UNIVERSIDADE DE LISBOA
FACULDADE DE LETRAS

Problema de Newcomb: Solução evidencialista ou causalista?

Diogo Andrade Ribeiro Fernandes

Orientador: Professor Doutor António José Teiga Zilhão

Tese especialmente elaborada para obtenção do grau de Doutor no ramo de Filosofia, na especialidade de Epistemologia e Filosofia da Ciência

2019

UNIVERSIDADE DE LISBOA
FACULDADE DE LETRAS

Problema de Newcomb: Solução evidencialista ou causalista?

Diogo Andrade Ribeiro Fernandes

Orientador: Professor Doutor António José Teiga Zilhão

Tese especialmente elaborada para obtenção do grau de Doutor no ramo de Filosofia, na especialidade de Epistemologia e Filosofia da Ciência

Júri:

Presidente: Doutor António Pedro Sangreman Mesquita, Professor Catedrático e membro do Conselho Científico da Faculdade de Letras da Universidade de Lisboa

Vogais:

- Doutor José António Díez Calzada, Professor Titular
  Department de Filosofia da Universidade de Barcelona, Espanha
- Doutor Erich Rast, Investigador Faculdade de Ciências Sociais e Humanas da Universidade Nova de Lisboa
- Doutor Bruno Miguel Jacinto, Investigador Júnior da FCT Faculdade de Ciências da Universidade de Lisboa
- Doutor António José Teiga Zilhão, Professor Associado com Agregação
  Faculdade de Letras da Universidade de Lisboa, orientador

Trabalho financiado por fundos nacionais através da Fundação para a Ciência e Tecnologia (FCT), no âmbito de Bolsa Individual de Doutoramento (SFRH/BD/79876/2011)

2019

Agradecimentos

Estou profundamente grato ao Professor Doutor António Zilhão, sem o incentivo e o apoio do qual esta dissertação nunca teria sido concluída.
Estou também profundamente grato à Fundação para a Ciência e a Tecnologia, pela bolsa de doutoramento (SFRH/BD/79876/2011) que me foi concedida.
Este trabalho é dedicado aos meus avós, Maria da Graça Soares e António Pedro Paiva de Andrade.

Resumo

O objectivo deste trabalho consiste em avaliar os méritos explicativos e normativos da teoria matemática da acção racional conhecida como teoria bayesiana, tal como foi apresentada por Leonard Savage, em *The Foundations of Statistics* (1954), e mais tarde reformulada por Richard Jeffrey em *The Logic of Decision* (1965). O ponto de partida consistirá na apresentação analítica do Problema de Newcomb, um dos puzzles decisórios mais conhecidos e discutidos, o qual põe em causa esses referidos méritos. A tentativa de solucionar este puzzle oferecerá a oportunidade para colocar em confronto duas interpretações distintas do princípio da maximização da utilidade subjectiva esperada: a interpretação evidencialista e a interpretação causalista. Divergindo estas interpretações quanto à identificação daqueles que são os factores relevantes para a determinação de uma fórmula da utilidade esperada, a discussão em redor dessa divergência será o objecto central da investigação.

Ao defender-se a necessidade de incorporar o conhecimento causal na construção de uma teoria da decisão adequada, defender-se-á a superioridade da teoria causal sobre a teoria evidencial. Essa superioridade é fortalecida pela constatação de que as estratégias revisionistas da teoria evidencial, a *tickle defense* e o ratificacionismo, não são capazes de lidar com certos problemas decisórios relevantes. Defender-se-á ainda que, num contexto decisório, a aplicação de um processo de revisão de crenças conhecido por *visualização* (Lewis 1976), alternativo à condicionalização bayesiana, permite neutralizar os contra-exemplos à teoria causal. Dada a opção pela versão contrafactual da teoria causal, constatou-se que, para efeitos de aplicação da teoria, é necessário adoptar-se uma resolução para a vagueza inerente à análise semântica de contrafactuais, resolução essa que dependerá de considerações pragmáticas. Defender-se-á, portanto, que a solução para o Problema de Newcomb dependerá do modo de resolução dessa vagueza, e que o confronto entre soluções distintas passará a dar-se no contexto da própria teoria causal.

Palavras-chave: Teoria evidencial da decisão, teoria causal da decisão, contrafactuais, ratificacionismo, vizualização.

Abstract

The aim of this dissertation is to evaluate the explanatory and normative merits of the mathematical theory of rational decision-making, known as Bayesian theory, as presented by Leonard Savage in *The Foundations of Statistics* (1954), and later reformulated by Richard Jeffrey in *The Logic of Decision* (1965). The starting point will be the analytical presentation of Newcomb's Problem, one of the most well-known and discussed decision-theoretical puzzles, which calls into question these merits. The attempt to solve this puzzle will offer the opportunity to confront two distinct interpretations of the principle of expected subjective utility maximization: the evidentialist interpretation and the causalist interpretation. As these interpretations diverge on the identification of those factors that are relevant to the determination of an expected utility formula, the discussion around this divergence will be the central object of the investigation.

In defending the need to incorporate causal knowledge into the construction of an adequate decision theory, I will defend the superiority of causal theory over evidential theory. This superiority is confirmed by the realization that revisionist strategies of the evidential theory, the tickle defense and ratificationism, are not capable of dealing with certain relevant decision-making problems. It will also be argued that, in a decision-making context, the application of a belief revision procedure known as *imaging* (Lewis 1976), an alternative to Bayesian conditionalization, defuses the counterexamples to the causal theory. Given the choice of the counterfactual version of the causal theory, it was found that, for the purposes of theory application, a resolution for the vagueness inherent in the semantic analysis of counterfactuals is required, a resolution which will depend on pragmatic considerations. It will be argued, therefore, that the solution to Newcomb's problem will depend on the way in which this vagueness is resolved, and that the confrontation between different solutions will take place in the context of the causal theory itself.

Keywords: Evidential decision theory, causal decision theory, counterfactuals, ratificationism, imaging.

Índice

Introdução

Consideremos uma situação com a qual alguns já se terão deparado. Amanhã será um dos dias mais importantes na nossa vida, o dia do exame final. Consumimos horas, dias e meses a estudar, pois se passarmos no exame, teremos dado um passo decisivo para obter aquele posto de trabalho com que sempre sonhámos. Mas, na noite do dia anterior, somos inesperadamente convidados para um jantar animado com alguns amigos, incluindo aquela pessoa especial que há muito desejamos conhecer. Sabemos, contudo, que, nestas ocasiões, temos tendência para beber mais do que a nossa conta, principalmente quando a bebida é uma condição necessária para uma interacção social mais proveitosa: queremos fazer boa figura. Em vez de ficarmos em casa a rever a matéria, decidimos ir na mesma ao jantar e a noite corre de acordo com o esperado: esquecemos por algumas horas a ansiedade habitual nestas ocasiões (a do exame e a do convívio), combinamos voltar a encontrar-nos com a pessoa especial, e chegamos a casa já de madrugada, poucas horas antes do exame, emocional e fisicamente exaustos. Nessa manhã não ouvimos o despertador e quando acordamos, desorientados, mal temos tempo de nos vestir e comparecer uma hora atrasados no local do exame. Mas, quando aí chegamos, algo de inesperado aconteceu: o professor adoeceu e o exame terá de ser adiado. Em suma, a vida não nos podia correr melhor, pensamos.

A decisão de ir ao jantar foi, ou não, a decisão certa? Olhando para o que *aconteceu*, a resposta terá de ser positiva, pois as *consequências* foram-nos extremamente favoráveis. Mas terá sido a acção racional? Para responder a esta questão teremos de considerar quais eram as nossas crenças e desejos *antes* de a decisão ter sido tomada. Apesar de nos apetecer ir à festa, a probabilidade de nos arrependermos era enorme. As consequências agradáveis da acção efectuada não compensariam a oportunidade perdida e a perspectiva de uma vida profissional arruinada. Além disso, a probabilidade de bebermos demasiado, caso decidissemos ir à festa, era demasiado elevada. Isto significa que muitas vezes tomamos decisões que se mostram ‘acertadas’, ainda que claramente irracionais, e muitas vezes tomamos decisões racionais que mais tarde se mostram ‘erradas’. Uma questão relevante, por exemplo, consiste em saber se temos motivos para nos arrependermos genuinamente, quando sabemos que tomámos uma decisão racional.

A teoria da decisão que será objecto do nosso estudo preocupa-se apenas com a racionalidade das nossas acções. Trata-se, portanto, de uma teoria que tenta formular, da

forma mais exacta possível, hipóteses acerca daquilo em que consiste agir racionalmente. Uma outra maneira de nos referirmos à racionalidade das nossas decisões é considerá-las como estando, ou não, justificadas. Essa justificação depende da combinação apropriada de certos aspectos relevantes, pertencentes a nós próprios e ao mundo. Ou seja, a teoria da decisão tem como objectivo estabelecer uma conexão entre os nossos desejos, ou preferências por determinados resultados, e aquelas que são *realmente* as acções mais apropriadas para as satisfazer, tendo em conta as nossas crenças acerca do mundo ou das maneiras como este pode contribuir para que esses resultados se produzam. Temos, assim, à nossa disposição todos os elementos relevantes para formalizar qualquer problema de decisão que se nos coloque: 1) um conjunto de acções alternativas, 2) os resultados ou consequências que são objecto das preferências do agente, e 3) os estados do mundo que, juntamente com as acções empreendidas, contribuem para que esses resultados se produzam.

Uma maneira de relacionar estes três elementos, embora um tanto ou quanto abstracta, consiste em definir as acções como funções de estados para consequências, funções que tomam como argumentos os estados do mundo relevantes e que apresentam como valores as consequências possíveis das acções. Na posse destes dados, podemos agora também visualizar qualquer problema de decisão. Existem várias maneiras de o fazer, mas a mais comum, e aquela que será aqui utilizada, é a matriz ou tabela de decisão. Consideremos um problema de decisão típico e familiar (com apenas duas acções e dois estados): vou sair de casa para dar um passeio e o tempo está nublado. Levo, ou não, o chapéu-de-chuva? Uma possível formalização e visualização do problema é a seguinte:

	Chove	Não chove
Não levar chapéu	passeio estragado	passeio muito confortável
Levar chapéu	passeio confortável	passeio desconfortável

As nossas crenças acerca da ocorrência dos estados do mundo relevantes admitem, naturalmente, graus de incerteza, e, nessa medida, é possível atribuir-lhes probabilidades.

Quando isto acontece, estamos perante um problema de decisão *em condições de incerteza*. A teoria da decisão que estudaremos lida apenas com *decisões em condições de incerteza* e com a necessidade de se formular uma regra de decisão que se aplique a estes problemas. Essas atribuições de probabilidade não podem ser feitas de forma arbitrária, daí que as hipóteses formuladas pela teoria, acerca do modo como as nossas crenças e desejos influenciam as nossas escolhas, devam ser matematicamente precisas. A teoria da decisão moderna tem como regra de decisão a *maximização da utilidade subjectiva esperada*; a sua hipótese fundamental é a de que os agentes agem racionalmente se, e somente se, agirem de forma a maximizar a utilidade sujectiva esperada; seleccionando, portanto, de entre as acções disponíveis, aquela cuja utilidade subjectiva esperada é a mais elevada.

O conceito de utilidade apresenta uma história venerável que iremos aqui ignorar. Para os efeitos do nosso estudo, tal conceito é utilizado para representar o quão boas, ou más, são as consequências das acções do ponto-de-vista do agente. Mais precisamente, o quão boas, ou más, são essas acções, relativamente umas às outras. Por forma a iniciarmos o estudo e compreender o modo através do qual cada agente calculará a utilidade esperada de cada uma das acções, é necessário compreender o conceito de *função de utilidade*. Considera-se que cada agente possui uma escala na qual se encontram ordenadas as suas preferências relativamente a todas as consequências das acções disponíveis. Essa ordenação é feita, portanto, de acordo com o quão desejáveis são essas consequências do ponto-de-vista do agente. É possível, então, definir uma *função de utilidade*, fazendo corresponder a cada um dos itens dessa escala um número real, o qual representará a utilidade desse item. Mais, essa função, para além de representar uma ordem, representa também a magnitude relativa da desejabilidade dessas consequências. A teoria oferecerá, obviamente, um método através do qual essa escala pode ser construída de uma forma matematicamente precisa. No caso do problema de decisão acima, tal escala pode ser representada, por exemplo, da seguinte maneira,

1. Passeio muito confortável (*melhor do que*) 18 *utiles*
2. Passeio confortável (") 15 "
3. Passeio desconfortável (") 0 "
4. Passeio estragado -20 "

O segundo conceito fundamental da teoria da utilidade esperada é o da *distribuição de probabilidades* do agente. As crenças deste acerca da ocorrência dos estados do mundo serão igualmente representadas por valores numéricos. No caso descrito, o agente sabe, por exemplo, que o céu está nublado e que o boletim meteorológico previu a ocorrência de chuva. Contudo, ele não tem a certeza de que vai chover, atribuindo apenas uma elevada probabilidade à ocorrência de tal fenómeno, daí que ele tenha de considerar as quatro consequências da matriz acima, e não apenas as que dizem respeito à coluna da direita. Se o agente atribuir uma probabilidade de 0,7 à ocorrência de chuva, então a probabilidade do estado que consiste em não-chover será necessariamente igual a 1 - 0,7. Ou seja, o conjunto dos estados do mundo relevantes para um determinado problema de decisão deverá constituir uma partição: uma colecção de estados mutuamente exclusivos e conjuntamente exaustivos; ou seja, pelo menos um terá de ser verdadeiro e apenas um pode ser verdadeiro.

Estamos, assim, na posse dos dois conceitos fundamentais da teoria da utilidade esperada. Dada a *função de utilidade do agente* e a sua *distribuição de probabilidades*, é agora possível calcular a utilidade subjectiva esperada de cada uma das acções: multiplica-se a probabilidade da ocorrência de cada um dos estados do mundo pela utilidade de cada uma das consequências determinadas por esses estados, e somam-se os produtos obtidos nessas multiplicações. Esta média ponderada constitui a utilidade esperada (UE) da acção. Por exemplo, a UE de não levar o chapéu é -8,6; e a UE de levar o chapéu é 10,5. Segundo a teoria, o agente racional é aquele que leva consigo o chapéu-de-chuva.

