o vetor de atributos acumulados observados pelo agente até o tempo t_{t_G}.

Ao demonstrar uma questão, dois períodos $\Gamma_1 = 0, 1, \ldots, N$ e $\Gamma_2 = 0, 1, \ldots, N$ são demonstrados pelo agente e observados pelo usuário. Durante a demonstração, o agente encontra-se no período Γ em um determinado tempo $t^{\Gamma}_{t_G}$. O vetor de atributos $\boldsymbol{\mu}_{\text{Ag}}(t^{\Gamma}_{t_G})$ indica os atributos acumulados observados pelo agente até o tempo $t^{\Gamma}_{t_G}$, isto é, os atributos já ocorridos até aquele momento e não passíveis de mudança. Por outro lado, o vetor de atributos esperados $\bar{\boldsymbol{\mu}}^{\pi^{\Gamma}}_{\text{Ag}}(s_{t_G}, t^{\Gamma}_{t_G})$ indica as expectativas para o período Γ com relação ao futuro.

Ao considerar replanejamento, a distribuição de probabilidade $\Pr(r|\pi^1, \pi^2, u)$ deve ser trocada por $\Pr(r|\pi^1, \pi^2, \mathbf{x}_{t_G}, u)$, onde $\mathbf{x}_{t_G}$ é o estado definido para o PMD estendido, pois ele representa a situação atual da questão que está sendo demonstrada. Diferentemente da seção anterior, o cálculo de $\Pr(r|\pi^1, \pi^2, \mathbf{x}_{t_G}, u)$ não pode ser feito *a priori*, pois o seu valor depende do vetor de atributos acumulados $\boldsymbol{\mu}_{\mathrm{Ag}}(t^{\Gamma}_{t_G})$. Dessa forma, essa aproximação deve ser calculada a cada tomada de decisão, dependendo do ocorrido.

Uma aproximação $\widehat{P}^{\mathtt{melhor},\pi^1_q,\pi^2_q,\mathbf{x}_{t_G},u}$ para a distribuição de probabilidades $\Pr(r|\pi^1_q, \pi^2_q, \mathbf{x}_{t_G}, u)$ pode ser calculada utilizando vetores de atributos esperados e matrizes de covariâncias por:

$$\widehat{P}^{\mathtt{melhor},\pi^1_q,\pi^2_q,\mathbf{x}_{t_G},u} = \Pr(x > 0|x = N(\langle \mathbf{w}_u, \Delta\boldsymbol{\mu}^{\pi^1_q,\pi^2_q,u}_{\mathbf{X}_{t_G}}\rangle, \sigma_u + \sigma^{\pi^1_q,\pi^2_q,u}_{\mathbf{x}_{t_G}}), \qquad (7.3)$$

onde:

$$\Delta\boldsymbol{\mu}^{\pi^1_q,\pi^2_q,u}_{\mathbf{X}_{t_G}} = \begin{cases} [\mathbf{x}_{t_G}(\boldsymbol{\mu}^1) + \bar{\boldsymbol{\mu}}^{\pi^1_q}_{\mathrm{Ag}}(\mathbf{x}_{t_G}(s), \mathbf{x}_{t_G}(t^{\Gamma}))] - \bar{\boldsymbol{\mu}}^{\pi^2_q}_{\mathrm{Ag}} & , \text{ se } \mathbf{x}_{t_G}(\Gamma) = 1 \\ \mathbf{x}_{t_G}(\boldsymbol{\mu}^1) - [\mathbf{x}_{t_G}(\boldsymbol{\mu}^2) + \bar{\boldsymbol{\mu}}^{\pi^2_q}_{\mathrm{Ag}}(\mathbf{x}_{t_G}(s), \mathbf{x}_{t_G}(t^{\Gamma}))] & , \text{ se } \mathbf{x}_{t_G}(\Gamma) = 2 \end{cases},$$

$$\sigma^{\pi^1_q,\pi^2_q,u}_{\mathbf{x}_{t_G}} = \begin{cases} \mathbf{w}_u^{\mathsf{T}}[\Sigma^{\pi^1_q}(\mathbf{x}_{t_G}(s), \mathbf{x}_{t_G}(t^{\Gamma})) + \Sigma^{\pi^2_q}]\mathbf{w}_u & , \text{ se } \mathbf{x}_{t_G}(\Gamma) = 1 \\ \mathbf{w}_u^{\mathsf{T}}\Sigma^{\pi^2_q}(\mathbf{x}_{t_G}(s), \mathbf{x}_{t_G}(t^{\Gamma}))\mathbf{w}_u & , \text{ se } \mathbf{x}_{t_G}(\Gamma) = 2 \end{cases},$$

e $\bar{\boldsymbol{\mu}}^{\pi}_{\mathrm{Ag}}(s,t)$ e $\Sigma^{\pi}(s,t)$ são obtidos como na seção 5.1.

No caso em que a aproximação $\widehat{P}^{\mathtt{melhor},\pi^1_q,\pi^2_q,\mathbf{x}_{t_G},u}$ é calculada utilizando simulação de Monte de Carlo, deve-se gerar amostras de vetores de atributos, mas considerando o estágio onde a demonstração de uma questão se encontra. Então:

- se $\mathbf{x}_{t_G}(\Gamma) = 1$ deve-se obter
  - uma seqüência de n_1 amostras de vetores de atributos $\boldsymbol{\mu}^{\pi^1_q}_1, \boldsymbol{\mu}^{\pi^1_q}_2, \ldots, \boldsymbol{\mu}^{\pi^1_q}_{n_1}$ obtidos da variável aleatória $\mathbf{x}_{t_G}(\boldsymbol{\mu}^1) + \boldsymbol{\mu}_{\mathrm{Ag}}(\mathbf{x}_{t_G}(s), \mathbf{x}_{t_G}(t^{\Gamma}), \pi^1_q)$, onde $\boldsymbol{\mu}_{\mathrm{Ag}}(s,t,\pi)$ é obtido como na seção 5.1, e
  - uma seqüência de n_2 amostras de vetores de atributos $\boldsymbol{\mu}^{\pi^2_q}_1, \boldsymbol{\mu}^{\pi^2_q}_2, \ldots, \boldsymbol{\mu}^{\pi^2_q}_{n_2}$ obtidos da distribuição $\Pr(\boldsymbol{\mu}_{\mathrm{Ag}}|\pi^2_q)$;
- se $\mathbf{x}_{t_G}(\Gamma) = 2$ deve-se obter
  - uma seqüência com apenas uma amostra ($n_1 = 1$) de vetor de atributos $\boldsymbol{\mu}^{\pi^1_q}_1 = \mathbf{x}_{t_G}(\boldsymbol{\mu}^1)$ e
  - uma seqüência de n_2 amostras de vetores de atributos $\boldsymbol{\mu}^{\pi^2_q}_1, \boldsymbol{\mu}^{\pi^2_q}_2, \ldots, \boldsymbol{\mu}^{\pi^2_q}_{n_2}$ obtidos da variável aleatória $\mathbf{x}_{t_G}(\boldsymbol{\mu}^2) + \boldsymbol{\mu}_{\mathrm{Ag}}(\mathbf{x}_{t_G}(s), \mathbf{x}_{t_G}(t^{\Gamma}), \pi^2_q)$.

A distribuição de probabilidades $\Pr(\texttt{melhor}|\pi_q^1, \pi_q^2, \mathbf{x}_{t_G}, u)$ pode ser aproximada por:

$$\widehat{P}^{\texttt{melhor},\pi_q^1,\pi_q^2,\mathbf{x}_{t_G},u} = \frac{1}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \Pr(x > 0 | x = N(\langle \mathbf{w}_u, \boldsymbol{\mu}_i^{\pi_q^1} - \boldsymbol{\mu}_j^{\pi_q^2} \rangle, \sigma_u)). \quad (7.4)$$

Como o custo de replanejamento da política pode ser muito alto, pode-se considerar a opção de realizar tal planejamento de forma mais esparsa. Em um caso extremo, o replanejamento pode ser feito apenas após demonstrar o primeiro comportamento por inteiro, devendo-se calcular $\Pr(r|\boldsymbol{\mu}_{\mathrm{Ag}}^1, \pi_q^2, u)$. No algoritmo 2 apresenta-se uma versão mais geral do algoritmo 1, onde o replanejamento pode ocorrer de forma arbitrária e tal ocorrência é definida pela função $\mathrm{Replaneja}(t^\Gamma, \Gamma)$.

7.3 Experimentos

Os objetivos dos experimentos neste capítulo é comparar as duas técnicas para formular questões aqui apresentada: utilizando políticas fixas e utilizando replanejamento. Além disso, apresenta-se a especificação completa de um agente para solucionar o problema de EPCO.

A comparação entre as diferentes técnicas de formulação de questões consiste em extrair as preferências do usuário utilizando as duas técnicas para formulação de questões e verificar o desempenho de ambas as técnicas. Também serão extraídas as preferências do usuário formulando questões aleatoriamente, de modo a verificar o ganho obtido com o custo computacional empregado na formulação de questões mais informativas.

7.3.1 Tarefa do Agente

A tarefa que se deseja programar no agente é a mesma já introduzida no capítulo 5, isto é, tarefa de se deslocar de um ponto inicial a um ponto final e ao mesmo tempo maximizar ou minimizar a ocorrência de alguns atributos de forma a satisfazer as preferências do usuário.