Note-se que não se disse que o agente racional leva consigo uma gabardina. Isto porque, tal como o conjunto dos estados deve possuir certas características, o mesmo acontece com o conjunto das acções. Algumas características incontroversas das acções consistem, por exemplo, nas seguintes: têm de ser pelo menos duas, têm de ser executáveis, incompatíveis entre si e também conjuntamente exaustivas. Supõe-se, então, que vestir uma gabardina não constituía uma opção para o agente ou que não seria incompatível com as duas acções disponíveis.

Um ponto que é importante considerar previamente é o da justificação do princípio da maximização da utilidade esperada. Mais uma vez, verifica-se uma preocupação com o rigor matemático e as abordagens modernas, desde o início do séc. XX, têm consistido em tentativas de axiomatização da teoria. A ideia fundamental consiste em definir um conjunto de restrições estruturais, ou axiomas, os quais deverão regular as preferências

dos agentes por consequências cuja obtenção é incerta, e mostrar que um agente cujo comportamento de escolha obedeça a esses axiomas, agirá de uma maneira compatível com o princípio da maximização da utilidade esperada. Ou seja, um tal agente agirá atribuindo probabilidades numéricas a estados do mundo e valores de utilidade a consequências. A normatividade da teoria que promove o princípio em causa está, portanto, relacionada com a aceitação desses axiomas, a qual se encontra em disputa, e a qualificação de 'racional' ou 'irracional' refere-se ao respeito, por parte dos agentes, por esses axiomas.

A avaliação que aqui se fará do valor epistemológico da teoria focar-se-á, portanto, na questão da sua normatividade, e não no seu eventual valor descritivo. Não só a avaliação deste último ponto depende da observação de dados empíricos – o que se encontra fora do alcance deste estudo - como também é importante termos à nossa disposição regras de decisão com valor normativo, caso se verifique que o nosso comportamento fica aquém dos parâmetros mais básicos de racionalidade. Tendo em conta o tipo de justificação adoptado pelos seus principais proponentes, qualquer abordagem aos fundamentos da teoria necessita de considerar os argumentos a favor e contra esses parâmetros básicos ou axiomas. Nesta medida, embora as questões descritivas não constituam o enfoque principal, é importante considerar alguns casos em que o comportamento dos agentes não parece coadunar-se com esses axiomas e averiguar se existem, ainda assim, razões para se continuar a designá-los como racionais. Por outro lado, uma teoria da decisão que imponha normas dificilmente implementáveis por agentes comuns, e não apenas por agentes idealmente racionais, não se mostrará de grande utilidade.

Resta, apenas, especificar melhor os dois conceitos fundamentais aqui apresentados: utilidade e probabilidade. Quanto ao primeiro, como se referiu, o que está em causa é o quão desejável é um certo item do ponto-de-vista de um agente; ou seja, este conceito é sempre utilizado para referir uma perspectiva individual ou subjectiva. Segue-se daqui que a utilidade de um determinado item varia de agente para agente, consoante o valor numérico que cada um destes lhe atribui, valor esse que representará a magnitude relativa da sua desejabilidade.

Quanto ao conceito de probabilidade envolvido, este pode também ser interpretado de várias maneiras. No contexto da teoria da decisão que iremos tratar, as atribuições numéricas que representam as crenças do agente acerca da ocorrência dos estados do mundo resultam de uma avaliação subjectiva por parte desse mesmo agente. Podem

eventualmente existir dados empíricos importantes relacionados com a probabilidade dessas ocorrências – por exemplo, estatísticas que nos informam acerca da probabilidade objectiva de sucesso de uma cirurgia – e que o agente utilizará para definir as suas preferências. Contudo, mesmo quando esses dados não existem, o agente atribuirá uma estimativa subjectiva à possibilidade de ocorrência dos estados do mundo. Este é um procedimento natural que se encontra relacionado com a natureza incerta da maioria das nossas crenças, principalmente daquelas que se mostram relevantes na maioria dos nossos problemas de decisão. Como se verá, uma teoria normativa da decisão que faça uso de uma concepção objectiva de probabilidade irá impor exigências performativas demasiado complicadas para poderem ser levadas a cabo por agentes comuns. Quando uma determinada teoria da decisão emprega os conceitos de utilidade subjectiva e de probabilidade subjectiva, essa teoria é designada como uma teoria *bayesiana* da decisão. Os conceitos-chave apresentados nesta introdução são suficientes para se fazer uma abordagem informada aos conteúdos apresentados neste estudo. A perspectiva que se tenta aqui oferecer pretende ser não apenas detalhada, mas também panorâmica, cobrindo todo o período de desenvolvimento da teoria da decisão, desde os seus primórdios no séc. XVII, até aos seus desenvolvimentos mais recentes. Contudo, a exposição histórica tem apenas uma função introdutória e, mesmo esta, pretende-se dinâmica, desenvolvendo-se sempre através da consideração dos problemas que a teoria vai enfrentando.

O estudo encontra-se divido em três partes. Na primeira serão apresentados os fundamentos da teoria, desenvolvendo-se alguns conceitos-chave – como a construção da função de utilidade e da distribuição de probabilidades do agente – mas também, e mais importante, será analisada a justificação do poder normativo da teoria, sob a forma de uma discussão do método axiomático e dos argumentos a favor e contra os axiomas da preferência. Cada uma das secções da primeira parte é autónoma e pode ser lida separadamente, sem que com isso se perca a informação nelas contida.

Ao contrário do que acontece com o tratamento destas temáticas no contexto da ciência económica, em que os conceitos operativos da teoria são normalmente definidos em termos do comportamento observável dos agentes, pretende-se aqui abordar os conceitos de desejo e crença com o intuito de oferecer racionalizações e explicações para as causas desse mesmo comportamento. Considerando que nem sempre o comportamento de escolha oferece toda a evidência necessária para se efectuarem inferências acerca dos estados internos dos agentes, será ainda discutida na primeira parte a adequação de uma

interpretação behaviorista da teoria. A consideração de uma história e de uma justificação da teoria, e da interpretação dos seus conceitos fundamentais, é aquilo que permite atribuir uma perspectiva panorâmica ao estudo aqui desenvolvido.

Na segunda parte inicia-se a perspectiva detalhada da problemática fundamental que motivou este estudo. Será feita uma apresentação analítica do Problema de Newcomb (§4.1) e dos argumentos tradicionais a favor de cada uma das possíveis soluções (Robert Nozick 1969). Este problema destruiu quaisquer consensos que pudessem existir acerca da interpretação do princípio da maximização da utilidade esperada, especialmente em redor do princípio da maximização da utilidade *condicional* esperada, de Richard Jeffrey (1964). Esta interpretação recomendava que a probabilidade dos estados do mundo relevantes fosse condicionada à execução de cada uma das acções, permitindo assim dar conta das relações de dependência probabilística existentes entre estas e aqueles. Como a consideração dessa dependência permitiria averiguar que evidência poderiam as nossas acções oferecer-nos acerca da ocorrência dos estados do mundo, a teoria foi designada como *evidencial*.

A tentativa de resolução deste problema motivou o desenvolvimento das várias teorias *causais* da decisão. Estas teorias incorporam nexos de causalidade entre as acções e os estados do mundo no cálculo da utilidade esperada, permitindo assim discriminar entre os casos em que a dependência probabilística é apenas do tipo evidencial e os casos em que essa dependência pode ser também do tipo causal. A teoria causal apresentada será a de Gibbard e Harper (1978). De acordo com uma sugestão de Robert Stalnaker, esta teoria propõe que se utilize não a probabilidade condicional de estados do mundo, mas sim a probabilidade incondicional de proposições contrafactuais que dão conta do poder causal de acções sobre esses estados. Dada a importância das condicionais contrafactuais no contexto da teoria causal, a segunda parte do estudo consistirá em grande medida numa análise da semântica destas proposições. Será aqui desenvolvida a ideia de Terence Horgan (1981), segundo a qual existem duas interpretações possíveis dessas contrafactuais – standard e retroactiva - cada uma delas dando origem a uma ou outra das soluções para o problema, respectivamente, a solução bicaixista e a solução monocaixista. Favorecer-se-á, para efeitos do cálculo da probabilidade de contrafactuais, a interpretação retroactiva, e defender-se-á, para efeitos de determinação do seu valor de verdade, que uma semântica alternativa à de David Lewis (1973a), baseada na ideia de que estas proposições possuem um valor semântico intermédio - que consiste num grau de crença

condicional (Van Fraassen 1976) - pode fortalecer o argumento monocaixista e oferecer algumas razões para adoptarmos uma interpretação retroactiva/causal dessas contrafactuais, favorecendo-se desse modo uma solução monocaixista. O objectivo do argumento desenvolvido consistirá, portanto, em defender uma solução monocaixista no contexto da teoria causal.

A última secção da segunda parte aborda o Problema de Newcomb sob a perspectiva distinta da Teoria dos Jogos, mostrando-se aí que este problema é estruturalmente semelhante a um outro famoso puzzle da racionalidade prática, o Dilema do Prisioneiro. Argumentar-se-á a favor de uma solução cooperativa para o Dilema, equiparando-se depois essa solução à escolha monocaixista no Problema de Newcomb. Ao contrário do que acontece na primeira parte, a segunda tem uma progressão bem definida e a compreensão de cada secção depende da informação veiculada pelas secções anteriores.

Na terceira parte serão discutidos os problemas enfrentados pela teoria causal. Será apresentada uma reformulação da teoria evidencial que tem como objectivo permitir-lhe oferecer os resultados correctos para um certo tipo de problema de decisão, cuja solução apenas a teoria causal podia oferecer. Essa reformulação, designada por *ratificacionismo* (Jeffrey 1983), não se mostrará completamente isenta de problemas. Serão também oferecidos alguns contra-exemplos recentes à teoria causal, concebidos por Andy Egan (2007). Argumentar-se-á que esses contra-exemplos não são nocivos à teoria causal e que a adopção da estratégia ratificacionista, enquanto princípio de racionalidade prática, conduz a uma situação de instabilidade deliberativa que não é salutar no que respeita ao valor normativo da teoria. Essa situação de instabilidade resulta da adopção, proposta por Joyce (2012), do método da condicionalização para efeitos de actualização de crenças no decorrer do processo deliberativo. Propor-se-á, ao invés, a adopção para o mesmo efeito do método de David Lewis (1976), designado por ‘visualização’ [*imaging*]. Finalmente, oferecer-se-á uma versão do princípio da dominação que seja compatível com a possibilidade da interpretação retroactiva das contrafactuais envolvidas num problema de decisão.

Parte Um - Fundamentos

1. Teoria da utilidade esperada: uma história

1.1. Utilidade subjetiva esperada

Qualquer abordagem introdutória à teoria da utilidade esperada tem de tornar clara a origem da força normativa da teoria. Impõe-se, portanto, uma resposta à questão acerca do motivo pelo qual devemos maximizar a utilidade esperada. Uma das maneiras de alcançar essa compreensão consiste em seguir a história do desenvolvimento da teoria, nomeadamente, saber qual o tipo de problema para a resolução do qual a teoria foi criada e desenvolvida. Esses problemas, como sabemos, são de natureza instrumental, ou seja, queremos saber o que temos de fazer para alcançar um certo bem ou evitar um qualquer mal. Para cumprir este objectivo, de acordo com os autores da *Lógica de Port Royal* de 1662, é necessário considerar,

'não apenas o bem ou o mal em si mesmos, mas também a probabilidade com que estes poderão, ou não, ocorrer; e considerar, de uma forma geométrica, a proporção que estas coisas têm entre si quando tomadas em conjunto'.[1]

A ideia central das teorias da decisão modernas encontra-se aqui claramente expressa: perante um conjunto de bens cuja obtenção é incerta, a satisfação das nossas preferências deve ter em conta não apenas a magnitude relativa desses bens, mas também a probabilidade da qual depende a sua obtenção. Entenda-se aqui 'geometricamente' como referindo-se ao tipo de combinação que se deve verificar entre as referidas magnitudes e as referidas probabilidades, ou seja, de forma multiplicativa.

As bases intuitivas da teoria da probabilidade que se encontram subjancentes a este princípio foram descobertas por Blaise Pascal em 1654, no decorrer de uma correspondência com o matemático Pierre de Fermat, e na sequência da discussão do chamado 'problema dos pontos'. Um conhecido jogador, o Chevalier de Méré, colocou a Fermat o seguinte problema: considere-se um jogo com dois jogadores que consiste em

[1] Citação em Joyce (1999: 9) de *La logique, ou l'art de penser*. Esta obra consiste num extenso manual de lógica, publicado anonimamente em 1662, por Antoine Arnauld e Pierre Nicole, e para o qual contribuiu em larga parte Blaise Pascal.

lançar ao ar uma moeda não-viciada, no máximo cinco vezes, ganhando quem obtiver um maior número de *caras* ou *coroas*. Suponha-se que o resultado dos três primeiros lançamentos foi *cara/coroa/coroa*. Sabendo-se que a quantia total que pode ser ganha é, por exemplo, 50 Euros, e que o jogo tem de ser interrompido ao fim dos três lançamentos, como se deve repartir o dinheiro de forma justa entre os dois jogadores? Uma repartição equitativa parece prejudicar o jogador que apostou *coroas*, pois ele tem maior possibilidade de ganhar os 50 Euros; por outro lado, também não parece justo ele ficar com a totalidade do pote, pois o jogador que apostou *caras* ainda tem possibilidade, ainda que inferior, de ganhar. Para que uma repartição seja justa, o jogador que apostou *coroas* tem de receber um valor entre 25 e 50 Euros, e o jogador que aposto *caras* tem de receber um valor entre 0 e 25 Euros. Mas como determinar o valor exacto? Este é o 'problema dos pontos'.

Pascal compreendeu que o conjunto de todas as possibilidades ou formas de conclusão do jogo – e, por sinal, de todos os jogos – constitui uma partição de eventos mutuamente exclusivos e colectivamente exaustivos, ou seja, apenas um se pode verificar e pelo menos um tem de se verificar. Seja *E* um evento complexo que corresponde à serie dos dois lançamentos finais, *O* o evento simples que consiste em 'sair *coroas*', e *A* o evento simples que consiste em 'sair *caras*'. Uma partição óbvia é, por exemplo, a seguinte:

$$\{E_{AA,}\, E_{AO,}\, E_{OA,}\, E_{OO}\}.$$

Naturalmente, o jogador que apostou *caras* ganhará o pote se, e somente se, ocorrer o primeiro de entre estes quatro eventos. Por outro lado, para que o jogador que apostou *coroas* ganhe o pote, basta que ocorra um de entre os três eventos seguintes, ou seja, basta que saia *coroas* apenas uma vez.

Pascal transformará esta sua intuição no princípio fundamental da teoria da probabilidade da sua autoria:

Dado um conjunto $\{E_{1,}E_{2,}E_{3\ldots}E_n\}$ de possibilidades exaustivas e mutuamente exclusivas, a cada uma das quais pode ser atribuída uma determinada probabilidade, a soma de todas essas probilidades terá o valor 1; ou seja, $\Sigma_i\, p\,(E_i) = p\,(E_1) + p\,(E_2) + p\,(E_3) \ldots p\,(E_n) = 1$.

Daqui segue-se que a soma das probabilidades dos eventos da partição, que representa todas as possibilidades de conclusão do jogo dos pontos, terá também valor 1; ou seja,

$$(E_{AA} + E_{AO} + E_{OA} + E_{OO}) = 1.$$

Obtido este resultado, a solução para o problema dos pontos resulta da combinação deste princípio da probabilidade com a ideial central, acima mencionada, da teoria da decisão: a parte do pote a que cada um dos jogadores tem direito equivale, portanto, ao valor que resulta da combinação (geométrica/multiplicativa) entre os valores recebidos em cada um dos eventos da partição e a probabilidade de cada um desses mesmos eventos se verificar. Sabendo-se que cada um dos eventos da partição tem uma probabilidade ¼ de ocorrer (a moeda não está viciada), o valor a que tem direito o jogador que apostou coroas, que vencerá o jogo nos três últimos eventos da partição, calcula-se da seguinte maneira:

$$p\,(E_{AA}) \times 0 + p\,(E_{AO}) \times 50 + p\,(E_{OA}) \times 50 + p(E_{OO}) \times 50 = 37{,}5.$$

Já o jogador que apostou caras, que ganhará o jogo se, e somente se, sair caras nos dois últimos lançamentos da moeda, terá direito ao valor que é calculado da seguinte maneira:

$$p\,(E_{AA}) \times 50 + p\,(E_{AO}) \times 0 + p\,(E_{OA}) \times 0 + p(E_{OO}) \times 0 = 12{,}5.$$

Estes dois valores correspondem àquilo que se designa como o *resultado esperado* do jogo ou, neste caso, o *valor esperado* de uma aposta; ou seja, quando confrontado com um novo jogo que consiste em terminar o jogo original inacabado, um qualquer jogador pode escolher comprar uma aposta em como ganhará 50 Euros com uma probabilidade ¾ por 37,5 Euros, ou uma aposta em como ganhará 50 Euros com uma probabilidade de ¼ por 12,5 Euros. Como se verá, um jogador racional será indiferente entre receber 37,5 Euros ou aceitar a primeira aposta, e entre receber 12,5 Euros ou aceitar a segunda aposta. Pode-se constatar intuitivamente a correcção da solução de Pascal, ao considerar-se que a soma dos *valores esperados* de cada uma das apostas corresponde à totalidade do pote, ou seja, estamos perante um jogo de soma-zero, em que nada se perde e nada se ganha, para além do valor que um ou outro dos jogadores receberá, caso o jogo seja jogado até

ao fim. Além disso, a solução de Pascal implica também que, caso o jogo parasse ao fim de dois lançamentos, tendo saído uma vez caras e outra coroas, os jogadores deveriam dividir o pote entre si de forma equitativa.

A solução de Pascal, estabelecendo que as apostas devem ser avaliadas segundo o seu *valor esperado,* constitui o primeiro passo para a formalização das teorias da decisão modernas. À luz de conceitos posteriormente adquiridos, é possível resumir, de forma retrospectiva, o essencial das descobertas de Pascal:

'Se um jogador acredita que *p* é a função de probabilidade objectiva correcta para estados do mundo, e se o mesmo estiver interessado numa aposta *A*, apenas enquanto meio para aumentar a sua fortuna total, então ele deverá ser indiferente entre a perspectiva de aceitar *A* ou a perspectiva de receber um certo valor em dinheiro *E*, tal que *E* é igual ao valor esperado de *A*' (Joyce 1999: 18).

Dois aspectos deste princípio devem ser realçados. Em primeiro lugar, ainda se está longe de chegar a um dos aspectos necessários à elaboração de uma teoria bayesiana, ou seja, ainda não se encontra disponível uma interpretação subjectiva do conceito de probabilidade. É perfeitamente natural que os jogadores do séc. XVII fizessem as suas apostas com base numa compreensão, pelo menos implícita, da noção de frequência relativa, considerando empiricamente as proporções dos casos ou eventos favoráveis, como sair cara ou coroa, para a totalidade de eventos do mesmo tipo, como os vários lançamentos de uma moeda ao ar. Faz, portanto, sentido atribuir a Pascal, retrospectivamente, um princípio em que o conceito de probabilidade está a ser interpretado de modo objectivo, seja como frequência relativa, seja de acordo com a concepção clássica de Laplace[2].

[2] De acordo com a interpretação frequentista é possível testar empiricamente uma frase de atribuição de probabilidade observando a frequência relativa com que o evento alvo de atribuição – por ex. sair cara – ocorre numa longa série de observações do fenómeno – lançamentos – a que está associado e verificar se essa frequência relativa é igual ao valor da probabilidade atribuída. Esta ideia tem como base o pressuposto de que a frequência relativa do evento converge na direcção de um limite, quando a série de observações tende para infinito. A concepção clássica tem como base o Princípio da Indiferença: supondo que existem $n > 1$ possibilidades colectivamente exaustivas e mutuamente exclusivas, e que estas possibilidades são indistinguíveis entre si, a não ser pelo nome (ou enumeração, se quisermos), então deve ser atribuída a cada uma delas uma probabilidade de $1/n$. Este é um princípio *a priori*, de onde se segue que a concepção em causa, ao contrário da que tem por base a ideia de frequência relativa, não tem qualquer conteúdo empírico. Para uma discussão detalhada destas concepções, ver Zilhão (2012).

O segundo aspecto a considerar é a importante qualificação, de acordo com a qual o jogador está interessado nas apostas apenas como meios para aumentar a sua riqueza. O prazer que pode ser obtido pelo próprio acto de jogar não está a ser tido em conta na avaliação do *valor esperado* das apostas, nem, por exemplo, o prazer alcançado por uma eventual propensão para correr riscos elevados. A ser tido em conta, este factor teria de ser incluído, de alguma maneira, no cálculo do *valor esperado* das apostas; mas, dado que este valor está a ser equiparado àquilo que se designa como o *valor monetário esperado*, a esse prazer teria de ser atribuído um determinado valor em dinheiro, o que é pouco plausível, ou então outro critério teria de ser criado de modo a incluir a satisfação de jogar no *valor esperado* da aposta. O conceito de utilidade pode precisamente capturar o grau de satisfação, ou de felicidade, que um determinado bem, como uma certa quantia de dinheiro, pode trazer, satisfação essa que poderá, ou não, ser equiparada ao *valor monetário esperado*. Neste ponto estamos apenas a considerar o contexto idealizado dos jogos de azar, aos quais se aplicam naturalmente tanto as concepções objectivas de probabilidade, quanto a noção de *valor monetário esperado*, enquanto critério de avaliação da desejabilidade das apostas. Contudo, na vida real não é de todo plausível aplicar este critério como princípio orientador da nossa racionalidade instrumental. Mesmo quando estão em causa decisões relacionadas com o mundo dos negócios, em que as análises de custos e benefícios assumem enorme importância, a noção de *valor monetário esperado* pode mostrar-se inadequada; por exemplo, quando o esforço individual, ou colectivo, para assegurar uma quantia extra se torna demasiado oneroso, a utilidade de encetar esse esforço pode ser menor do que a utilidade de não fazer nada, mesmo que no primeiro caso a nossa fortuna seja maior; ou, por exemplo, quando uma empresa sacrifica lucros a curto prazo em prol de considerações humanitárias, sociais, educacionais, etc.

O conceito de utilidade não se encontrava, contudo, à disposição de Pascal. De modo a poder compreender-se a sua introdução na história do progresso da teoria da decisão, é necessário, em primeiro lugar, conhecer o resultado de um argumento favorável à tese de Pascal; ou seja, um argumento para mostrar que o preço de uma aposta é idêntico ao seu *valor monetário esperado*. Apesar de Pascal nunca o ter apresentado, é possível, de forma retrospectiva, apresentar esse argumento (Savage 1954; Kreps 1988). A sua estrutura é a seguinte: introduzindo-se um conjunto de premissas que estabelecem as características estruturais do conjunto de apostas disponíveis, características essas que regulam as

preferências de um agente racional por essas mesmas apostas, é possível construir uma função de utilidade única *u*, tal que o preço *€* de uma aposta *A* consiste na solução da equação $u(€) = U(A)$, consistindo $U(A)$ no *valor monetário esperado* de $u(€)$, calculado relativamente a uma função de probabilidade *p* para estados do mundo *E*, tal que, quando esses estados são enumeráveis, $U(A) = \Sigma_A\, u(A).p(E)$.

Ora, isto significa que a 'utilidade monetária esperada' de cada aposta (ou o seu *valor monetário esperado*) é idêntica à utilidade do seu preço. Já a utilidade do seu preço não é mais do que a utilidade da respectiva quantia de dinheiro. A função em causa consiste, portanto, numa função de utilidade para quantias de dinheiro. Portanto, um apostador racional, cujas preferências satisfaçam as premissas do argumento, estabelecerá os preços das apostas da seguinte maneira: 1) calculando a *utilidade monetária esperada* de cada aposta *A* e 2) encontrando um valor em dinheiro *€* cuja utilidade seja idêntica à $U(A)$.

Uma maneira de visualizar o resultado do argumento consiste em entender que a construção da função de utilidade *u* equivale à construção de uma escala de intervalos, na qual é medido o valor relativo das apostas. Os detalhes desta construção serão apresentados mais adiante, mas para já basta saber o seguinte: determinando-se arbitrariamente um certo valor 1 para a aposta 'mais preferida' e o valor 0 para a aposta 'menos preferida', as restantes apostas têm de se situar entre estes dois pontos da escala, tal que a sua posição seja relevante não só ordinalmente, mas também cardinalmente. Quando estamos a lidar com decisões em condições de incerteza, não basta saber que um determinado bem é preferido a outro, é necessário também saber se ele é suficientemente bom para que se corra o risco de procurar obtê-lo, daí que uma escala meramente ordinal não seja suficiente para nela basearmos os nossos cálculos de utilidade. Uma escala de intervalos mede, portanto, o tamanho relativo dos intervalos do ranking de preferências de um agente.

O sistema métrico e o sistema imperial de medidas são exemplos típicos de escalas de intervalos equivalentes, uma resultando da outra através da multiplicação por uma constante positiva, o que constitui um tipo específico de transformação linear positiva. Duas funções de utilidade *u* e *u'* representam a mesma ordenação de preferências de um agente se, e somente se, existem duas constantes *a* e *b*, tal que $u' = a.\, u+b$. A segunda consiste, assim, numa transformação linear positiva da primeira. É, portanto, neste sentido que estas funções de utilidade se classificam como únicas.

As escalas de temperatura Celsius e Farenheit são particularmente elucidativas. Temos de ter cuidado quando afirmamos que um corpo que esteja a 50ºC está duas vezes mais quente do que um corpo que esteja a 25ºC. Ainda que isto seja verdade para a escala Celsius, não faz sentido afirmar que um corpo que esteja a 122º F(50ºC) está duas vezes mais quente do que um corpo que esteja a 77ºF (25ºC). O que tem relevância cardinal é o *ratio* entre diferenças de temperaturas, o qual tem de ser idêntico para escalas equivalentes; ou seja, (50ºC – 25ºC)/ (25ºC – 0ºC) = 1, tal como (122ºF – 77ºF)/ (77ºF - 32ºF) = 1[3]. Do mesmo modo, afirmar que entre dois valores de utilidade $u(A)$ e $2u(A)$, o segundo é duas vezes mais desejável que o primeiro, é algo que apenas se pode afirmar no contexto de uma única função de utilidade. Em suma, existem várias funções compatíveis com a equação acima apresentada, $u(€) = U(A)$, ou várias escalas de intervalos equivalentes, todas elas transformações lineares positivas umas das outras, nas quais os *ratios* entre diferenças equivalentes são invariáveis. A conclusão do 'argumento de Pascal' pode ser apresentada da seguinte maneira: se um jogador satisfizer um conjunto de premissas que regulam a sua preferência relativamente a um conjunto de apostas, existirá, então, uma escala de intervalos única, e equivalentes transformadas, nas quais estará representado o ranking de preferências do agente por essas apostas.