Para simular as dificuldades com relação às diferentes observações entre usuário e agente, apenas três atributos foram considerados para observação do agente: `distância`, `curva geral` e `colisão`. O usuário foi simulado por uma função utilidade linear com base no conjunto completo de atributos: `distância`, `tempo`, `curva geral`, `curva fechada` e `colisão`. O agente deve encontrar uma função utilidade $u_{\mathrm{Ag}}(\boldsymbol{\mu}_{\mathrm{Ag}})$ que lhe permita avaliar uma trajetória.

Algoritmo 2: Pseudo-algoritmo para EPCO com base em replanejamento utilizando o arcabouço de PMD e considerando que as preferências do usuário são lineares.

Data: Π, $\mathcal{W}$, Σ, N
define $\mathcal{U} = \{(\mathbf{w}, \sigma) | \mathbf{w} \in \mathcal{W} \wedge \sigma \in \Sigma\}$ e $\Pr(\mathbf{w}, \sigma) = \frac{1}{|\mathcal{U}|}$ para todo $(\mathbf{w}, \sigma) \in \mathcal{U}$;
define $Q = \{(\pi^1, \pi^2) | \pi^1, \pi^2 \in \Pi\}$;
repeat
  inicializa o primeiro comportamento fazendo $\boldsymbol{\mu}^1_{\text{Ag}} = \mathbf{0}$;
  for $t = 0$ *até* $N - 1$ **do**
    if $\text{Replaneja}(t, 1)$ ou $t = 0$ **then**
      cria o estado estendido $\mathbf{x} = (s_t, t, \boldsymbol{\mu}^1_{\text{Ag}}, \mathbf{0}, \Pr(\mathbf{w}, \sigma), 1)$;
      estima $\widehat{P}^{r,\pi^1,\pi^2,\mathbf{x},(\mathbf{w},\sigma)}$ para todo $(\mathbf{w}, \sigma) \in \mathcal{U}$;
      forall $q = (\pi^1_q, \pi^2_q) \in Q$ **do**

$$V^q \leftarrow -\sum_{r \in \mathcal{R}_q} \min_{\pi \in \Pi_{\text{NonDom}}} \sum_{(\mathbf{w},\sigma) \in \mathcal{U}} \text{Regret}(\pi, \mathbf{w}) \widehat{P}^{r,\pi^1_q,\pi^2_q,\mathbf{x},(\mathbf{w},\sigma)} \Pr(\mathbf{w}, \sigma)$$

      end
      escolhe $q^* = \max_{q \in Q} \text{Info}(q)$;
    end
    observa o estado do ambiente s_t e executa a ação $a_t = \pi^1_{q^*}(s_t, t)$;
    observa os atributos $\phi_{\text{Ag}}(s_t, a_t)$;
    atualiza $\boldsymbol{\mu}^1_{\text{Ag}} \leftarrow \boldsymbol{\mu}^1_{\text{Ag}} + \phi_{\text{Ag}}(s_t, a_t)$;
  end
  inicializa o segundo comportamento fazendo $\boldsymbol{\mu}^2_{\text{Ag}} = \mathbf{0}$;
  for $t = 0$ *até* $N - 1$ **do**
    if $\text{Replaneja}(t, 2)$ **then**
      cria o estado estendido $\mathbf{x} = (s_t, t, \boldsymbol{\mu}^1_{\text{Ag}}, \boldsymbol{\mu}^2_{\text{Ag}}, \Pr(\mathbf{w}, \sigma), 2)$;
      estima $\widehat{P}^{r,\pi^1,\pi^2,\mathbf{x},(\mathbf{w},\sigma)}$ para todo $(\mathbf{w}, \sigma) \in \mathcal{U}$;
      forall $q = (\pi^1_q, \pi^2_q) \in Q$ **do**

$$V^q \leftarrow -\sum_{r \in \mathcal{R}_q} \min_{\pi \in \Pi_{\text{NonDom}}} \sum_{(\mathbf{w},\sigma) \in \mathcal{U}} \text{Regret}(\pi, \mathbf{w}) \widehat{P}^{r,\pi^1_q,\pi^2_q,\mathbf{x},(\mathbf{w},\sigma)} \Pr(\mathbf{w}, \sigma)$$

      end
      escolhe $q^* = \max_{q \in Q} \text{Info}(q)$;
    end
    observa o estado do ambiente s_t e executa a ação $a_t = \pi^2_{q^*}(s_t, t)$;
    observa os atributos $\phi_{\text{Ag}}(s_t, a_t)$;
    atualiza $\boldsymbol{\mu}^2_{\text{Ag}} \leftarrow \boldsymbol{\mu}^2_{\text{Ag}} + \phi_{\text{Ag}}(s_t, a_t)$;
  end
  usuário escolhe qual comportamento foi melhor emitindo a resposta r;
  forall $(\mathbf{w}, \sigma) \in \mathcal{U}$ **do**

$$\Pr(\mathbf{w}, \sigma) \leftarrow \frac{\Pr(r | \boldsymbol{\mu}^1_{\text{Ag}}, \boldsymbol{\mu}^2_{\text{Ag}}, \mathbf{w}, \sigma) \Pr(\mathbf{w}, \sigma)}{\sum_{(\mathbf{w},\sigma) \in \mathcal{U}} \Pr(r | \boldsymbol{\mu}^1_{\text{Ag}}, \boldsymbol{\mu}^2_{\text{Ag}}, \mathbf{w}, \sigma) \Pr(\mathbf{w}, \sigma)}$$

  end
until *extração de preferências satisfatória* ;
agente retorna o vetor recompensa esperado $\mathbf{w}_{\text{E}} = \sum_{(\mathbf{w},\sigma) \in \mathcal{U}} \Pr(\mathbf{w}, \sigma) \mathbf{w}$

A função utilidade $u_{\mathrm{Ag}}(\boldsymbol{\mu}_{\mathrm{Ag}})$ deve representar as preferências do usuário e o processo de EPCO será utilizado para determinar tal função. Então, o seguinte ciclo repete-se:

- agente formula uma questão $q = (\pi_q^1, \pi_q^2)$,
- agente age no ambiente durante o período Γ^1, enquanto agente observa os atributos $\boldsymbol{\mu}_{\mathrm{Ag}}^1$ e usuário observa o comportamento ψ_{Us}^1,
- agente age no ambiente durante o período Γ^2, enquanto agente observa os atributos $\boldsymbol{\mu}_{\mathrm{Ag}}^2$ e usuário observa o comportamento ψ_{Us}^2, e
- usuário compara os comportamentos ψ_{Us}^1 e ψ_{Us}^2 e emite para o agente qual é o melhor comportamento segundo suas preferências.

Como o usuário é simulado por uma função utilidade linear, tal função utilidade pode ser representada por um vetor recompensa $\mathbf{w}_{\mathrm{Us}}$ que atribui um peso para cada um dos cinco atributos observados pelo usuário e um comportamento ψ real e completo ocorrido no ambiente é mapeado em um vetor de atributos $\boldsymbol{\mu}_{\mathrm{Us}}(\psi)$, ou seja, $\psi_{\mathrm{Us}}^1 = \boldsymbol{\mu}_{\mathrm{Us}}(\psi^1) = \boldsymbol{\mu}_{\mathrm{Us}}^1$ e $\psi_{\mathrm{Us}}^2 = \boldsymbol{\mu}_{\mathrm{Us}}(\psi^2) = \boldsymbol{\mu}_{\mathrm{Us}}^2$. A resposta que o usuário emite ao agente depende: dos atributos observados $\boldsymbol{\mu}_{\mathrm{Us}}^1$ e $\boldsymbol{\mu}_{\mathrm{Us}}^2$, do vetor de recompensa $\mathbf{w}_{\mathrm{Us}}$ e, ainda, de uma variável aleatória normal ω_{Us} com média 0 e variância σ_{Us}. Então, a probabilidade de o usuário responder `melhor`, isto é, $\psi_{\mathrm{Us}}^1 \succ \psi_{\mathrm{Us}}^2$, é dada por:

$$\Pr(r = \texttt{melhor} | \psi_{\mathrm{Us}}^1, \psi_{\mathrm{Us}}^2) = \Pr(x > 0 | x = N(\langle \mathbf{w}_{\mathrm{Us}}, \boldsymbol{\mu}_{\mathrm{Us}}^1 - \boldsymbol{\mu}_{\mathrm{Us}}^2 \rangle, \sigma_{\mathrm{Us}})).$$

7.3.2 Especificação de um Agente para EPCO

Um agente que realize a EPCO deve ter especificado alguns módulos em sua arquitetura. Primeiro, como será realizada a inferência mediante uma questão formulada, comportamentos observados e resposta emitida pelo usuário. Segundo, um conjunto de funções utilidade deve ser especificado para que inferências sobre ele possam ser utilizadas para testar aderência das respostas do usuário às preferências que tal função representa. Terceiro, uma técnica deve ser especificada para formular questões. Esses três módulos são discutidos a seguir.

Inferência: Diferentes Observações e Factibilidade da EP

Na tarefa utilizada nos experimentos, ocorre que os atributos utilizados pelo agente não são completos com relação às preferências que o agente deve representar, isto é,

as preferências do usuário. No entanto, no capítulo 6, condições foram delineadas de modo a possibilitar a EPCO mesmo quando o conjunto de atributos do agente não é completo. Utilizando os dados obtidos no capítulo 5, pode-se observar a dependência que o valor de um dado atributo $i \in \Xi_{\mathrm{Us}} - \Xi_{\mathrm{Ag}}$ possui com os atributos $j \in \Xi_{\mathrm{Ag}}$, isto é, a correlação entre os atributos não observados pelo agente e os atributos observados pelo agente. Na figura 7.1, para cada par de atributos (i, j) são exibidos os valores esperados obtidos para cada política não dominada[2]. As colunas representam os atributos que são observados pelo agente (eixo horizontal), enquanto as linhas representam os atributos não observados pelo agente (eixo vertical).

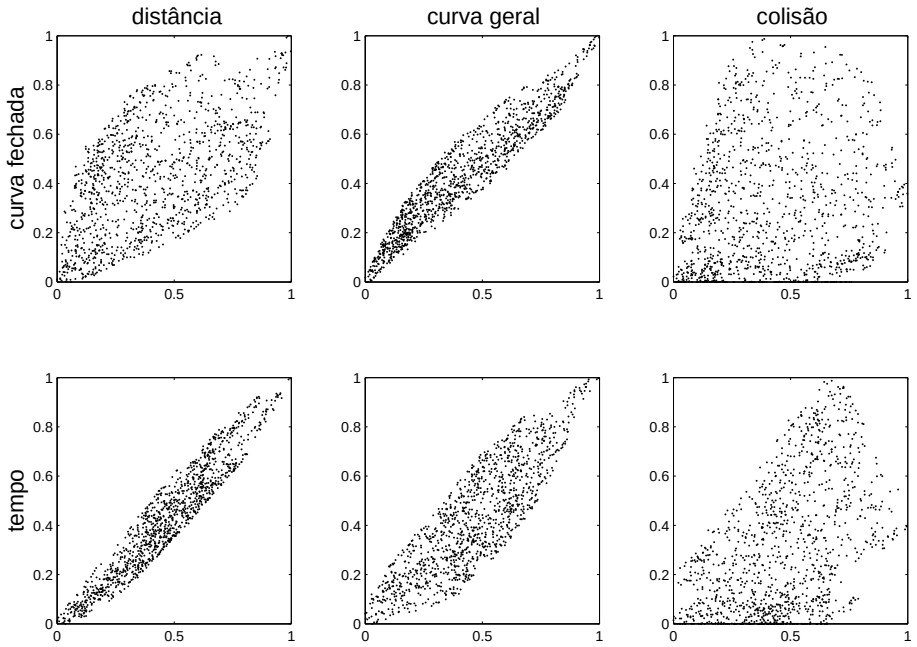

Figura 7.1: Correlação entre atributos observados e não observados pelo agente.

Para que o agente possa modelar adequadamente as preferências do usuário, os atributos considerados pelo usuário e não observados pelo agente devem possuir uma alta correlação com os atributos observados pelo agente. Nos gráficos da figura 7.1 pode-se notar que os atributos `curva fechada` e `tempo` podem ser determinados, com um erro limitado, respectivamente pelos atributos `curva geral` e `distância` e essa determinação é feita de forma linear. Com os valores dos atributos normalizados entre 0 e 1, o erro máximo para o par `curva fechada-curva geral` é 0,1853, enquanto o desvio padrão é 0,0020. Para o par `tempo-distância`, o erro máximo é 0,1598 e o desvio padrão 0,0017. Dessa forma, o vetor de atributos do agente expande-se em uma hiperesfera pequena no espaço de vetor de atributos do usuário e a probabilidade $\Pr(r|\pi^1, \pi^2, \boldsymbol{\mu}^1_{\mathrm{Ag}}, \boldsymbol{\mu}^1_{\mathrm{Ag}}, u)$ pode ser aproximada por $\Pr(r|\boldsymbol{\mu}^1_{\mathrm{Ag}}, \boldsymbol{\mu}^1_{\mathrm{Ag}}, u)$.

[2]Para construir a figura 7.1 foram utilizados os dados do capítulo 5 referentes ao cálculo das políticas não dominadas utilizando $\epsilon = 0,0001$.

Funções Utilidade Candidatas

Os fatos de que a função utilidade do usuário é linear e que o agente consegue representar os atributos do usuário baseado em uma transformação linear de seus próprios atributos permitem considerar apenas funções lineares também para o agente. Mesmo assim, deve-se enumerar, dentre os possíveis valores no espaço contínuo de vetores recompensas do agente, aqueles que serão considerados para representar as funções utilidade candidatas.

Se o usuário fosse determinista e pudessem ser consideradas questões hipotéticas sobre a preferência real do usuário entre políticas, no final do processo de EP, restariam apenas funções utilidade que guiassem o agente para a política ótima segundo o usuário (SILVA; COSTA; LIMA, 2006). Por outro lado, no capítulo 5 especificou-se um algoritmo para determinar todas as políticas que podem ser políticas ótimas, mesmo antes de especificar uma função utilidade para avaliar tais políticas, ou seja, as políticas não dominadas Π_{NonDom}. Então, escolhe-se para toda política não dominada π uma função utilidade candidata u que tenha como política ótima a política π (ABBEEL; NG, 2004; NG; RUSSELL, 2000; RUSSELL, 1998).

Ao escolher as funções utilidade candidatas, para cada política não dominada π, considerou-se o problema de encontrar um vetor recompensa $\mathbf{w}^\pi$ tal que $\langle \mathbf{w}^\pi, \bar{\boldsymbol{\mu}}^\pi_{\text{Ag}} - \bar{\boldsymbol{\mu}}^{\pi'}_{\text{Ag}} \rangle > 0$ para todo $\pi' \in \Pi_{\text{NonDom}}$. Então, define-se:

$$\mathbf{w}^\pi = \arg\max_{\mathbf{w} \in W_\pi} \sum_{\pi' \in \Pi_{\text{NonDom}}} \langle \mathbf{w}, \bar{\boldsymbol{\mu}}^\pi_{\text{Ag}} - \bar{\boldsymbol{\mu}}^{\pi'}_{\text{Ag}} \rangle,$$

onde $W_\pi = \{\mathbf{w} | \langle \mathbf{w}, \bar{\boldsymbol{\mu}}^\pi_{\text{Ag}} - \bar{\boldsymbol{\mu}}^{\pi'}_{\text{Ag}} \rangle > 0 \ \forall \pi' \in \Pi_{\text{NonDom}}, \ \langle \mathbf{w}, \bar{\boldsymbol{\mu}}^* - \bar{\boldsymbol{\mu}}^0 \rangle = 1$ e . Essa definição encontra um vetor recompensa $\mathbf{w}^\pi$ que garante π como política ótima e também garante que a qualidade de π segundo $\mathbf{w}^\pi$ quando confrontada com as outras políticas possua uma distância média significativa.

Mas, não basta considerar apenas funções utilidade deterministas, pois além de o usuário simulado ser não determinista, o agente não possui uma observação completa e determinista, pois os atributos observados para um par (s, a) podem ser diferentes a cada execução, e, ao final do processo, pode acontecer de todas funções utilidade candidatas apresentarem aderência nula às respostas do usuário. Assim, como no modelo do usuário, deve-se também considerar um desvio padrão σ_{Ag} para as funções utilidade candidatas, possibilitando que todas as funções utilidade candidatas apresentem algum grau de aderência às respostas do usuário.

Nos experimentos em questão, utilizando um erro $\epsilon = 0,01$ e apenas os três atributos observados pelo agente para encontrar o conjunto Π_{NonDom} (ver capítu-

lo 5), 62 políticas não dominadas foram encontradas, dando origem a 62 vetores recompensas. Além disso, considerou-se três diferentes possíveis valores para σ_u (0,02; 0,05 e 0,10), resultando em 186 funções utilidade candidatas.

Formulando Questões

Como foi visto neste capítulo, ao formular questões, ou seja, escolher duas políticas para serem executadas, se uma questão informativa é desejada, é necessário conhecer a distribuição de probabilidade $\Pr(r|\pi_q^1, \pi_q^2, u)$. Calcular analiticamente tal distribuição não é trivial, pois é exigido que todos os possíveis desdobramentos de uma política sejam enumerados. Então, aqui será comparada as duas técnicas de estimativa para tal distribuição apresentadas na seção 7.1.1: com base em vetores de atributos esperados e matrizes de covariância (equação (7.1) e equação (7.3)) e com base em simulação de Monte Carlo (equação (7.2) e equação (7.4)).

Note que, apesar do método de Monte Carlo produzir apenas uma estimativa de $\Pr(r|\pi_q^1, \pi_q^2, u)$, quanto maior o número de amostras utilizadas para construir tal estimativa, mais próxima tal estimativa estará da distribuição de probabilidades real. Neste experimento foram utilizadas 50 amostras para cada política de uma questão. Por outro lado, os cálculos de vetores de atributos esperados e matrizes de covariâncias podem ser feitos de forma exata se algumas condições de independência forem contempladas. No entanto, ao utilizar tais valores, deve-se assumir um modelo para a distribuição dos comportamentos de modo a se estimar $\Pr(r|\pi_q^1, \pi_q^2, u)$, e tal estimativa dependerá diretamente de tal escolha. No experimento aqui realizado foram consideradas distribuições normais.

Uma vez estimada a distribuição de probabilidades $\Pr(r|\pi_q^1, \pi_q^2, u)$, deve-se escolher quais serão os tipos de questões avaliadas para formular uma questão. Considerou-se dois métodos: o método com base em políticas fixas, isto é, ao formular uma questão determina-se duas políticas fixas as quais o agente utilizará para demonstrar os dois comportamentos que serão comparados pelo usuário (algoritmo 1), e o método com base em replanejamento de políticas, que replaneja políticas com base nos comportamentos parciais obtidos pelo agente durante a execução de uma política (algoritmo 2).