De uma maneira retrospectiva, identificou-se o *valor monetário esperado* de uma aposta com a *utilidade* de uma certa quantia de dinheiro, mas a tese de Pascal continua na sua essência inalterada: o preço de uma aposta é equivalente ao seu valor monetário esperado. Devemos, contudo, questionar-nos se esta tese está correcta. Se em certos casos, tal como no problema dos pontos, ou nos jogos de azar em geral, esta tese surge como uma solução natural, existem outros em que o resultado da sua aplicação surge claramente como contra-intuitivo. Por exemplo, um jogador racional, de acordo com a tese, será indiferente entre aceitar uma aposta em que ganhará ou perderá, com a mesma probabilidade, um milhão de Euros, e não aceitar a aposta. Acredito que a maioria de nós preferirá não aceitar a aposta. Em termos modernos diríamos que a utilidade negativa de perder um milhão é muito maior do que a utilidade de ganha-lo. Contudo, alguém que já possui, digamos, vinte milhões, poderá aceitar a aposta sem qualquer hesitação. Dir-se-ia, portanto, que a utilidade de uma certa quantia de dinheiro varia, afinal, de indivíduo para indivíduo.

[3] Não confundir aqui escalas de intervalos com escalas de *ratios*. Embora a maioria defenda que a utilidade apenas pode ser medida numa intervala de intervalos, existem economistas e matemáticos que sustentam que é possível medi-la numa escala de *ratios*. Em última instância, o debate depende da nossa perspectiva metafísica acerca da natureza da utilidade. Para uma introdução a este debate, ver Peterson (2009: 106-10).

Considere-se um outro caso em que, de acordo com a tese, um jogador será indiferente entre as seguintes apostas, todas elas com o mesmo preço: ganhar um milhão com probabilidade 1/2, dois milhões com probabilidade 1/4, quatro milhões com probabilidade 1/8, oito milhões com probabilidade 1/16, etc. Arrisco-me a afirmar que a maioria de nós preferirá tentar ganhar um milhão com probabilidade 1/2, a tentar ganhar, pelo menos, dezasseis milhões com probabilidade 1/32, ou trinta e dois milhões com probabilidade 1/64, supondo que o preço da aposta é idêntico para todos os casos. Estes resultados sugerem fortemente que a tese de Pascal poderá estar errada.

Uma maneira de colocar o problema de forma mais dramática passa por considerar o chamado 'paradoxo de St Petersburg', descoberto no séc. XVIII pelo matemático suíço Daniel Bernoulli. Não se trata estritamente de um paradoxo em sentido lógico, pois não se deriva uma contradição da aplicação da tese de Pascal; contudo, da análise do problema, segue-se uma recomendação que, tal como nos exemplos acima, nos surge como extremamente contra-intuitva. O paradoxo resulta do homónimo jogo de St. Petersburgo. Uma moeda não-viciada é lançada ao ar até sair caras; o jogador ganhará um prémio de valor 2^n, em que *n* corresponde ao número de lançamentos realizados até sair caras. O valor do prémio pode ser medido tanto em quantias de dinheiro, como em unidades de utilidade. Por exemplo, o jogador ganhará trinta e dois Euros se a moeda for lançada cinco vezes antes de sair caras. Para que a banca possa lucrar, o preço do jogo terá sempre de ser superior a dois Euros. A pergunta que se impõe é a de saber quão superior é a quantia de dinheiro que estamos dispostos a pagar para jogar, ou seja, qual é *o preço justo desta aposta*. Se este for idêntico, de acordo com a tese de Pascal, ao valor esperado do jogo, então, na suposição de que a banca tem fundos ilimitados, deveríamos estar dispostos a pagar uma quantia infinita para jogar. Considere-se o cálculo do *valor esperado* do jogo:

$$\sum_{n=1}^{\infty} \quad (1/2)^n \cdot 2^n = \infty.$$

Este resultado é claramente absurdo. Um jogador teria, por exemplo, de estar disposto a pagar um milhão de Euros para jogar, ainda que a moeda tivesse de calhar coroas dezanove vezes seguidas para o prémio superar esse valor.

Existem várias propostas de solução para este paradoxo, algumas delas bastante semelhantes entre si. Por exemplo, em 1745, e em resposta a Bernoulli, Buffon propôs que, para efeitos práticos, simplesmente se ignorassem resultados altamente improváveis,

do tipo 'ser atingido por um cometa' ou 'sair coroas vinte vezes seguidas'.[4] Muito mais recente é a proposta do Prémio Nobel da Economia, Kenneth Arrow (1970), para quem as preferências de um agente devem ser representadas por uma função de utilidade limitada ou finita, ou seja, para além de um determinado valor qualquer acrescento de um certo bem não implica um equivalente aumento de utilidade.[5] Isto implica, claro, que a utilidade, ou *valor monetário esperado*, do jogo de St. Petersburg é finito.

Mas uma das soluções mais plausíveis do paradoxo foi descoberta por Cramer, um matemático também do séc. XVIII, o qual sugeriu uma hipótese com enorme poder explicativo no campo da economia: que o dinheiro tem utilidade marginal decrescente. O significado desta expressão será melhor compreendido através de um exemplo do que através de uma definição formal. Suponha-se que acrescentos de mil Euros vão sendo oferecidos a um agente que começa com os bolsos vazios. Os primeiros mil Euros terão para esse agente uma determinada utilidade, digamos 1000 utiles; um segundo acrescento de mil Euros, que resultam na acumulação de uma fortuna duas vezes superior, já não terá para esse agente a mesma utilidade que o primeiro, mas sim, digamos, 990 utiles; um terceiro acrescento de mil Euros, perfazendo uma fortuna três vezes superior à inicial, já não terá para esse agente a mesma utilidade que o segundo acrescento, e muito menos o terceiro, mas sim, digamos, 970 utiles. Continuando da mesma maneira, cada novo acrescento gerando cada vez menor utilidade, chegar-se-á a um ponto em que um novo acrescento gerará virtualmente zero utiles para o agente. Se visualizarmos o gráfico de uma qualquer função de utilidade que represente o fenómeno da utilidade marginal decrescente, constataremos que será descrita uma curva côncava, com declive descendente à medida que cresce, até que, eventualmente, se transformará numa linha recta, como no caso do jogo de St. Petersburg. Cramer sugeriu um valor estimativo de mais ou menos dez milhões ou 2^{24} ducados, como o limite acima do qual qualquer acrescento monetário não implicaria qualquer aumento de utilidade (em Bernoulli 1738).

---

[4] Na análise de riscos utiliza-se um princípio chamado *de minimis*, que em substância equivale à proposta de Buffon. Por exemplo, uma seguradora que vende um seguro de vida ignora a probabilidade de o cliente ser atropelado ao sair da agência.

[5] Quando uma função é finita existe um número real M, tal que $|f(x)| \leq M$. Neste caso a função é limitada 'por cima'. Mas também existem funções limitadas 'por baixo', se $f(x) \geq M$, para todo o x *em* X. A meu ver, é perfeitamente plausível que a função de utilidade de um agente seja limitada por cima; não creio, contudo, que essa função deva ser limitada por baixo, admitindo a hipótese de que pode existir uma desutilidade infinita. Uma determinada aposta, ou acção disponível, pode ter como consequência possível a morte do agente ou, por exemplo, de um seu familiar. Em ambos os casos, a desutilidade da morte poderá ser tal que, por maior que seja a utilidade de outra consequência possível dessa acção, o agente nunca decidirá executá-la. Isto não impede, claro, que existam agentes, ou apostadores, para quem a possibilidade de ficar rico, por exemplo, supera em muito a desutilidade que atribuem ao estado de morte.

A melhor descrição do fenómeno, ou pelo menos a mais poética, é apresentada por Joyce (1999, p. 33):

'O dinheiro, contrariamente ao que diz o velho adágio, pode comprar a felicidade (…), contudo, a taxa de câmbio entre os dois não é constante. O "valor de felicidade" de um dado dólar depende de quantos outros restarão depois de ser gasto; o dinheiro vale sempre mais para um pobre do que para um príncipe. Ao identificar os preços justos com os valores esperadas, Pascal estava implicitamente a negar este facto. Do seu ponto-de-vista, o valor de um dólar extra é invariável de jogador profissional para jogador profissional, *independentemente do quão grande ou pequena seja a fortuna de cada um*'.[6]

No jogo de St. Petersburg, o erro da tese da tese de Pascal consiste em implicar que um certo resultado 2^{n+1} é duas vezes mais desejável do que um resultado 2^n, ou que 2^{n+19} é duas vezes mais desejável do que 2^{n+18}. De acordo com a hipótese da utilidade marginal descrescente, a utilidade que 2^{n+1} nos permite obter é sempre menor do que o dobro da utilidade de 2^n. Quando *n* tem um valor pouco elevado, o acréscimo de utilidade que 2^{n+1} permite obter é pouco inferior ao dobro de 2^n, mas quando chegamos, por exemplo, a 2^{n+19} o acréscimo de utilidade que este resultado acarreta é muito inferior ao dobro da utilidade de 2^{n+18}.

Esta é uma hipótese a respeito da psicologia dos jogadores ou dos agentes em geral. Como essa psicologia é variável, um Euro a mais para um ganacioso terá sempre, dentro de limites, um acréscimo de utilidade idêntico, comece ele com mil ou com um milhão; assim como, para um avaro, perder um Euro é quase o mesmo do que perder mil. Uma boa maneira de compreender o fenómeno da utilidade marginal decrescente consiste em compará-lo com o fenómeno de aversão ao risco. Estes dois conceitos, embora distintos, encontram-se relacionados, pois se um determinado agente tiver aversão ao risco, isso significa que para ele o dinheiro tem uma utilidade marginal decrescente. Ou seja, se o risco que esse agente está disposto a correr para ganhar uma determinada quantia não é o mesmo que ele se encontra disposto a correr para, por exemplo, ganhar em seguida o dobro dessa mesma quantia, então podemos daí concluir que para ele o dobro tem uma utilidade menor do que a sua metade vezes dois.

[6] A expressão 'jogador profissional' remete para o contexto em que estas teorias estavam na altura a ser discutidas, e refere-se simplesmente a alguém cuja única procupação é maximizar os seus ganhos e minimizar a suas perdas.

Este é um resultado de modo algum vago e que tem uma expressão técnica bastante clara. Ao contrário do jogador profissional de Pascal, que era indiferente entre o *status quo* e perder ou ganhar um milhão com a mesma probabilidade, o agente de Cramer e Bernoulli não estará, de modo algum, disposto a correr aquilo que a maioria de nós designaria como um risco irreflectido. Por exemplo, o agente com aversão ao risco preferirá sempre mil Euros certos a uma chance em duas de ganhar dois mil ou zero, seguindo-se daí que, para ele, a diferença de utilidade entre mil e zero é superior à que existe entre dois mil e mil. Para que a aposta se torne atractiva para o agente, a probabilidade de ganhar dois mil tem de ser grande o suficiente, de modo a compensar a redução do aumento de utilidade de mil para dois mil. É claro que isto não se aplica apenas a valores monetários: um agente com aversão ao risco, interessado em laranjas, preferirá sempre receber três laranjas a uma aposta cujos prémios consistem em ganhar seis ou zero laranjas com probabilidades idênticas.

O contraste entre a função de utilidade de um agente com aversão ao risco e a função de utilidade de um jogador profissional, que resulta do argumento a favor da tese de Pascal, é notória. Enquanto a primeira, como vimos, forma uma linha côncava, a função de Pascal forma uma linha recta que não pára de crescer. Assim, o argumento para a tese de Pascal incluirá necessariamente uma premissa que dá conta da atitude do jogador perante o risco, neste caso uma atitude de neutralidade perante o mesmo. Isto também pode ser dito de uma outra maneira: para Pascal o dinheiro tem uma utilidade marginal constante. Portanto, para o agente com aversão ao risco, o preço de uma aposta será sempre inferior ao valor esperado da mesma; contudo, para o agente que encontra felicidade na simples procura do risco, o preço de uma aposta é superior ao seu valor esperado.

Embora a aversão ao risco seja uma característica natural da grande maioria dos agentes, o seu grau varia de uns para outros. Do mesmo modo, consoante o valor que cada um atribui ao dinheiro, a sua função de utilidade para este item terá um declive maior ou menor; mais precisamente, quanto maior for o declive, maior será a aversão ao risco. (Afinal há quem goste mais de dinheiro do que outros). Nesta medida, não existe uma função única de utilidade para dinheiro que se aplique a todas as pessoas indiferentemente. Este facto é de extrema importância, dada a conclusão a que pretendemos chegar, nomeadamente, que o preço de uma aposta equivale à *sua utilidade subjectiva esperada*, e já não ao seu *valor monetário esperado*, como era o caso na teoria de Pascal. A teoria da decisão não pretende ditar normas quanto ao gosto por dinheiro dos agentes ou quanto

às suas preferências em geral. Estas constituem apenas um dos dados do problema, o qual consiste em saber como é possível maximizar a sua utilidade *subjectiva* esperada.

A função de utilidade *u* representa os desejos de um agente por determinados bens, incluindo apostas e quantias em dinheiro. Quando esses desejos obedecem a certas restrições estruturais, o agente faz as suas escolhas maximizando a sua utilidade subjectiva esperada. Essas restrições funcionam como critérios de racionalidade instrumental, daí que se considere racional o agente que age de modo a maximizar a sua utilidade subjectiva esperada. Os proponentes da moderna teoria da decisão apresentam essas restrições sob a forma de axiomas, os quais devem ser satisfeitos pelo agente aquando da formação do seu ranking ou ordenação de preferências.[7] Isto significa que os desejos do agente podem ser modelados através de uma ordenação de preferências que estabelecerá a desejabilidade comparativa de qualquer par de apostas à disposição.

Assim como para a defesa da tese de Pascal foi possível apresentar um argumento, em que se demonstrava um teorema a partir da aplicação de uma série de restrições estruturais sobre as apostas de um jogador profissional, também agora será possível apresentar um argumento para a defesa da teoria da utilidade subjectiva esperada, em que se demonstra também um teorema a partir de uma série de axiomas ou restrições estruturais sobre as preferências dos agentes em geral. Uma apresentação desse argumento é a de John von Neumann e Oskar Morgenstern no clássico *Theory of Games and Economic Behavior* (1953)[8]:

Se um ranking de preferências satisfizer um certo conjunto de axiomas, então existe uma função de utilidade u para apostas A, cujo operador associado $U(A) = \Sigma_A\, u(A).p(E)$ *representa de forma ordinal essas preferências, na medida em que A é preferida a A' se, e somente se* $U(A) > U(A')$.