Enquanto na formulação de questões utilizando políticas fixas, $\Pr(r|\pi_q^1, \pi_q^2, u)$ pode ser calculado antes do início de interação entre agente e usuário, quando replanejamentos são utilizados é essencial que este cálculo possa ser feito em tempo real. O método de Monte Carlo possui uma dependência quadrática na quantidade de amostras utilizadas ao calcular $\Pr(r|\pi_q^1, \pi_q^2, u)$ e no modelo de distribuição ado-

tado. Ao utilizar vetores de atributos esperados e matrizes de covariâncias, pode-se calcular *a priori* tais valores (vetores e matrizes), restando apenas uma dependência no modelo de distribuição adotado no cálculo de $\Pr(r|\pi_q^1, \pi_q^2, u)$.

No caso do método de Monte Carlo, considerou-se o replanejamento apenas ao demonstrar o segundo comportamento ($\mathrm{Replaneja}(t, 1) = 0$ para todo t), reduzindo a apenas uma dependência linear na quantidade de amostras utilizadas. Mesmo assim, se o replanejamento for feito para cada ação a ser executada, pode-se inviabilizar a formulação de questões em tempo real. Então, no caso do método de Monte Carlo, o replanejamento da questão é realizado apenas de cinco em cinco ações a serem executadas, permanecendo com políticas fixas nas escolhas de ações intermediárias ($\mathrm{Replaneja}(t, 2) = 1$ para $t = 0, 5, 10, 15, \ldots$ e $\mathrm{Replaneja}(t, 2) = 0$ caso contrário). No entanto, ao utilizar vetores de atributos esperados e matrizes de covariâncias, o replanejamento é feito em todas as escolhas de ações na demonstração do segundo comportamento ($\mathrm{Replaneja}(t, 1) = 0$ e $\mathrm{Replaneja}(t, 2) = 1$ para todo t).

7.3.3 Resultados

O experimento realizado neste capítulo compara as duas técnicas para estimar a distribuição de probabilidades $\Pr(r|\pi_q^1, \pi_q^2, u)$ e os dois métodos utilizados para formular questões, gerando as configurações `MC-ND`, `Med-ND`, `MC-Re` e `Med-Re`, onde `MC` representa a técnica de simulação de Monte Carlo, `Med` representa a técnica baseada em vetores de atributos esperados, `ND` representa formulação de questões baseada em políticas não dominadas (políticas fixas) e `Re` representa a formulação de questões baseada em replanejamento. Então, para cada usuário simulado foram combinadas duas a duas cada uma das possibilidades para extrair as preferências do usuário. A EPCO é realizada por 50 questões, pois após essa quantidade de questão não ocorre melhoria nas preferências extraídas. Ainda, como referência para avaliar o quanto as comparações dos comportamentos obtidos são efetivas, foi considerado um método que seleciona aleatoriamente entre as políticas não dominadas ao formular questões no processo de EPCO.

Um usuário simulado Us é representado por um vetor recompensa $\mathbf{w}_{\mathrm{Us}}$ e um desvio padrão σ_{Us}. 50 diferentes vetores recompensas gerados aleatoriamente foram considerados, sendo combinados com três níveis de desvio padrão (0,00; 0,02 e 0,05), dando origem a 150 usuários diferentes. Cada um desses usuários tiveram suas preferências extraídas por 5 vezes para cada configuração do processo de EPCO.

Os gráficos da figura 7.2 demonstram a evolução da média de arrependimento para cada uma das configurações de EPCO e desvio padrão do usuário. Em uma pri-

meira análise pode-se comparar a formulação de questões aleatória com os métodos informativos. Nos gráficos fica claro que no início do processo de EPCO, os métodos informativos reduzem o arrependimento muito mais rápido que as questões aleatórias. No entanto, quando o desvio padrão das respostas do usuário é baixo, ao final do processo de EPCO, o método aleatório chega a obter arrependimento abaixo de métodos informativos. Isso pode ocorrer quando as respostas esperadas calculadas pelo agente, são muito diferentes das respostas esperadas de fato do usuário, produzindo um viés que impossibilita a redução do arrependimento. Ao formular questões aleatórias, o único viés está contido no conjunto de políticas não dominadas, permitindo que o agente siga reduzindo o arrependimento esperado.

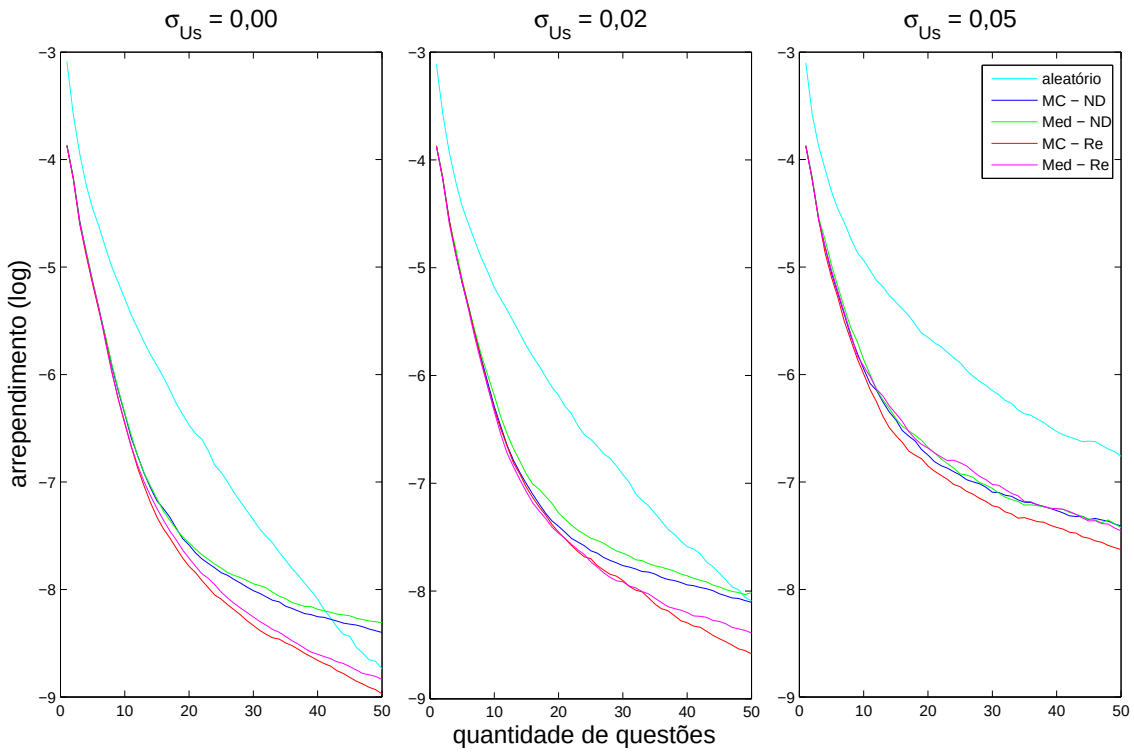

Figura 7.2: Arrependimento esperado no processo de EPCO.

Uma segunda análise consiste em comparar os métodos informativos entre si. Independentemente do método utilizado, no início do processo, o arrependimento esperado decai em taxas parecidas. Diferenças entre os métodos apenas são notadas no fim do processo de EPCO, quando os métodos baseados em replanejamento apresentam um menor arrependimento que os métodos baseados em políticas fixas. Ainda assim, essa diferença só é evidente quando ocorre um baixo desvio padrão nas respostas do usuário. Com relação às estimativas utilizadas para $\Pr(r|\pi_q^1, \pi_q^2, u)$, isto é, a comparação entre `MC-ND` e `Med-ND` ou `MC-Re` e `Med-Re`, ambas apresentaram desempenhos similares, com uma leve tendência de superioridade para a estimativa

baseada na simulação de Monte Carlo.

O arrependimento esperado serve como uma estimativa do desempenho de uma política escolhida pelo agente. Na figura 7.3 compara-se as políticas ótimas escolhidas pelo agente em cada um dos métodos. Nesses gráficos pode-se ver que, mesmo que haja redução no arrependimento durante todo o processo de EPCO, um desempenho próximo à otimalidade é obtido no início do processo, antes mesmo que 10 questões fossem formuladas. A avaliação média de uma tomada de decisão antes de extrair as preferências do usuário – a avaliação média das políticas não dominadas para cada um entre os 150 usuários – é de 0,4964, enquanto a média de avaliação das políticas ótimas, escolhidas por cada usuário entre as 62 políticas não dominadas, é de 0,6133. Após a décima questão, a média de avaliação das políticas ótimas, encontradas pelo a gente, entre todos os métodos informativos é de 0,6124 para σ_{Us} =0,00, 0,6119 para σ_{Us} =0,02 e 0,6101 para σ_{Us} =0,05.

Ainda, analisando a figura 7.3, pode-se comparar o desempenho do método aleatório contra os métodos informativos. O método aleatório apresentou um resultado sempre pior que os métodos informativos, mesmo quando o arrependimento no método aleatório está abaixo dos métodos informativos. Isso mostra que a medida de arrependimento esperado, quando o agente não possui informações completas, não indica de forma absoluta o desempenho do agente, pois a relação entre arrependimento esperado e o desempenho do agente, após realizar a EPCO, parece depender das questões formuladas.

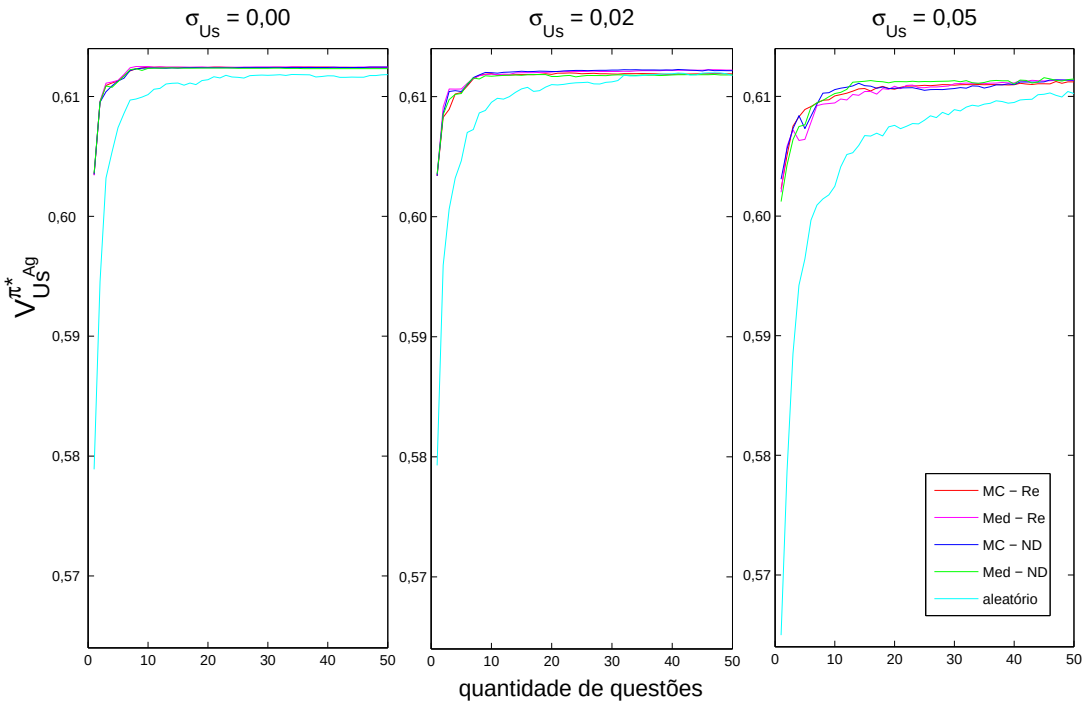

Figura 7.3: Avaliação segundo o usuário da política encontrada pelo agente.

Uma última análise que será feita é com relação ao desvio padrão do usuário. Note que, independente do desvio padrão do usuário ou do desvio padrão considerado pelo agente, a política ótima é sempre a mesma, pois ambos utilizam funções utilidade lineares. No entanto, o usuário também pode aprender, além de um vetor de recompensas, um desvio padrão que represente as preferências do usuário. Na tabela 7.1 é exibida, para cada desvio padrão do usuário, a porcentagem das funções utilidade candidatas com maior aderência às respostas do usuário para cada um dos desvios padrões considerados pelo agente (0,02, 0,05 e 0,10).

Tabela 7.1: Distribuição para os desvios padrões aprendidos junto ao usuário.

	Métodos Informativos			Método Aleatório		
σ_{Us}	0,00	0,02	0,05	0,00	0,02	0,05
σ_{Ag}=0,02	**0,882**	**0,642**	0,258	**0,644**	0,352	0,056
σ_{Ag}=0,05	0,089	0,231	0,214	0,244	**0,396**	0,384
σ_{Ag}=0,10	0,029	0,127	**0,528**	0,112	0,252	**0,560**

No método aleatório, observa-se uma tendência de inferir um desvio padrão para o agente sempre um pouco maior que o desvio padrão do usuário, sendo que tal acréscimo pode ser atribuído ao desvio padrão com relação aos atributos não observáveis pelo agente. No entanto, nos métodos informativos, essa tendência é combinada com a tendência de escolher desvios padrões próximos a 0,00. Essa segunda tendência deve-se ao fato de métodos informativos descartarem questões que não são bem diferenciadas pelas funções utilidade candidatas com maior aderência, isto é, questões onde as probabilidades de respostas atribuídas por cada função utilidade candidata com alta aderência seja próximo a 0,5. Dessa forma, para a função utilidade candidata ideal, as respostas emitidas pelo usuário tenderão a serem consistentes, favorecendo as funções utilidade com baixos desvios padrões. Ainda, devido ao fato de se considerar linearidade, para tomada de decisões ótimas não há a necessidade de diferenciar entre funções utilidade com alto ou baixo desvio padrão, sendo que as de baixo são sempre favorecidas quando poucas questões são consideradas, como nos experimentos realizados.

7.4 Considerações Finais

Os dois métodos utilizados para formular questões apresentados neste capítulo mostraram-se eficientes ao extrair as preferências do usuário. Mesmo que o método baseado em replanejamento, em teoria, garanta um melhor desempenho, se os atributos observados em um ambiente forem contínuos e possuírem distribuições mais

prováveis próximo a uma média, como ocorre em uma distribuição normal, o método que utiliza políticas fixas pode apresentar um desempenho similar. Enquanto os métodos baseados em replanejamento apresentam um custo computacional muito maior, nos experimentos aqui apresentados, esse custo não se justifica frente aos resultados obtidos. Caracterizar ambientes nos quais seja essencial esse replanejamento para uma efetiva EPCO é fundamental, para decidir qual dos métodos utilizar, ou ainda, se um PMD estendido é necessário.

O conjunto de atributos observados pelo agente utilizado nos experimentos também aproxima a eficiência dos dois métodos utilizados para estimar a distribuição de probabilidades $\Pr(r|\pi_q^1, \pi_q^2, u)$. Enquanto a estimativa por simulação de Monte Carlo pode ser sempre melhorada com a adição de novos exemplos de simulação, a utilização de vetores de atributos esperados e matrizes de covariâncias dependem diretamente do tipo de distribuição suposto ao estimar as respostas do usuário. Ao escolher entre um desses dois métodos, é necessário realizar uma análise preliminar para verificar o quanto a distribuição de atributos observados pode ser identificada com alguma distribuição analítica de fácil tratamento.

Por último, a consideração do arrependimento esperado para avaliar a qualidade das preferências extraídas junto ao usuário não se mostrou uma avaliação absoluta para o desempenho do agente utilizando tais preferências quando avaliado pelo usuário. Os experimentos demonstraram que, mesmo quando o arrependimento esperado é menor, nem sempre o desempenho é melhor. Por outro lado, quando a EPCO foi feita utilizando o arrependimento esperado para formular questões, obteve-se o melhor desempenho. Uma melhor análise teórica sobre todo esse processo é necessária, relacionando as tomadas de decisões sob incertezas não só à distribuição de probabilidades sobre as funções utilidade obtidas, mas também no método utilizado ao formular questões ao usuário.

8 CONCLUSÃO E TRABALHOS FUTUROS

Nesta tese, o problema de EPCO foi apresentado. Argumentou-se que o cenário utilizado nesse problema pode tornar a EP mais confortável do ponto de vista do usuário, no sentido de que o usuário necessita realizar menor esforço cognitivo para responder às questões que lhe são submetidas. Além disso, o usuário tem a oportunidade de sofrer as conseqüências das opções disponíveis, ou melhor dizendo, tem a oportunidade de avaliar conseqüências já sofridas por ele.

Além de ser desejável um menor esforço cognitivo do usuário, quando este escolhe entre conseqüências sofridas, tais escolhas podem ser mais fidedignas em relação às suas preferências. Primeiro, ao diminuir o esforço cognitivo do usuário, permite-se que o usuário não necessite levar em conta este fator no compromisso com outros fatores, fazendo com que pouco esforço cognitivo possa resultar em um grande retorno para o usuário. Esse compromisso pode ser ainda menor se for levado em conta que o usuário pode sentir e perceber os efeitos que os comportamentos produzem nele mesmo, sem a necessidade de inferir quais serão as conseqüências em que os comportamentos podem resultar.

A maior limitação conceitual da EPCO é justamente o horizonte de efeitos do comportamento, que devem estar contido no mesmo tempo de execução do comportamento. Se a avaliação de um comportamento não se limitar a duração do comportamento, isto é, os efeitos de um comportamento dura mais tempo que o comportamento – por exemplo, o recebimento de uma multa uma semana após a trajetória ter sido realizada. Nesse caso, se o usuário avaliar o comportamento logo após o mesmo ter ocorrido, o usuário deverá inferir resultados futuros, e, neste caso, a EPCO não apresenta as suas melhores características.

Outra limitação é operacional: o fato de que o comportamento deve ser demonstrado e que alguns comportamentos deverão ser demonstrados até que um modelo adequado das preferências do usuário seja inferido. Se comportamentos devem ser demonstrados algumas ou, dependendo do caso, várias vezes, o custo de demonstrar comportamentos não deve superar os benefícios obtidos após definir adequadamente as preferências do usuário. Em casos mais extremos a experiência só pode ser realizada

uma única vez, o que inviabiliza o uso de EPCO.

Os argumentos aqui utilizados em prol da EPCO foram feitos com base principalmente no trabalho de Plott (1996) que veio trazer uma contribuição no sentido de exigir-se um maior controle nos experimentos realizados para que a hipótese de preferência revelada possa ser considerada verdadeira, e é nesse sentido que o cenário aqui apresentado facilita a EP. Esta tese não foi desenvolvida no sentido de provar tais argumentos, uma vez que tal discussão envolve não apenas a área computacional, mas também outras áreas como a Psicologia e a Economia.