[7] Esta ordenação pode ser completa ou incompleta, disso dependendo a força normativa do argumento a favor da hipótese da utilidade esperada (acerca da existência da função *u*); essa ordenação será completa quando entre dois itens, o agente prefere um ao outro ou é indiferente entre elas; e incompleta quando tal não se aplica. Que a ordenação tem de ser completa é um dos axiomas referidos. A questão da sua satisfação é importante e complexa e será tratada mais adiante.

[8] Na verdade, foi Frank Ramsey, um filósofo e matemático de Cambridge, contemporâneo de Bertrand Russel, quem primeiro apresentou um argumento para o mesmo efeito, no seu 'Truth and Probability', publicado postumamente em 1931. Von Neumann e Morgenstern não conheciam o trabalho de Ramsey antes de apresentarem a sua defesa da teoria da utilidade esperada. O tratamento da utilidade de Ramsey é, de certa foma, mais complexo, pois encontra-se interconectado com a sua teoria subjectiva da probabilidade. Já o método de von Neumann e Morgenstern para a construção da função de utilidade faz uso de probabilidades objectivas. Esse método será apresentado mais adiante.

Este argumento tem a seguinte interpretação: se acreditarmos que os axiomas da preferência constituem verdadeiros requisitos de racionalidade instrumental, então temos também de acreditar que existe uma função de utilidade, com relevância cardinal, que mede a força relativa dos desejos de um agente, o qual ordena as suas preferências satisfazendo esses axiomas. Como uma aposta *A* é preferida a outra *A'* se, e somente se, a utilidade de *A* for superior à utilidade de *A'*, segue-se daí, naturalmente, o princípio fundamental da teoria da utilidade subjectiva esperada: um agente racional é aquele que faz as suas escolhas *como se* estivesse a maximizar a sua utilidade subjectiva esperada. A interpretação deste *como se* está relacionada com a discussão acerca da natureza descritiva da teoria e será tratada mais adiante.

Resumindo, a moderna teoria da utilidade esperada é herdeira da teoria da expectativa matemática, desenvolvida pelos primeiros probabilistas, como Pascal, no contexto específico dos jogos de azar; tal teoria recomendava que em determinadas escolhas o agente racional deveria maximizar o *valor monetário esperado* de uma aposta. Vimos que, aplicando retroactivamente o conceito mais tardio de utilidade, o valor esperado de uma aposta, de acordo com a teoria da expectativa matemática, era equivalente à utilidade do seu preço ou, o que é mesmo, da respectiva quantia de dinheiro. Contudo, exemplos como o jogo de St. Petersburg mostram que esta teoria leva a conclusões absurdas ou inaceitáveis face àquela que é a realidade comportamental de agentes que aceitamos como racionais. Foi por este motivo que Bernoulli introduziu o conceito de *valor moral*, tentando capturar a ideia de que o valor que atribuímos a um determinado bem, incluindo apostas e quantias de dinheiro, não é necessariamente equivalente ao valor monetário objectivo. O factor psicológico que se encontra por detrás desta ideia é o mesmo que está na base da hipótese de Cramer, segundo a qual para a maioria de nós o dinheiro tem uma utilidade marginal decrescente. Sendo o conceito de *valor moral* equivalente ao de *utilidade subjectiva*, temos, assim, a definição de racionalidade associada à teoria da decisão moderna: maximizar a *utilidade subjectiva esperada*.

O contexto desta história da teoria da decisão tem sido o dos jogos de azar. Tem-se falado principalmente de jogadores profissionais que tentam maximizar os seus ganhos. De acordo com o que até agora se considerou, o que torna a teoria da utilidade esperada (na versão de von Neumann e Morgenstern) particularmente apropriada para lidar com essas situações, é o facto de as probabilidades objectivas dos estados do mundo serem conhecidas. Quando se lançam dados, quando se atiram moedas ao ar ou se faz rodar a

roleta, as probabilidades das consequências são objectivamente conhecidas e podem substituir imediatamente as respectivas variáveis *E* no cálculo da utilidade esperada.

Até agora tem-se dado atenção apenas a um dos dois factores da equação: os valores da função *u*. Contudo, o que a teoria nos diz é que o agente atribui às consequências não apenas valores de utilidade, mas também probabilidades numéricas. Mas como poderá o agente fazê-lo? Ou seja, quando se trata de decidir coisas como, por exemplo, que seguro de vida subscrever, por que curso universitário enveredar, ou se, face ao risco de humilhação, vale a pena convidar uma certa pessoa para sair. A teoria tem, portanto, de se poder aplicar a tomadas de decisão reais, no mundo real fora dos casinos, em condições de incerteza. Deste modo, tal como a função de utilidade representa os desejos do agente, a sua função de probabilidade tem de poder representar as suas crenças, incluindo as crenças acerca da ocorrência de coisas como a possibilidade de uma morte precoce, a capacidade para compreender certos conceitos filosóficos ou a rejeição de um convite para jantar. Em suma, falta introduzir nesta história a concepção subjectiva de probabilidade, da qual advém a qualificação da moderna teoria da decisão como bayesiana

1.2. Probabilidade subjetiva

O bayesianismo é a perspectiva segundo a qual a noção de probabilidade pode ser interpretada de maneira subjectiva, especificando o grau de crença de um agente numa determinada proposição; ou seja, a atribuição de probabilidades consiste numa forma de medir os graus de crença do agente. Recusa-se, assim, a ideia de que as probabilidades se referem a propriedades objectivas das coisas ou a fenómenos físicos observáveis. A teoria bayesiana da confirmação, por exemplo, diz-nos que novo grau de crença é racional adoptar quando obtemos novos dados observacionais que confirmam as nossas hipóteses, tendo em conta os nossos graus de crença anteriores. Trata-se, pois, de uma teoria que nos permite atribuir um valor quantitativo preciso ao modo como um determinado corpo de informação observacional serve para confirmar racionalmente as nossas hipóteses científicas.

Frank Ramsey (1931) foi um dos primeiros a construir uma teoria subjectiva da probabilidade. Ao constatar a estreita conexão entre crenças, desejos e acções, Ramsey

compreendeu também o seguinte: se, por exemplo, estivermos dispostos a apostar em *Brigadier Gerard* para ganhar a corrida, então também estamos dispostos a apostar na verdade da crença segundo a qual *Brigadier Gerard* vai ganhar a corrida. Se estivermos dispostos a apostar numa hipótese de 8/10 em *Brigadier Gerard*, então o nosso grau de crença em como ele irá ganhar é de 0.8 e, respectivamente, 0.2 em como ele irá perder.

Um céptico poderá argumentar que uma teoria da probabilidade baseada em meros palpites não pode ser aplicada com seriedade ao tratamento de questões científicas. Contudo, De Finetti (1964) conseguiu provar que se as nossas atribuições subjectivas de probabilidade forem coerentes entre si, então elas encontram-se de acordo com os axiomas do cálculo de probabilidades. Mais, também pôde ser demonstrado que, se as nossas atribuições subjectivas de probabilidade estiverem de acordo com o cálculo de probabilidades, então elas têm de ser coerentes. Logo, a conformidade com os axiomas do cálculo de probabilidades é uma condição necessária e suficiente para a coerência subjectiva das nossas atribuições de probabilidade ou, o que é o mesmo, para a coerência dos nossos graus de crença. Uma das ideias fundamentais do bayesianismo consiste, assim, na determinação de um importante critério de racionalidade para agentes e detentores de crenças em geral, que permite classificá-los como coerentes ou racionais, caso as suas crenças exemplifiquem a referida conformidade.

Mas o céptico pode ainda insistir: 'E se sairmos da pista de corridas sem nada nos bolsos? Os nossos palpites, apesar de coerentes, estavam todos errados. Não existindo qualquer base empírica que os sustente, a coerência das nossas atribuições subjectivas de probabilidade não parece ser um critério de racionalidade suficientemente forte e restritivo para poder ser aplicado tanto à investigação científica como à tomada de decisões'.

Um dos proponentes da teoria frequentista da probabilidade, Richard von Mises (1936), expressa de forma clara esta ideia de imprecisão ligada à atribuição de probabilidades por parte do indivíduo não treinado. Para von Mises, a teoria das probabilidades tratar-se-ia de uma ciência natural, e as nossas intuições ou estimativas subjectivas de probabilidades nada teriam que ver com o conhecimento quantitativamento rigoroso do limite da frequência relativa de um certo evento numa sequência de eventos repetitivos. Esses ditos palpites seriam o equivalente às nossas estimativas subjectivas acerca da temperatura ambiente, ou à avaliação de distâncias 'a olho', quando comparadas com as medições rigorosas produzidas por um termómetro ou por receptores GPS.

Contudo, em primeiro lugar, o conhecimento decorrido da avaliação de estatísticas não elimina completamente o elemento de incerteza relacionado com as nossas previsões, as quais são parte essencial de qualquer processo de tomada de decisões. Segundo, uma das motivações para desenvolver uma teoria subjectiva da probabilidade consistiu, precisamente, na constatação da impossibilidade de dar conta, se formos fiéis a uma interpretação frequentista, da ideia de que é razoável atribuir probabilidades a acontecimentos únicos e irrepetíveis, entre os quais se podem encontrar certas consequências observáveis das teorias científicas. Será, pois, a aplicação regrada do Teorema de Bayes que irá disciplinar a atribuição de probabilidades subjectivas para efeitos de confirmação de hipóteses científicas, tarefa para a qual a interpretação frequentista não oferecia um método. Consideremos a formulação seguinte do Teorema de Bayes:

$$P(H \mid E) = \frac{P(H)\,P(E \mid H)}{P(E)}$$

em que $P(H|E)$ se lê como 'probabilidade da hipótese dada a evidência'; $P(H)$ como 'probabilidade prévia da hipótese'; $P(E|H)$ como 'probabilidade da evidência dada a hipótese'; e $P(E)$ como 'probabilidade prévia da evidência'. Assim, uma certa hipótese é confirmada por uma certa evidência se, e somente se, $P(H|E) > P(H)$; e uma hipótese é infirmada por uma certa evidência se, e somente se, $P(H|E) < P(H)$.

Na posse dos termos fundamentais, podemos agora afirmar que o uso regrado do teorema irá disciplinar a atribuição de probabilidades prévias às nossas hipóteses científicas e, desse modo, permitir-nos responder ao céptico. Considere-se um exemplo utilizado para mostrar como funciona o processo de confirmação (Zilhão 2010a: 329). Uma das consequências que a Teoria da Relatividade Geral de Einstein previa era a de que a trajectória da luz encurvaria perto do Sol. Em 1919 foi realmente observado este fenómeno. Podem-se, assim, atribuir à hipótese (H) e à evidência (E) valores de probabilidade que nos permitem aplicar o teorema de Bayes. Por exemplo, dado o carácter controverso, à data, da teoria de Einstein, podemos talvez dizer que $P(H) = 0.2$; como a evidência constitui aqui uma consequência lógica da hipótese, então $P(E) \geq P(H)$, daí $P(E) = 0.3$; como a evidência (E) se segue dedutivamente da hipótese (H), então $P(E|H) = 1$. Substituindo estes dados no teorema resulta que $P(H|E) = 2/3$.

O que este resultado mostra é que, a partir da observação realizada, houve um incremento de probabilidade da hipótese ($P(H|E) > P(H)$). Ou seja, de um grau subjectivo de crença de 0.2 na teoria de Einstein, passou-se para um grau subjectivo de crença de 0.66. A confiança em como a teoria seria verdadeira saiu bastante reforçada depois da observação de uma das suas consequências previstas, o que corresponde a uma confirmação da hipótese. É claro que se se tivesse observado que o fenómeno previsto não se verificava, a evidência recolhida estaria em contradição com uma das consequências dedutivas da teoria e, portanto, $P(E|H)$ teria de ser igual a zero. Aplicando este resultado ao teorema, teríamos que $P(H|E) = 0$. Neste caso a hipótese teria sido infirmada ($P(H|E) < P(H)$).

O processo de confirmação pode ser levado a cabo aplicando-se a regra da condicionalização: o valor obtido, a probabilidade da hipótese dada a observação, irá substituir o valor da probabilidade prévia da hipótese em caso de nova aplicação do teorema, dada a descoberta de novos dados observacionais (confirmatórios ou infirmatórios). Ou seja, nessa nova aplicação do teorema, o valor prévio da hipótese seria de 0.66. O aspecto ampliativo da teoria consiste, assim, em esperar que o valor prévio da hipótese vá aumentando progressivamente, confirmando cada vez mais a hipótese, convergindo para o valor 1.

A regra da condicionalização é, ela própria, uma maneira de responder a um dos desafios enfrentados pela teoria: o problema da suposta arbitrariedade das atribuições de um valor à probabilidade prévia da hipótese. Ou seja, dois indivíduos poderão atribuir probabilidades completamente diferentes à hipótese inicial e ambos terem uma justificação racional para acreditarem nessa hipótese em determinado grau. A resposta mais comum consiste em apelar para a noção de convergência. Mesmo começando com probabilidades 'heréticas', será possível modificá-las à medida que a nova informação se vai acumulando, através da aplicação sucessiva da regra da condicionalização (Jeffrey 1964). Ou seja, a aplicação da regra fará com que as probabilidades iniciais do perito e do leigo se aproximem progressivamente até convergirem num determinado valor que corresponderia ao grau de crença que é racionalmente mais adequado. No caso do apostador inexperiente, ele pode mesmo obter informação preciosa na enorme colecção de dados que os seus colegas frequentistas vão recolhendo. Mas, se nos casos em que a informação vai surgindo com frequência, e é de tipo quantitativo, a perspectiva de convergência é favorável, já nos casos em que é difícil obter novos dados, e a prova é de um tipo mais impreciso, a convergência pode tornar-se num processo bastante demorado

ou até num objectivo utópico. Em suma, a teoria subjectiva da probabilidade não é perfeita, apresentando também os seus limites.

Concluindo, a utilização ou tratamento da probabilidade subjectiva, no contexto da elaboração de complexas teorias de decisão, tem como referência a obra de Ramsey (1931), de Leonard Savage, *The Foundations of Statistics* (1954), e de Richard Jeffrey, *The Logic of Decision* (1964). O método utilizado para determinar essas probabilidades tem como base a ideia de que os nossos graus de crença estão directamente relacionados com o nosso comportamento observável, nomeadamente, com as preferências que revelamos nas nossas escolhas por consequências cuja obtenção é incerta.

A comparação entre os respectivos méritos e insuficiências destas duas teorias não cabe nos limites deste trabalho, contudo algumas observações poderão ser úteis. Enquanto Savage faz uma distinção entre, por um lado, os objectos do desejo instrumental e não-instrumental – respectivamente, acções e consequências - e os objectos de crença por outro – estados do mundo – a teoria de Jeffrey utilizará proposições que descrevem estados de coisas como objecto tanto dos desejos, como das crenças do agente. Como tal, enquanto em Savage a distribuição de probabilidades do agente se faz sobre um domínio de eventos, em Jeffrey essa distribuição faz-se sobre um domínio de proposições, o que implica, neste último caso, uma redefinição do cálculo de probabilidades para as funções de verdade da lógica proposicional. Uma das vantagens da utilização de proposições prende-se com a forma como é constituída a partição de acções à disposição do agente: em vez de definidas como funções de *estados* para *consequências*, as acções passam simplesmente a constituir proposições que o agente pode escolher tornar verdadeiras em determinada circunstância. Isto permite ao agente encarar os problemas de decisão que tem pela frente de uma forma mais realista, o que é importante quando pretendemos tratar de decisões no *mundo real*.

O fulcro da teoria de Savage é semelhante ao de von Neumann e Morgenstern: impor restrições estruturais às preferências do agente e demonstrar que caso um agente satisfaça esses axiomas, então agirá *naturalmente* como se estivesse a maximizar a sua utilidade subjectiva esperada. A diferença entre os dois métodos consiste no facto de o método de von Neumann/Morgenstern pressupor que as probabilidades dos resultados das apostas ou lotarias são objectivas. O teorema da *representação* de Savage garante a existência de uma função de probabilidade que *representa* formalmente os graus de crença do agente na obtenção de certas consequências, e também a existência de uma função de utilidade

única na qual é estabelecida a força relativa dos seus desejos por essas mesmas consequências.[9]

Existe também um teorema da representação para a teoria de Jeffrey, embora, neste caso, as conclusões não são tão fortes quanto alguns desejariam, pois não é possível garantir que as funções de probabilidade e utilidade de um único agente sejam únicas. Isto pode constituir um problema para quem acredita, sob um pressuposto comportamentalista, usualmente partilhado por bastantes economistas, que a única maneira de determinar as crenças de um agente é olhando para as suas preferências ou, o que para estes consiste no mesmo, para o seu comportamento de escolha, o qual nos é dado por simples observação.

2. Explicação e behaviorismo

2.1. O silogismo prático

Todos estão de acordo com a distinção da teoria da utilidade esperada entre *descritiva* e *normativa*. É possível defender que a teoria *descreve* de forma verdadeira o modo como os agentes agem em situações de incerteza. Nessa medida, a teoria descreveria factos acerca do comportamento dos agentes e, como tal, a determinação desse valor descritivo seria uma tarefa empírica, a qual consistiria em averiguar se os agentes satisfazem os axiomas da preferência. É também possível defender que, apesar de a teoria não descrever o modo de agir dos seres humanos, os seus princípios devem, contudo, ser aplicados, caso os agentes desejem agir racionalmente. Existem argumentos cujo objectivo consiste, precisamente, em mostrar de que modo os agentes podem sair prejudicados caso não sigam os princípios da teoria.

Os dois pontos-de-vista não têm, contudo, de ser incompatíveis. É razoável supor que, na maioria das vezes, as pessoas agem racionalmente de modo a satisfazer as suas preferências, maximizando a sua utilidade subjectiva esperada e respeitando os axiomas da teoria. Seria difícil conciliar a negação deste pressuposto com a observação de que a maioria de nós consegue levar vidas razoavelmente bem-sucedidas ou, em geral, com a

[9] Martin Peterson (2009: 147-151) oferece uma interpretação conceptual muito acessível dos seis axiomas de Savage e remete para uma versão menos complexa da demonstração em Kreps (1988). Muito esclarecedor é também o artigo da Stanford Encyclopedia of Philosofy sobre teoria da decisão (Katie Steele e H. Orri Stefánsson (2016: 19-39).

observação do simples facto de estarmos vivos. Contudo, a questão normativa parece ser prioritária, pois caso se descubra que as pessoas agem ocasionalmente de forma irracional, convém termos em nossa posse os princípios através dos quais as possamos ajudar a alterar o seu comportamento. Mas, seja qual for a nossa prioridade, temos sempre de garantir que as nossas hipóteses, descritivas ou normativas, tenham realmente valor epistemológico, permitindo-nos formular leis gerais acerca do comportamento humano, real ou desejável.

Os meados do século XX foram um período dourado no desenvolvimento da teoria da decisão. Foi nele que, como se viu, Savage publicou *The Foundations of Statistics* (1954), obra na qual, depois de Ramsey (1931) e von Neumman e Morgenstern (1953), foi apresentada uma outra justificação axiomática do princípio da maximização da utilidade esperada. Neste período fazia-se sentir a influência do positivismo lógico no modo como se entendia que devia funcionar a metodologia da análise científica do comportamento humano. Portanto, seja a relação de preferência definida para eventos ou proposições, essa relação pode ser interpretada de uma forma mentalista, atribuindo ao agente desejos enquanto estados mentais, ou de uma forma behaviorista, tratando essas preferências apenas como disposições para agir desta ou daquela maneira, disposições essas manifestadas pelo seu comportamento de escolha. A questão que se nos coloca é, portanto, a de saber se uma interpretação behaviorista da teoria bayesiana da decisão possui realmente valor teórico ou explicativo.

A origem desta dúvida prende-se, em primeiro lugar, com o tipo de linguagem que utilizamos e com os compromissos que parecemos assumir ao utilizá-la. Por exemplo, afirmar que um agente é indiferente entre duas apostas apenas se acreditar que a sua escolha não gerará outras consequências para além dessas apostas (ou que essas consequências não terão qualquer efeito quanto à desejabilidade das mesmas), parece implicar a atribuição ao agente de certos estados mentais que não são passíveis de observação através de um tipo específico de comportamento.

Em segundo lugar, a teoria da *preferência revelada*, proposta em 1938 pelo economista Paul Samuelson, e ainda hoje fortemente dominante no contexto da teoria económica, não aceita como possibilidade a existência de objectos ou apostas de valor incomensurável. Segundo esta teoria, dadas ao agente duas alternativas mutuamente exclusivas e exaustivas, se o agente escolhe uma delas, então um observador pode concluir que a alternativa escolhida é, para o agente, pelo menos tão valiosa quanto a não escolhida. Este

aspecto, que resulta da perspectiva comportamentalista de Samuelson, pode ser compreendido quando se considera a dificuldade em descortinar, ou mesmo conceber, um traço comportamental do agente que seja objectivamente observável e que corresponda à sua atitude perante alternativas incomensuráveis. Se, por exemplo, entre duas alternativas, nenhuma é escolhida - podemos facilmente conceber cenários do tipo '*throw the fat lady off the bridge*'[10] - dever-se-ia concluir que, afinal, não estamos perante uma escolha entre alternativas mutuamente exclusivas e exaustivas; ou seja, existirá uma outra alternativa que não se encontra contemplada ou que é possível escolher uma alternativa sem excluir a outra. No contexto da teoria da utilidade esperada, dir-se-á que todas as alternativas são comparáveis de acordo com a função de utilidade do agente.

Mas, por outro lado, recordando o princípio de Pascal, considere-se que um agente é indiferente entre uma determinada aposta *A* e o seu *preço justo* $*a*, conclusão a que se chegou observando o seu comportamento de escolha: num jogo de casino esta escolha é-lhe colocada sucessivamente e ele vai variando entre uma ou outra alternativa. Suponha-se, ainda, que aquilo em que o agente realmente acredita é que $*a* é ligeiramente melhor do que *A*; mas ele acredita também que escolhendo *A* poderá virá a ser recompensado no futuro imediato, caso a sorte lhe vá sorrindo. Ou seja, apenas com base no seu comportamento, não é possível distinguir se o agente é indiferente entre as alternativas *por elas mesmas* ou se o seu comportamento tem por base as crenças acima mencionadas.[11] Este tipo de explicação para o comportamento do agente, natural e plausível, não se encontra ao dispor da abordagem comportamentalista.

Não se pense, contudo, que a interpretação behaviorista é característica especial de um qualquer estádio de desenvolvimento da teoria da decisão. Esta interpretação aplica-se não só ao estádio rudimentar da teoria enquanto teoria dos *preços justos de apostas*, oferecida por Pascal, mas também às versões modernas da teoria, com os seus axiomas da preferência. Considerem-se, por exemplo as duas condições que caracterizam o *preço justo de apostas*, nas suas duas versões, mentalista e behaviorista:

a) O agente encara a perspectiva de receber uma qualquer soma superior a $*a* como *estritamente mais desejável* do que a perspectiva de aceitar *A*.

[10] 'Empurro a senhora gorda da ponte e salvo todos os restantes, incluindo eu próprio, ou não a empurro e morremos todos'.

[11] Esta possibilidade é mencionada por Joyce (1999: 21).

b) O agente encara a perspectiva de receber uma qualquer soma inferior a $*a* como *estritamente menos desejável* do que a perspectiva de aceitar *A.*

Estas duas condições, numa interpretação behaviorista, transformam-se nas seguintes:

c) Se for *apresentada ao agente a escolha* entre *A* e uma qualquer soma superior a $*a*, então *o agente escolherá* essa soma.
d) Se for *apresentada ao agente a escolha* entre *A* e uma qualquer quantia inferior a $*a*, então *o agente escolherá A.*[12]

Da mesma forma, os teoremas da representação poderiam ser apresentados sem se mencionar preferências enquanto estados mentais dos agentes, mas antes como disposições para se escolher entre eventos ou proposições alternativas, disposições essas que não violariam os axiomas da preferência; ou seja, que o agente é racional se, e somente se, estiver disposto a escolher *A* em vez de *A*' se, e somente se, a utilidade esperada de *A* for superior à utilidade esperada de *A*'. A ideia é a de que todos os termos essenciais à definição do princípio da maximização da utilidade subjectiva esperada, como a representação de *graus de crença* através de uma função de probabilidade ou a representação da *magnitude relativa dos desejos* através de uma função de utilidade, podem ser interpretados através da observação do comportamento observável do agente. Esta perspectiva é atraente para economistas, pesquisadores de mercado ou psicólogos que desejem tirar conclusões a partir desse comportamento observável. Mas quando pretendemos fazer inferências acerca das preferências dos agentes a partir do seu comportamento, sem admitir que essas inferências constituem induções, reaparecerem as dúvidas quanto ao poder explicativo da teoria. Estas inferências são essenciais para *explicar* ou *fazer racionalizações* acerca do comportamento, ou seja, responder a questões do tipo 'Por que razão fez ele isto ou aquilo?' A resposta que imediatamente se afigura consiste em fazer referência às crenças e desejos do agente. Além disso, é difícil justificar a irracionalidade de uma determinada escolha sem referir que o desejo de executar uma certa acção não estava de acordo com as crenças e outros desejos do agente. Contudo, de acordo com os positivistas, a relação entre os desejos e as disposições para agir é analítica,

[12] Estas quatro condições são uma tradução, sem alteração de substância, das apresentadas por Joyce (1999: 13-20.

ou seja, o agente apresenta os desejos descritos, por exemplo, em a) e b) se, e somente se, apresentar as disposições descritas em a') e b'). Mas se esta relação é analítica, se se tratar de uma mera necessidade conceptual, onde poderemos então encontrar uma explicação científica da acção, ou da irracionalidade desta, que não seja circular? De modo a responder a esta pergunta, talvez seja importante começar por analisar o conceito de explicação.

O conceito que iremos utilizar é aquele que é capturado pelo modelo hempeliano de explicação, designado como *nomológico-dedutivo* (Hempel 1965). Os termos são desde logo elucidativos. Classifica-se como dedutivo porque se considera que uma explicação científica tem a forma de um argumento, em que o evento a explicar, o *explanandum*, aparece no lugar da conclusão; e nomológico (de *nomos*) porque no lugar das premissas, o *explanans*, deve constar, juntamente com uma certa descrição das condições iniciais que se pressupõe serem condicionantes do fenómeno a explicar, a enunciação de uma lei de carácter geral. Como a hipótese explicativa que iremos considerar é a da chamada Psicologia Popular, em que a combinação prévia de certas atitudes proposicionais do agente, crenças e desejos, tem de se verificar com via a explicar a acção que este empreende, um argumento nomológico-dedutivo apresenta o seguinte aspecto:

1. O agente *a* tinha (um desejo *d* ∧ uma crença *c*)
2. Qualquer agente que tenha (*d* ∧ *c*) faz *f*

Logo, *a* fez *f*

Esta concepção de explicação tem a vantagem de permitir acomodar dois elementos que intuitivamente se considera que devem constar numa explicação científica. Primeiro, que uma explicação deve consistir numa resposta à pergunta 'Porque é que *f* é o caso?', ou seja, que compreender um fenómeno é conhecer as suas causas. Neste caso, que a combinação de atitudes proposicionais atribuída ao agente se verificou e que cada uma delas, juntamente com a conclusão *f*, recai sob um *tipo específico* de evento (cujas instâncias, por definição, são repetíveis), encontrando-se, por isso, implícito o apelo a uma lei geral, de acordo com a qual sempre que uma instância do tipo da causa se verifica, verificar-se-á também uma instância do tipo do efeito. O segundo elemento que se deixa capturar por este modelo, e que se reveste de importância em qualquer concepção científica do mundo, é a noção de previsibilidade; ou seja, que explicar e compreender os

factos implica a capacidade de prever a sua ocorrência, sempre que os antecedentes adequados se manifestem, de modo a poder-se manipular os fenómenos em causa, caso isso seja possível. Esta relação lógica entre explicação e previsão encontra-se implícita na ideia de que, sendo o *explanandum* uma consequência lógica do *explanans*, a verificação da conjunção das premissas constituiria uma condição suficiente para a ocorrência do primeiro.[13]

Finalmente, duas condições devem verificar-se para que um argumento tenha realmente um carácter explicativo e não apenas a aparência de uma explicação. A primeira ficou já clara, consistindo na necessidade de incluir nas premissas uma lei de carácter geral que dê conta das relações causais que o argumento tenta capturar. Segundo, que a conclusão não se siga de um subconjunto das premissas, deixando de fora qualquer uma delas; ou seja, é necessário que a verdade da conclusão seja estabelecida pelo conjunto total das premissas contidas no *explanans*, caso contrário a presença neste de uma ou outra premissa seria gratuita ou dispensável. Tal pode acontecer, por exemplo, no caso em que na descrição das condições relevantes, que podem consistir na atribuição de propriedades ao agente, esteja já implícita a lei geral em que se subsume a relação causal presente no mundo.