Por outro lado, este cenário, apresentado sob um arcabouço teórico, permitiu analisar formalmente um problema comum a qualquer cenário onde EP é utilizada: o problema da observação/interpretação das questões. A forma como as questões são realizadas nesta tese propicia uma menor dependência da interpretação do usuário em suas respostas, mas uma completa dependência da observação do mesmo. O segundo problema, o problema de demonstração de um comportamento, nem mesmo é contemplado tradicionalmente em cenários de EP, os quais se preocupam principalmente com a forma como uma questão é posta, mas, uma vez escolhida esta forma, assumem que nenhum problema existe na demonstração da mesma. Na EPCO, por outro lado, a demonstração de um comportamento depende do ambiente onde o agente está demonstrando os comportamentos.

Os resultados obtidos nesta tese consideraram, em sua grande maioria, que as preferências do agente podem ser representadas por uma função utilidade linear. Ao mesmo tempo que a área de Economia raramente considera funções utilidade simples como essas, funções utilidade lineares são o cerne dos PMDs. Esse arcabouço não só vem sendo perpetuado por sua simplicidade e facilidade no tratamento matemático, mas também por adequar-se a problemas de interesse que se deseje modelar. Dessa forma, embora muitas vezes funções utilidade lineares não representem exatamente as preferências do usuário, elas podem ser uma boa aproximação para tais preferências.

A escolha de funções utilidade lineares para tratar o problema nessa tese permitiu que alguns avanços teóricos fossem alcançados para esse caso específico. Embora as soluções apresentadas não se adequem a todos os casos, os resultados aqui alcançados servem como um guia para o tipo de resultado teórico que deve ser almejado com usuários mais complexos. Por outro lado, o modelo de usuário utilizado ao longo da tese permite que o mesmo seja inconsistente, possibilitando que mesmo quando não houver aderência das preferências do usuário às funções utilidade lineares, seja possível encontrar funções utilidade lineares que melhor se adequem às informações obtidas.

Na seção 8.1, as contribuições obtidas com esta tese são resumidas, enquanto na seção 8.2 são delineadas possibilidades de trabalhos futuros.

8.1 Contribuições

As contribuições alcançadas nesta tese foram:

- proposta do problema de EPCO e sua formalização no arcabouço de PMD;
- propriedades de um PMD
  - definição de um algoritmo para obtenção da matriz de covariância dos vetores de atributos obtidos com a execução de uma política no arcabouço de PMD;
  - definição de um algoritmo para obtenção das políticas não dominadas independentes do vetor recompensa para um PMD;
  - definição de um algoritmo para obtenção da política que resulta em qualquer vetor de atributos esperados convexo a um conjunto de vetores de atributos esperados referentes a políticas conhecidas;
- diferentes observações entre agente e usuário
  - definição da regra de inferência para quando existe conhecimentos completos a respeito do ambiente e observações do agente e do usuário;
  - definições de condições para que a EPCO seja factível quando não se consideram conhecido o modelo de observação do usuário e conhecimento de comportamentos completos no ambiente;
  - definições de condições para que a EPCO sejá factível quando se considera usuários com preferências modeladas por atributos com independência aditiva e neutralidade ao risco;
- formulação de questões
  - definição de uma técnica para formulação de questões com base em políticas fixas; e
  - definição de uma técnica para formulação de questões com base em políticas fixas e replanejamento.

As propriedades e algoritmos obtidos de forma genérica para PMDs simplificam a formulação de questões na EPCO, pois o uso do arcabouço de EPCO apresenta um

problema de custo computacional muito mais elevado para formulação de questões. Os trabalhos de EP consideram um arcabouço com poucas decisões, enquanto o arcabouço de PMD, por ser um arcabouço de decisões seqüenciais, resulta em cenário muito mais complexo em termos de decisões possíveis e seus desdobramento.

Então, mostrou-se como um cenário não apropriado em termos de custo computacional, após um pré-processamento, pode ser reduzido a um problema tratável computacionalmente. Devido à estrutura apresentada por PMDs e a noção de vetores de atributos esperados de uma política, pode-se reduzir o espaço de políticas estacionárias a um poliedro de vetores de atributos esperados. Mesmo que vetores de atributos esperados não permitam representar de forma fidedigna as propriedades de políticas, esta mudança de representação permite descartar o interior desde poliedro e utilizar apenas sua superfície no processo de formular questões. Uma aproximação ainda maior, que foi utilizada nesta tese, é considerar apenas os vértices de tal poliedro.

Um segundo ponto é que, a partir do momento em que um número reduzido de políticas é definido para utilizar-se na formulação de questões, pode-se considerar outras representações para essas políticas de forma que facilitem a formulação de questões. Os dois métodos empregados nessa tese foram: a) utilizar o vetor de atributos esperados de uma política agregado a uma matriz de covariância dessa mesma política e b) utilizar um conjunto de comportamentos amostrados de uma política obtidos por simulação de Monte Carlo.

Embora os resultados obtidos para redução do problema na formulação de questões tenham sido específicos para as condições que foram colocadas nessa tese, eles permitem delinear uma metodologia para modelos de ambientes e usuários mais complexos. Primeiro, através de alguma representação limitada das questões factíveis, reduz-se as questões a serem consideradas. Segundo, com um número reduzido de questões, pode-se obter uma representação mais completa das questões para que as mesmas possam ser corretamente avaliadas na formulação de novas questões.

A análise feita para o problema de diferentes observações entre agentes e usuários mostrou que a possibilidade de extrair de fato as preferências do usuário depende inversamente da probabilidade de ocorrência de cada um dos comportamentos em uma dada política e de quão completa é a observação do agente com relação a uma observação completa do ambiente. Considerando tal relação, apresentou-se uma estratégia para garantir a possibilidade de EPCO na qual o espaço de políticas candidatas a ótimas em um ambiente seja reduzido, mas que a política ótima dentro desse conjunto de políticas candidatas represente de forma adequada as preferências do usuário. Essa

redução dá-se ao limitar as políticas candidatas a políticas que exibam com uma taxa mínima todos os comportamentos possíveis. Essa limitação impede que políticas, que apresentem comportamentos ruins, tirem vantagem de preferências obtidas sob políticas boas, mas com baixa taxa de demonstração de comportamentos ruins.

Mesmo que esta solução apresente um conceito de otimalidade quando as interpretações do agente e do usuário de uma questão possuam viés, tal solução exige condições que podem ser difíceis de serem obtidas em alguns ambientes não deterministas, chegando-se ao absurdo de ser exigido que apenas políticas totalmente aleatórias possam ser consideradas ao tomar uma decisão. Por outro lado, existem cenários onde estas condições apresentam-se muito restritas, e condições mais relaxadas possam ser contempladas. Um desses cenários é quando a noção de atributos e funções utilidade lineares se aplicam. A noção de atributos permite construir um espaço métrico de interpretações do usuário, enquanto a noção de função utilidade linear permite dizer que duas interpretações próximas, possuam também respostas próximas. Nesse caso, basta que, dada uma interpretação do agente, a maioria das interpretações do usuário estejam próximas uma às outras.

O segundo problema trazido pelo cenário de EPCO é o de como demonstrar um comportamento arbitrário para o usuário. Devido a consideração de ambientes não deterministas, um comportamento pode ser obtido apenas com uma probabilidade de ocorrência. Levar em conta as distribuições de probabilidades de todas políticas factíveis, pode não ser viável computacionalmente. Mas, devido ao ambiente escolhido nesta tese, a quantidade de políticas a serem consideradas pôde ser reduzida.

Ainda que um comportamento arbitrário não possa ser obtido, se os atributos observados podem ser levados em consideração à cada tomada de decisão seqüencial, um comportamento exibido pode ser corrigido para chegar o mais próximo possível da arbitrariedade. No entanto, ao adotar esta estratégia, deve-se considerar políticas em um espaço de estados estendidos, que contempla o estado do ambiente e o comportamento demonstrado até o momento de cada decisão.

Embora a formulação com um espaço de estados estendido possa ser ótima do ponto de vista de EP, ela será custosa computacionalmente, pois o espaço de estado para realizar o planejamento será muito maior. Uma opção intermediária foi reformular uma questão baseada apenas em políticas estacionárias, isto é, após exibir um comportamento parcial, uma questão pode ser reformulada baseado no ocorrido e no esperado de cada política estacionária a partir do estado que o agente se encontra. Mesmo que teoricamente soluções que consideram replanejamento ou espaço de estados estendidos possam produzir questões mais informativas quando comparadas

a soluções com políticas estacionárias, na prática, o ganho pode não se justificar e depende do ambiente, como ocorreu nos experimentos aqui realizados.

8.2 Trabalhos Futuros

Além das restrições impostas em cada uma das fases do processo de EPCO que, como já foi comentado, podem ser relaxadas, produzindo resultados teóricos mais genéricos, outras abordagens e problemas não discutidos nessa tese podem ser objetos de pesquisas futuras.

Enquanto aqui optou-se apenas por utilizar questões relativas, nos trabalhos tradicionais de EP o uso de questões baseadas em loterias é muito comum. Dessa forma, torna-se necessário uma versão do problema de EPCO apresentado nesta tese, no qual tal tipo de questão possa ser aplicada. Uma opção seria considerar uma questão relativa formada por dois conjuntos de comportamentos, na qual o usuário deve optar entre qual conjunto de comportamentos é mais desejável. Então, a interpretação dessa questão poderia ser feita como na loteria, sendo que um conjunto de comportamentos representa uma loteria e cada comportamento nesse conjunto ocorre com probabilidade uniforme.