Considere-se o seguinte exemplo: quando queremos explicar em que consiste a acção racional, é natural que a lei a incluir no *explanans* exprima aquela relação entre as condições iniciais (as crenças e desejos do agente) e a acção executada que permita classificar esta última como racional. Assim, é natural que na atribuição ao agente, no conjunto das premissas, da propriedade de ser racional, esteja já incluída a lei que, em conjunto com as condições iniciais, permite derivar a conclusão. Ou seja, se essa premissa adicional consistir numa proposição que contém já a lei segundo a qual ser racional é fazer *f* quando se tem o desejo *d* e a crença *c*, então é possível derivar o *explanandum* apenas da conjunção entre a premissa que inclui as condições iniciais e a premissa que atribui ao agente a propriedade de ser racional, tornando-se desnecessária a premissa que contém a lei geral. Será, portanto, no quadro desta concepção nomológico-dedutiva de explicação que avaliaremos de seguida algumas das hipóteses de construção de um argumento com real valor explicativo.

---

[13] Se quisermos, podemos clarificar e considerar que, de acordo com uma concepção conhecida de causalidade (Mill 1919), cada uma das crenças e desejos do agente, e cada uma das restantes condições relevantes – como o agente não se encontrar atado a uma cadeira ou que o mundo entretanto não vai acabar, etc. - constituem condições necessárias da acção, a conjunção das quais constitui uma condição suficiente do fenómeno que se pretende explicar.

No contexto de uma abordagem positivista, o problema que uma teoria do comportamento humano, que se pretende empiricamente testável, enfrenta é o seguinte: como avaliar a verdade das premissas incluídas no *explanans*, sendo que estas consistem na atribuição ao agente de certas atitudes proposicionais? O modo de resolver este problema depende, em parte, do modo como uma atitude proposicional é caracterizada. De um ponto de vista relacionalista, as atitudes proposicionais são caracterizadas como uma relação que se verifica entre um sujeito e um qualquer objecto, podendo este último ser caracterizado de maneiras diferentes. Ora, as atitudes proposicionais possuem um conteúdo semântico, ou seja, são portadoras de sentido, logo as duas hipóteses de caracterizar o segundo elemento da relação são as seguintes: ou consistem em proposições ou em frases. Se supusermos que se trata do primeiro caso, temos de conceber a possibilidade de entidades físicas, que existem no espaço e no tempo, se encontrarem numa relação com entidades abstractas que não existem no espaço-tempo. Invariavelmente, as atitudes proposicionais assim concebidas são classificadas como estados mentais. Dado que os estados mentais não podem ser objecto de observação empírica, não podem também, por conseguinte, constituir tipos de coisas que possamos definir como causas da acção, as quais funcionam de acordo com leis empiricamente verificáveis. Restaria, assim, a segunda hipótese, proposta por Carnap (1953), de que as atitudes proposicionais consistem numa relação entre um sujeito e uma frase. Embora se possa atribuir um valor de verdade às frases da linguagem natural (que se supõe serem aquelas que estão em causa quando se trata de desejos e crenças), existe o problema de saber como se pode determinar o valor de verdade da atribuição de atitudes proposicionais, pelo menos enquanto forem entendidas como estados mentais.

A alternativa que se irá analisar tenta resolver este problema eliminando do discurso científico, nomeadamente daquele que é empregue na explicação de uma acção, toda a referência a estados mentais. Por exemplo, se a crença de um agente consistir na relação entre este e uma frase da língua natural, faz sentido conceber que essa atribuição de crença seja verdadeira quando o agente está disposto a usar uma frase que representa o conteúdo da sua crença de forma adequada. Por exemplo, se alguém acreditar que ‘o Sporting é o melhor clube de Portugal’, estará então disposto a dar o seu assentimento a uma frase como ‘O Sporting é melhor do que o Benfica’; este comportamento linguístico é algo de observável e, eventualmente, cientificamente controlável, ao contrário da respectiva crença, entendida como estado mental.

O projecto que se propõe eliminar os conceitos mentais, como crenças e desejos, do discurso psicológico científico, e substituí-los por disposições comportamentais, é o behaviorismo. As motivações para abraçar este projecto podem, todavia, ser de ordem distinta. Para um behaviorista como Skinner (1953), por exemplo, a atitude é mais metodológica do que propriamente ontológica. A sua preocupação prende-se mais com o rigor do discurso científico - com a sua capacidade de controlar variáveis externas que operam sobre a acção humana, de efectuar previsões acertadas, de evitar explicações *ad hoc* ou redundantes - do que propriamente com um cepticismo acerca de uma ontologia de estados mentais. O seu argumento é bastante persuasivo. Mesmo que possamos efectuar inferências do comportamento observável para estados mentais e pressupor a eficácia causal destes, teremos sempre de recuar na ordem das causas para fora da mente do agente, para variáveis externas que influem no seu comportamento. Ora, estas variáveis não podem ser eliminadas de uma teoria psicológica da acção sem que o seu valor explicativo fique extremamente reduzido.

Gilbert Ryle (1949), por seu lado, motivado pelas suas dúvidas quanto à natureza do mental, introduziu o conceito de propriedade disposicional. A sua motivação não parece apoiar-se tanto num pressuposto verificacionista, mas sim nos argumentos que o próprio apresenta para caracterizar o suposto erro categorial em que, segundo ele, os dualistas de substâncias teimam em cair. O exemplo que apresenta poderia ser utilizado por um funcionalista para descrever aquilo em que consistiria a natureza do mental. Quando um palestrante convidado pede para que lhe seja mostrada a Universidade, o cicerone mostra-lhe o campus universitário, os edifícios das faculdades, a reitoria, explica-lhe o modo de funcionamento da universidade e as relações que existem entre as suas partes, etc., e, no final, o indivíduo a quem tudo isto foi mostrado, questiona ainda: 'Sim, muito bem, mas onde está a Universidade?' Ora, o erro aqui evidenciado seria o mesmo que os dualistas cometeriam ao conceberem a mente e os estados mentais como coisas essencialmente diferentes destas partes, das suas funções e das relações que mantêm umas com as outras. Nas palavras de Ryle, 'o fantasma dentro da máquina'. Independentemente do mérito dos seus argumentos a favor de uma visão materialista da mente e contra o apelo a estados mentais como causas explicativas da acção, a sua concepção de crenças e desejos é fundamental para todo o projecto behaviorista/disposicionalista de explicação da acção. Para Ryle, tal como quando falamos da mente de um indivíduo estamos realmente a falar do seu comportamento e do modo como as suas actividades são coordenadas, também

quando falamos nos seus desejos e crenças estamos a referir-nos a disposições para agir de uma determinada maneira; ou seja, atribuir a um indivíduo uma crença ou um desejo é descrever esse indivíduo como alguém que se comportaria de determinada maneira, caso certas condições se verificassem. A verificação, digamos, da atribuição a alguém de uma dada propriedade disposicional não é dada por uma frase categórica que descreveria a ocorrência de um qualquer estado do mundo; por exemplo, uma frase do tipo 'É verdade que Maria possui a propriedade de *desejar ver o Sporting sagrar-se campeão* se, e somente se, existir um estado mental que consiste em Maria desejar ver o Sporting sagrar-se campeão'. No caso de uma propriedade disposicional, a verificação dependeria da verdade da conjunção de várias condicionais como as seguintes: 'Maria deseja ver o Sporting campeão se, e somente, Maria festejar a vitória do Sporting se o Sporting for campeão, apoiar o Sporting nos jogos em se irá decidir se o Sporting será, ou não, campeão, etc.'. Esta propriedade poderia ser classificada como uma propriedade disposicional de ordem superior, pois é definida através de outras propriedades disposicionais de ordem inferior, como 'festejar' ou 'apoiar'; e poderia também classificar-se como ampla, pois admite uma larga variedade de padrões comportamentais na sua definição. Fica por determinar se algum desses padrões comportamentais pode constituir uma condição suficiente para determinar a adequação da atribuição da propriedade ou se existe uma conjunção de padrões, cada um deles uma condição necessária, que possa assumir esse papel.

Como vimos acima, a inclusão da propriedade 'ser racional' numa das premissas do silogismo prático, definida de um modo disposicionalista como a acção que alguém empreenderá caso tenha um desejo e uma crença como as que eram referidas na premissa contendo as condições iniciais, tornava o argumento explicativamente vazio pelo facto de se poder prescindir da lei geral que conecta causalmente as condições iniciais com a conclusão. Além disso, na medida em que é possível interpretar a relação entre as premissas e a conclusão como contingente, faz sentido pensar que, sendo concebível existirem casos em que as premissas são verdadeiras e a conclusão é falsa, algo fica igualmente explicado, nomeadamente, a irracionalidade do agente. No contexto da filosofia de Aristóteles, em que se aceitava como pressuposto a definição de homem como animal racional, fazia sentido incluir a atribuição dessa propriedade nas premissas do argumento. Mas, tanto no caso da teoria de Aristóteles, como no contexto das abordagens modernas ao problema da explicação da acção, é sempre bom possuirmos uma razão que

nos permita explicar a possibilidade de um agente agir irracionalmente, seja devido à fraqueza da vontade ou a um erro cognitivo na aplicação da lei geral ao caso particular, como no caso de Aristóteles, ou aceitando-se simplesmente a possibilidade de se agir contra a própria crença acerca da melhor maneira de agir numa determinada situação. A inclusão da propriedade 'ser racional' nas premissas do silogismo prático, emprestaria, assim, um sabor de contradição à análise da irracionalidade oferecida pela própria teoria explicativa em causa.

Considere-se, então, de forma mais detalhada, um argumento com a forma do silogismo prático:

1. (*a* desejava uma cerveja) ∧ (*a* cria que deslocar-se até junto do frigorífico e abri-lo era a coisa a fazer para obter uma cerveja)
2. (∀x) {[(*x* deseja uma cerveja) ∧ (*x* crê que deslocar-se até junto do frigorífico e abri-lo é a coisa a fazer para obter uma cerveja)] → (*x* desloca-se até junto do frigorífico e abre-o).

Logo, *a* deslocou-se até junto do frigorífico e abriu-o

Parece haver aqui uma interdependência entre as definições disposicionais de desejo e de crença. Qual será, por exemplo, a condição experimental mais relevante para definir a propriedade disposicional que consiste em *Maria desejar uma cerveja*? A resposta mais óbvia parece consistir no apelo à disposição para ter uma crença: 'se Maria crê que a coisa a fazer para obter uma cerveja é deslocar-se até ao frigorífico e abri-lo, então Maria deseja uma cerveja se, e somente se, se deslocar até ao frigorífico e o abrir'. Do mesmo modo, para a crença relevante, a condição experimental que parece revelar o comportamento mais apropriado parece consistir na disposição para desejar algo: 'se Maria deseja uma cerveja, então Maria crê que a coisa a fazer para obter uma cerveja é deslocar-se até ao junto do frigorífico e abri-lo se, e somente se, se deslocar até junto do frigorífico e o abrir'. Ambas estas frases têm a forma de frases de redução bilateral, as quais Carnap e Hempel consideraram ser adequadas para introduzir no discurso científico uma propriedade disposicional: A → (B ↔ C), em que A consiste na condição experimental sob a qual se deverá verificar o comportamento adequado; B na atribuição da propriedade em causa; e C na verificação empírica do comportamento que se pretende explicar. Ora, nas duas frases acima consideradas, parece difícil negar que no lugar de C temos um

comportamento que constitui uma condição suficiente para atribuir a alguém tanto o desejo como a crença considerada, nomeadamente, a acção que consiste em deslocar-se até ao frigorífico e abri-lo.

O problema é o seguinte: considere-se a frase da forma acima considerada, por meio da qual é possível atribuir a Maria o desejo de beber uma cerveja: A = Maria crê que ir até junto do frigorífico e abri-lo é a forma de obter uma cerveja, B = Maria desloca-se até junto do frigorífico e abre-o, e C = Maria deseja uma cerveja. Ora, se B pode ser concebida como uma condição necessária para atribuição da propriedade em C, teremos então também uma frase com a seguinte forma: $A \rightarrow (B \rightarrow C)$. Mas, pela regra da importação, podemos derivar uma outra frase com a forma $(A \wedge B) \rightarrow C$: 'se Maria crê que deslocar-se até junto do frigorífico e abri-lo é a coisa a fazer para obter uma cerveja e Maria deseja uma cerveja, então Maria desloca-se até junto do frigorífico e abre-o'.