Ao utilizar conjunto de comportamentos, o problema de analisar a demonstração e formulação de questões torna-se muito mais complexo. No entanto, quando comportamentos podem ser integrados, é possível que pelo menos um comportamento médio possa ser obtido. Considerando esse fato, pode-se utilizar a estratégia de abordagem de um determinado vetor de atributos esperados desejado (MANNOR; SHIMKIN, 2004; SHIMKIN; SHWARTZ, 1993; BLACKWELL, 1956), possibilitando que exista um conjunto de loterias que possam ser obtidas arbitrariamente com um erro pequeno.

Outro ponto interessante é que a medida de arrependimento considera as possíveis decisões a serem tomadas pelo agente para formular as questões, sendo que as questões são realizadas apenas até garantir a otimalidade da decisão. Se deseja-se transferir as preferências obtidas em um ambiente para um segundo ambiente, é importante mensurar a validade de tal transferência na tomada de decisão nesse segundo ambiente. Trabalhar com funções utilidade intermediando a obtenção de uma política ótima, apresenta duas grandes vantagens: o uso de conhecimentos *a priori* sobre as preferências do usuário representado por restrições e a abstração da descrição de um problema. Enquanto o uso da primeira vantagem foi largamente utilizado durante a tese, na consideração de funções utilidade lineares, o segundo não foi abordado. Silva, Costa e Lima (2006) apresentam resultados empíricos, onde se realiza a transferência

das preferências do avaliador obtidas por um robô para um segundo robô também maximizá-las, demonstrando essa vantagem.

As preferências extraídas do usuário dependem de dois fatores. Primeiro, os comportamentos que podem ser demonstrados em um ambiente, pois qualquer informação só poderá ser obtida sob tais comportamentos limitando a resolução com que uma função utilidade pode ser obtida. Segundo, as políticas disponíveis para decisão, já que o objetivo do agente na EPCO é apenas encontrar uma função utilidade que lhe garanta uma política ótima. Considerando esses fatores, deve-se determinar uma medida de distância entre ambientes, de modo que uma distância pequena garanta uma transferência com sucesso. Silva, Costa e Lima (2006) mostram que esse sucesso ocorre de forma diferente em diferentes direções de transferência.

O processo de EPCO considerado nesta tese foi inspirado no processo de EP, isto é: formula-se uma questão, propõe-na ao usuário, obtém uma resposta do usuário e realiza-se inferência sobre as preferências do usuário. No entanto, tudo que se deseja é uma política que atenda às preferências do usuário, ou, de forma indireta, uma função utilidade que guie a tal política. Métodos que cheguem à mesma conclusão por outras vias também precisam ser estudados. A busca no espaço de políticas permite obter informações sobre o valor das políticas (HOOS; BOUTILIER, 2000), mesmo quando a observação do agente não é compatível com o objetivo que se deseja alcançar, bastando que avaliações diretas sobre as políticas possam ser obtidas.

Consideram o caso onde nenhuma informação é suposta sobre o usuário, Silva, Lima e Costa (2007) propõem uma busca aleatória no espaço de políticas, mas sempre utilizando o espaço de funções utilidade para guiar a busca. Se diferentes observações impossibilitam inferir a aderência das respostas do usuário a funções utilidade candidatas, elas podem auxiliar na escolha de uma vizinhança adequada na busca aleatória. No entanto, a busca aleatória, utilizando amostras de comportamentos de uma política, não garante que uma busca convirja para a decisão ótima (PRUDIUS; ANDRADóTTIR, 2004). Por outro lado, as freqüências de ocorrências das políticas obtidas ao longo do processo de tal busca podem representar a qualidade das políticas se um processo adequado de busca pode ser realizado (SANBORN; GRIFFITHS, 2008; HASTINGS, 1970).

O desenvolvimento das sugestões de trabalhos futuros aqui apresentadas propiciarão uma nova classe de aplicações para a EPCO. Mesmo assim, os resultados obtidos nesta tese contemplam o processo inteiro de EPCO, possibilitando o uso direto do arcabouço aqui desenvolvido em uma aplicação real de delegação a agentes artificiais. Ao mesmo tempo, o desenvolvimento de tal arcabouço foi feito de tal forma a

possibilitar a sua adaptação a cenários aqui não considerados.

REFERÊNCIAS BIBLIOGRÁFICAS

ABBEEL, P.; NG, A. Y. Apprenticeship learning via inverse reinforcement learning. **Proceedings of the Twenty-first International Conference on Machine Learning**. Alberta, Canada: Omnipress, 2004.

ACTIVMEDIA ROBOTICS. *Saphira's Manual*. Menlo Park, CA, 2001. Version 8.0a.

ALOYSIUS, J. A.; DAVIS, F. D.; WILSON, D. D.; TAYLOR, A. R.; KOTTEMANN, J. E. User acceptance of multi-criteria decision support systems: The impact of preference elicitation techniques. *European Journal of Operational Research*, v. 169, n. 1, p. 273–285, February 2006.

BASU, K. Determinateness of the utility function: Revisiting a controversy of the thirties. *Review of Economic Studies*, v. 49, n. 2, p. 307–11, April 1982.

BERNOULLI, D. Exposition of a new theory on the measurement of risk. *Econometrica*, v. 22, n. 1, p. 23–36, 1954.

BETTMAN, J.; LUCE, M.; PAYNE, J. Constructive consumer choice processes. *Journal of Consumer Research*, v. 25, p. 187–217, December 1998.

BLACKWELL, D. An analogue for the minimax theorem for vector payoffs. *Pacific Journal of Mathematics*, v. 6, p. 1–8, 1956.

BLAVATSKYY, P. R. Stochastic expected utility theory. *Journal of Risk and Uncertainty*, Springer Netherlands, v. 34, n. 3, p. 259–286, June 2007. ISSN 0895-5646.

BLEICHRODT, H.; PINTO, J. L.; WAKKER, P. P. Making descriptive use of prospect theory to improve the prescriptive use of expected utility. *Management Science*, INFORMS, Institute for Operations Research and the Management Sciences (INFORMS), Linthicum, Maryland, USA, v. 47, n. 11, p. 1498–1514, 2001. ISSN 0025-1909.

BONET, B.; GEFFNER, H. Planning and control in artificial intelligence: A unifying perspective. *Applied Intelligence*, v. 14, n. 3, p. 237–252, 2001.

BOUTILIER, C. A pomdp formulation of preference elicitation problems. **Eighteenth national conference on Artificial intelligence**. Menlo Park, CA, USA: American Association for Artificial Intelligence, 2002. p. 239–246. ISBN 0-262-51129-0.

BOUTILIER, C. On the foundations of expected expected utility. **Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence (IJCAI-03)**. Acapulco: Morgan Kaufmann, 2003. p. 285–290.

BOUTILIER, C.; PATRASCU, R.; POUPART, P.; SCHUURMANS, D. Regret-based utility elicitation in constraint-based decision problems. **IJCAI-05, Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence**. Edinburgh, Scotland: Professional Book Center, 2005. p. 929–934. ISBN 0938075934.

BOUTILIER, C.; PATRASCU, R.; POUPART, P.; SCHUURMANS, D. Constraint-based optimization and utility elicitation using the minimax decision criterion. *Artificial Intelligence*, Elsevier Science Publishers Ltd., Essex, UK, v. 170, n. 8, p. 686–713, 2006. ISSN 0004-3702.

BRAGA, J.; STARMER, C. Preference anomalies, preference elicitation and the discovered preference hypothesis. *Environmental & Resource Economics*, v. 32, p. 55–89, 2005.

BRAZIUNAS, D. *Computational Approaches to Preference Elicitation*. Department of Computer Science, University of Toronto, 2006.

BUFFETT, S.; FLEMING, M. W. Persistently effective query selection in preference elicitation. **IAT '07: Proceedings of the 2007 IEEE/WIC/ACM International Conference on Intelligent Agent Technology**. Washington, DC, USA: IEEE Computer Society, 2007. p. 491–497. ISBN 0-7695-3027-3.

CARENINI, G.; POOLE, D. Constructed preferences and value-focused thinking: Implications for ai research on preference elicitation. **Preferences in AI and CP: Symbolic Approaches: Papers from the AAAI Workshop**. Edmonton, Canada: American Association for Artificial Intelligence, Menlo Park, California, 2002. p. 9–15.

CHAJEWSKA, U.; GETOOR, L.; NORMAN, J.; SHAHAR, Y. Utility elicitation as a classification problem. **Uncertainty in Artificial Intelligence. Proceedings of the Fourteenth Conference (1998)**. San Francisco: Morgan Kaufmann Publishers, 1998. p. 79–88.

CHAJEWSKA, U.; KOLLER, D.; PARR, R. Making rational decisions using adaptive utility elicitation. **Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence**. Austin, Texas: AAAI Press / The MIT Press, 2000. p. 363–369. ISBN 0-262-51112-6.

CHANKONG, V.; HAIMES, Y. Y. *Multiobjective Decision Making: Theory and Methodology*. New York: North-Holland, 1983.

CHEN, L.; PU, P. *Survey of Preference Elicitation Methods*. Swiss Federal Institute of Technology in Lausanne (EPFL), 2004. IC/2004/67.

CUBITT, R. P.; STARMER, C.; SUGDEN, R. Discovered preferences and the experimental evidence of violations of expected utility theory. *Journal of Economic Methodology*, v. 8, n. 3, p. 385–414, 2001.

DENNIS, B.; KOCHERLAKOTA, S.; SAWANT, A.; TATEOSIAN, L.; HEALEY, C. G. Designing a visualization framework for multidimensional data. *IEEE Comput. Graph. Appl.*, IEEE Computer Society Press, Los Alamitos, CA, USA, v. 25, n. 6, p. 10–15, 2005. ISSN 0272-1716.

DENNIS, B. M.; HEALEY, C. G. *A Survey of Preference Elicitation*. Knowledge Discovery Lab, Department of Computer Science, North Carolina State University, 2003. TR-2005-41.

DOSHI, F.; ROY, N. The permutable pomdp: fast solutions to pomdps for preference elicitation. **AAMAS '08: Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems**. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems, 2008. p. 493–500. ISBN 978-0-9817381-0-9.

GRUNE, T. The problems of testing preference axioms with revealed preference theory. *Analyse & Kritik*, Lucius & Lucius, Suttgart, v. 26, p. 382–397, 2004.

HA, V.; HADDAWY, P. Toward case-based preference elicitation: Similarity measures on preference structures. **Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence**. San Francisco: Morgan Kaufmann Publishers, 1998. p. 193–201.

HARRISON, G. W.; HARSTAD, R. M.; RUTSTROM, E. E. Experimental methods and elicitation of values. *Experimental Economics*, v. 7, n. 2, p. 123–140, 06 2004.

HASTINGS, W. K. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, v. 57, n. 1, p. 97–109, 1970.

HEY, J. D. Experimental investigations of errors in decision making under risk. *European Economic Review*, v. 39, n. 3-4, p. 633–640, April 1995.

HOOS, H. H.; BOUTILIER, C. Solving combinatorial auctions using stochastic local search. **Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence**. Austin, Texas: AAAI Press / The MIT Press, 2000. p. 22–29. ISBN 0-262-51112-6.

HUI, B.; BOUTILIER, C. Toward experiential utility elicitation for interface customization. **Proceedings of the 24th Conference in Uncertainty in Artificial Intelligence**. Helsinki, Finland: AUAI Press, 2008. p. 298–305. ISBN 0-9749039-4-9.

IYENGAR, V. S.; LEE, J.; CAMPBELL, M. Evaluating multiple attribute items using queries. **EC '01: Proceedings of the 3rd ACM conference on Electronic Commerce**. Tampa, Florida, USA: ACM, 2001. p. 144–153. ISBN 1-58113-387-1.

JUNG, S. Y.; HONG, J.-H.; KIM, T.-S. A formal model for user preference. **ICDM '02: Proceedings of the 2002 IEEE International Conference on Data Mining (ICDM'02)**. Washington, DC, USA: IEEE Computer Society, 2002. p. 235. ISBN 0-7695-1754-4.

JUNG, S. Y.; HONG, J.-H.; KIM, T.-S. A statistical model for user preference. *IEEE Trans. on Knowl. and Data Eng.*, IEEE Educational Activities Department, Piscataway, NJ, USA, v. 17, n. 6, p. 834–843, 2005. ISSN 1041-4347.

KEENEY, R. L.; RAIFFA, H. *Decisions with Multiple Objectives: Preferences and Value Tradeoffs*. New York: Wiley, 1976.

LAHAIE, S. M.; PARKES, D. C. Applying learning algorithms to preference elicitation. **EC '04: Proceedings of the 5th ACM conference on Electronic commerce**. New York, NY, USA: ACM, 2004. p. 180–188. ISBN 1-58113-711-0.

LUCE, R. D.; WINTERFELDT, D. von. What common ground exists for descriptive, prescriptive, and normative utility theories? *Manage. Sci.*, INFORMS, Institute for Operations Research and the Management Sciences (INFORMS), Linthicum, Maryland, USA, v. 40, n. 2, p. 263–279, 1994. ISSN 0025-1909.

MANNOR, S.; SHIMKIN, N. A geometric approach to multi-criterion reinforcement learning. *J. Mach. Learn. Res.*, MIT Press, Cambridge, MA, USA, v. 5, p. 325–360, 2004. ISSN 1533-7928.

NEUMANN, J. von; MORGENSTERN, O. *The Theory of Games and Economic Behaviour*. 2. ed. Princeton: Princeto University Press, 1947.

NG, A. Y.; RUSSELL, S. J. Algorithms for inverse reinforcement learning. **ICML '00: Proceedings of the Seventeenth International Conference on Machine Learning**. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2000. p. 663–670. ISBN 1-55860-707-2.

PATRASCU, R.; BOUTILIER, C.; DAS, R.; KEPHART, J. O.; TESAURO, G.; WALSH, W. E. New approaches to optimization and utility elicitation in autonomic computing. **Proceedings, The Twentieth National Conference on Artificial Intelligence and the Seventeenth Innovative Applications of Artificial Intelligence Conference**. Pittsburgh, Pennsylvania, USA: AAAI Press / The MIT Press, 2005. p. 140–145. ISBN 1-57735-236-X.

PAYNE, J.; BETTMAN, J.; JOHNSON, J. Behavioral decision research: a constructive processing perspective. *Annual Review of Psychology*, v. 43, p. 87–131, 1992.

PLOTT, C. R. Rational individual behaviour in markets and social choice processes: The discovered preference hypothesis. **Rational Foundations of Economic Behaviour**. London and New York: Macmillan and St. Martin's Press, 1996. p. 225–250.

PRUDIUS, A. A.; ANDRADóTTIR, S. Simulation optimization using balanced explorative and exploitative search. **WSC '04: Proceedings of the 36th conference on Winter simulation**. Washington, D.C.: Winter Simulation Conference, 2004. p. 545–549. ISBN 0-7803-8786-4.

PU, P.; FALTINGS, B.; TORRENS, M. User-Involved Preference Elicitation. **IJCAI'03 Workshop on Configuration**. Acapulco, Mexico: Morgan Kaufmann, 2003.

QIN, M.; BUFFETT, S.; FLEMING, M. W. Predicting user preferences via similarity-based clustering. **Canadian Conference on AI**. Canada: Springer, 2008. (Lecture Notes in Computer Science, v. 5032), p. 222–233. ISBN 978-3-540-68821-1.

ROSS, S. M. *Applied probability models with optimization applications*. San Francisco: Holden-Day, 1970.

RUSSELL, S. Learning agents for uncertain environments (extended abstract). **COLT' 98: Proceedings of the eleventh annual conference on Computational learning theory**. New York, NY, USA: ACM, 1998. p. 101–103. ISBN 1-58113-057-0.

RUSSELL, S.; NORVIG, P. *Inteligência Artificial*. 2nd. ed. Rio de Janeiro: Elsevier:Campus, 2004.

SANBORN, A.; GRIFFITHS, T. Markov chain monte carlo with people. In: PLATT, J.; KOLLER, D.; SINGER, Y.; ROWEIS, S. (Ed.). *Advances in Neural Information Processing Systems 20*. Cambridge, MA: MIT Press, 2008. p. 1265–1272.

SCHMIDT, U.; NEUGEBAUER, T. Testing expected utility in the presence of errors. *The Economic Journal*, v. 117, n. 518, p. 470–485, 03 2007.

SHIMKIN, N.; SHWARTZ, A. Guaranteed performance regions in markovian systems with competing decision makers. *IEEE Transactions on Automatic Control*, v. 38, p. 84–95, 1993. ISSN 0018-9286.

SILVA, V. F. d.; COSTA, A. H. R.; LIMA, P. Inverse reinforcement learning with evaluation. **IEEE International Conference on Robotics and Automation (ICRA'06)**. Orlando, FL: IEEE, 2006. p. 4246–4251.

SILVA, V. F. da; LIMA, P.; COSTA, A. H. R. Eliciting preferences over observed behaviours based on relative evaluations. **Proceedings of the IEEE/RSJ 2007 International Joint Conference on Intelligent Robots and Systems**. San Diego, CA: IEEE, 2007. p. 423–428.

SISKOS, Y.; SPYRIDAKOS, A. Intelligent multicriteria decision support: Overview and perspectives. *European Journal of Operational Research*, v. 113, n. 2, p. 236–246, March 1999.

STARMER, C. Developments in non-expected utility theory: The hunt for a descriptive theory of choice under risk. *Journal of Economic Literature*, v. 38, p. 332–382, 2000.

SUTTON, R. S.; BARTO, A. G. *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press, 1998.

THIEBAUX, S.; GRETTON, C.; SLANEY, J.; PRICE, D.; KABANZA, F. Decision-theoretic planning with non-markovian rewards. *Journal of Artificial Intelligence Research*, v. 25, p. 17–74, 2006.

TODOROV, A.; GOREN, A.; TROPE, Y. Probability as a psychological distance: Construal and preferences. *Journal of Experimental Social Psychology*, v. 43, p. 473–482, 2007.

VIAPPIANI, P.; FALTINGS, B.; PU, P. The lookahead principle for preference elicitation: Experimental results. **Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)**. Heidelberg, Germany: Springer Verlag, 2006. v. 4027 NAI, p. 378–389. ISBN 0302-9743.

WANG, T.; BOUTILIER, C. Incremental utility elicitation with the minimax regret decision criterion. **Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence (IJCAI-03)**. Acapulco: Morgan Kaufmann, 2003. p. 309–316.

WOOLDRIDGE, M. *An Introduction to Multiagent Systems*. Chichester, England: John Wiley & Sons, Inc., 2002. ISBN 047149691X.

ZILHãO, A. Psicologia popular, teoria da decisão e comportamento humano comum. *Annual Review of Psychology*, v. 1, n. 10, p. 22–42, May 2001.

ZUKERMAN, I.; ALBRECHT, D. W. Predictive statistical models for user modeling. *User Modeling and User-Adapted Interaction*, Kluwer Academic Publishers, Hingham, MA, USA, v. 11, n. 1-2, p. 5–18, 2001. ISSN 0924-1868.