Fica, assim, demonstrado que é possível obter-se a conclusão do silogismo prático a partir das condições iniciais que atribuem a Maria uma certa crença e um certo desejo, sem necessidade de inclusão da premissa que contém a lei geral. Logo, de acordo com os critérios apresentados, que determinam se um argumento tem, ou não, valor explicativo, podemos concluir que o modelo do silogismo prático, enquanto argumento para a explicação da acção, que inclui atribuições de crença e desejo ao agente definidas em termos behavioristas/disposicionalistas, não tem valor explicativo.[14] [15]

2.2. Construindo uma função de utilidade

Em alternativa ao modelo do silogismo prático, que se adequa perfeitamente à concepção hempeliana de explicação da acção enquanto tipo de argumento, existe uma teoria da

[14] Com efeito, apesar de o apelo a condições necessárias e suficientes para explicar a produção da acção ser bastante forte, podem conceber-se casos em que essas condições parecem realmente existir, enquanto desejos e crenças com eficácia causal, e, ainda assim, não serem suficientes para explicar a acção. Considere-se este exemplo de Davidson: 1) Édipo tinha o desejo de matar o seu pai, 2) Édipo tinha a crença, no momento em que teve a oportunidade, de qual era o meio para o conseguir, 3) Édipo matou o seu pai; todavia, a acção que consistiu em matar Laio não pode ser explicada pelas condições iniciais invocadas. Isto porque a razão pela qual Édipo matou Laio não consistiu no desejo de matar o seu pai. Daí o cepticismo de Davidson quanto à possibilidade de oferecer condições necessárias e suficientes para explicar a intencionalidade de uma acção em termos de conceitos como crenças, desejos e causas, i.e., o cepticismo quanto à possibilidade de oferecer explicações para a acção através de leis causais (Davidson 1974).

[15] O argumento exposto nesta secção, aplicado ao silogismo prático, consistiu numa versão abreviada do argumento original que se encontra no primeiro capítulo da parte um de *Animal Racional ou Bípede Implume*? (Zilhão 2010b).

decisão racional que pode preencher o papel de uma teoria explicativa da acção. Trata-se da moderna teoria bayesiana da decisão. A vantagem desta, relativamente à anterior, consiste em definir o conceito de racionalidade de uma forma rigorosamente quantificada. Portanto, se assim o desejarmos, podemos ter uma interpretação behaviorista do conceito de maximização da utilidade subjectiva esperada. Esta noção pode ser definida através da atribuição ao agente de uma série de comportamentos – preferências por apostas - sem necessidade de incluir na análise qualquer pressuposto mentalista, ou seja, qualquer atribuição de eficácia causal a estados mentais do agente.

Na sua origem, o conceito de utilidade esperada especificava o grau de satisfação (felicidade) que corresponde à obtenção da consequência de uma qualquer das acções que o agente pode empreender. Por exemplo, se é dado ao agente escolher entre A) um bitoque, B) uma lasanha de legumes e C) uma salada de frango, e o agente prefere B a C e C a A, então isso significa, *simpliciter*, que a acção que consiste em pedir uma lasanha de legumes é aquela que maximiza a utilidade do agente. A ordenação ou o estabelecimento do valor relativo das três alternativas equivale à construção da função de utilidade do agente. Suponha-se, contudo, que o agente se encontra num tipo de restaurante cuja salubridade o faz duvidar da capacidade que cozinheiro terá para servir uma boa lasanha de legumes ou até mesmo uma salada de frango livre de salmonelas. Nesse caso, pode acontecer que a acção que consiste em pedir um bitoque seja realmente aquela que, na ausência de alternativa, pode maximizar a utilidade do agente. Como sabemos, a racionalidade da acção passa assim a depender não apenas da função de utilidade, mas também dos estados do mundo relevantes cuja verificação, ou não, contribui para determinar as consequências da acção. Assim, ao efectuar a sua escolha, o agente deve ter em conta a sua crença acerca das capacidades do cozinheiro. Mas, tal como no caso do desejo a teoria bayesiana introduz o conceito mais preciso de *utilidade subjectiva esperada*, também para o caso das crenças (e, assim, temos as condições iniciais do argumento) a teoria bayesiana introduz o conceito de *distribuição de probabilidades do agente*; ou seja, o agente atribui subjectivamente uma determinada probabilidade aos estados do mundo relevantes que determinam as consequências de cada uma das acções disponíveis. Por exemplo, ao atribuir uma probabilidade de 70% à capacidade do cozinheiro de confeccionar uma boa lasanha de legumes, o agente está a exprimir de uma forma quantitativamente adequada a sua crença de acordo com a qual o cozinheiro é bastante competente.

Como vimos, ao contrário do que, por vezes, alguém menos familiarizado com esta interpretação subjectiva de probabilidade pode pensar, a distribuição de probabilidades do agente não é, por definição, arbitrária. É este um dos erros em que caem alguns daqueles que alegam que a teoria é implausível. Contudo, não é difícil verificar, através do comportamento observável dos agentes, que muitas das suas acções são de facto conformes aos axiomas do cálculo de probabilidades. Um jogador racional não irá fazer uma dupla aposta sobre a ocorrência de um evento em que as perdas decorrentes da não ocorrência suplantam sempre os ganhos obtidos pela ocorrência. Do mesmo modo, um veraneante racional, que atribui 98% de probabilidade ao estado do mundo que consiste em 'estar sol e calor amanhã' e 2% à probabilidade de 'estar frio e chover amanhã', não irá certamente para a praia vestido de gabardina.

A vantagem que a teoria da utilidade esperada tem de explicitar de uma forma mais precisa o conceito de racionalidade de uma acção, como complemento à doutrina do silogismo prático, pode ser melhor compreendida ao considerarmos que esta consegue incorporar na sua explicação a verificação efectiva da adequação da crença que o agente tem acerca de qual é o a melhor maneira de satisfazer as suas preferências. Ao utilizarmos os conceitos de utilidade subjectiva de uma acção, quantificando o grau de satisfação que uma determinada consequência traz ao agente, e de probabilidade subjectiva da ocorrência de um estado do mundo, conseguimos, através de um cálculo matemático simples, determinar qual a acção que melhor satisfará as suas preferências.

Uma coisa é calcular e decidir quando se encontram dados à partida os valores exactos necessários ao cálculo, outra é encontrar uma maneira de os agentes determinarem de uma maneira rigorosamente quantificada a sua função de utilidade. O método para esse efeito foi inventado por John von Neumann e Oskar Morgenstern (1953) e consiste no seguinte: suponhamos que a ementa atrás considerada esgota todas as possibilidades à disposição do agente. Este pode, então, começar por seleccionar os dois extremos da sua função de utilidade, conferindo uma utilidade de valor 1 à opção lasanha de legumes e uma utilidade de valor 0 à opção bitoque. A questão consiste em saber qual o valor exacto da utilidade subjectiva da opção salada de frango. Para esse efeito, imaginemos que o agente tem um amigo que lhe oferecerá garantidamente a salada de frango; contudo, a obtenção da melhor e da pior das três possibilidades, a lasanha e o bitoque, não depende da sua escolha imediata, mas sim de um sistema de lotaria em que o primeiro prémio é a lasanha e último o bitoque. Podemos supor que ao serem oferecidos bilhetes para esta

lotaria, em que a probabilidade de obtenção do melhor prémio vai progressivamente aumentando, chegar-se-á a um ponto em que o agente prefere correr o risco de perder a lotaria, e comer o bitoque, a comer de certeza a salada de frango. Se o bilhete de lotaria que ele aceitar lhe der, por exemplo, uma probabilidade de 0.5 de comer a lasanha e 0.5 de comer o bitoque, tal significará que a salada de frango tem uma utilidade subjectiva cujo valor é de 0.5. Este procedimento poderia ser repetido para qualquer outro prato numa ementa com mais opções do que a apresentada.

Qual é, então, a intuição de implausibilidade aqui presente? É aquela que consiste em pensar que os agentes racionais comuns, e não um subgrupo dotado de informação especializada e particularmente dotado, ou que tenha de escolher em condições bastante excepcionais, agem sempre em função do conhecimento das probabilidades objectivas das condições por intermédio das quais são definidos os valores da sua função de utilidade. Esta intuição não está relacionada com a suposição de que um dos papéis da teoria consiste em exigir do agente algo mais do que a simples satisfação dos axiomas da preferência. Nem que, se estivermos a considerar a questão descritiva, os agentes procedem de acordo com esse cálculo. No mínimo encontra-se implícita a ideia de que, pelo menos de uma forma subconsciente, existe da parte do agente uma ponderação das probabilidades dos estados relevantes, aquando da definição das suas preferências.

Ora, o que está aqui em causa é o conceito de probabilidade que é utilizado no método de von Neumann, o conceito empírico, objectivo, de probabilidade como frequência relativa: a probabilidade de um evento equivale ao limite para o qual tende o valor da frequência relativa com que esse evento se verificaria numa sucessão de eventos de tipo similar, quando essa sucessão se estende até ao infinito. A contenção de implausibilidade pode, então, ser definida do seguinte modo: o agente racional estaria para o agente comum, assim como o meteorologista estaria para o indivíduo que põe cuspo num dedo e aponta para o céu. A teoria seria demasiado exigente quanto aos seus pressupostos psicológicos, sendo tanto normativa como descritivamente irrealista. Daí a contenção de Hempel, segundo a qual a teoria da decisão possuiria apenas um valor ´crítico-normativo': algo que um técnico competente poderia talvez utilizar para analisar um problema de decisão e recomendar a melhor opção a tomar, mas não algo que pudesse ser realmente ser tomado como uma explicação psicológica da acção (ver Zilhão 2007, 2010b).

Ramsey (1931) criou um método para determinar a função de utilidade do agente e, ao mesmo tempo – no que difere do de von Neumann – determinar a sua distribuição de

probabilidades. No método de Ramsey, o conceito de probabilidade está a ser interpretado de maneira subjectiva, não se pressupondo a necessidade de entendimento, por parte do agente, de qualquer concepção técnica do conceito Considerem-se duas opções possíveis, às quais correspondem igualmente os dois valores extremos da função de utilidade do agente: a opção B, que consiste em ir jantar com Maria, tem valor de utilidade 1; e a opção A, que consiste em ir jantar com Joana, tem valor de utilidade 0 (ficar em casa não é opção). Tentemos, depois, encontrar uma condição P que torne o agente indiferente entre as duas seguintes apostas: a Aposta 1) se P se verificar, então B; se P não se verificar, então A; e Aposta 2) se P se verificar, então A; se P não se verificar, então B. Por outras palavras, o agente prefere ir jantar com Maria *simpliciter*. Contudo, ele não tem a certeza de que Maria vá aceitar o seu convite, ou seja, não se encontra disposto a receber uma recusa. Mas, se Maria aceitar, ele não tem qualquer dúvida quanto a preferir ir jantar com ela. Entretanto, um amigo informa-o de que Maria talvez aceite ir jantar com ele. Contudo, este amigo não é inteiramente de fiar e o agente não sabe se há de acreditar nesta informação. Assim, ele torna-se indiferente quanto à possibilidade de pedir a Maria para jantar com ele ou simplesmente ir jantar com Joana. Ora, tal só pode significar que ele atribui o valor de 0.5 à probabilidade de P, ou seja, à probabilidade de Maria aceitar o seu convite.

Suponhamos, agora, que existe um outro empreendimento à disposição do agente nessa noite, digamos C, tal que o agente seja indiferente entre, por exemplo, a acção 1) Se P se verificar, então B; Se P não se verificar, então A; e a Aposta 3) C, verifique-se P ou não (como a Aposta 1 tem uma utilidade idêntica à da Aposta 2, poder-se-ia ter usado esta em vez da primeira). Suponhamos, por exemplo, que C consiste em ir jantar com os seus três melhores amigos. Ora, se o agente é indiferente entre a apostas 1), 2) e 3), e se vimos que a utilidade de 1) e 2) só podem ser idênticas, pois é isso que significa ser indiferente entre elas, então a utilidade de 3) será igual às utilidades de 1) e 2). Mas como sabemos qual era a utilidade de 1) e 2)? A utilidade de 1), por exemplo, seria calculada somando-se a utilidade de A, multiplicada pela probabilidade de P, com a utilidade de B multiplicada pela probabilidade de não-P; ou seja, 1. ½ + 0. ½ = ½. Conclui-se, assim, que a utilidade de C é de ½, valor que se encontra exactamente entre a utilidade de A e a utilidade de B. De seguida, se quiséssemos refinar a nossa função de utilidade, poderíamos tentar encontrar uma acção cuja utilidade fosse ¾. Para isso, seria necessário descobrir uma situação em que o agente fosse indiferente entre a Aposta 3) se P se verificar, então B; se

P não se verificar, então C; e a Aposta 4) D, seja P ou não o caso. D poderia muito bem ser, por exemplo, ir jantar com os três melhores amigos e um deles levar uma amiga que o agente gostaria de conhecer.

Com isto, ficamos na posse da função de utilidade do agente. Se pretendêssemos calibrá-la ainda melhor, poderíamos fazê-lo através da descoberta de uma condição à qual atribuiríamos subjectivamente a probabilidade de 0.5[16]:

1) jantar com Maria

¾) jantar com os três melhores amigos e uma amiga nova

½) jantar com os três melhores amigos

¼) ?

0) jantar com Joana[17]

Ao contrário do método de von Neumman, que faz uso do sistema de apostas em jogos de azar, baseado numa interpretação de probabilidade como frequência relativa, este método parece encontrar-se mais perto do tipo de considerações que normalmente nos atribuímos e nos revemos a fazer quando optamos por realizar certas acções em detrimento de outras no nosso quotidiano. Assim, quanto ao valor descritivo da teoria, se acharmos que este se encontra relacionado com um certo processamento por parte do agente das variáveis em jogo, uma teoria que utiliza os conceitos de utilidade subjectiva e de probabilidade subjectiva não parece impor pressupostos psicológicos demasiado exigentes. Esses mesmos factores operacionais, como vimos, podem ser todos eles definidos de uma forma comportamentalista. Se no esquema do silogismo prático tínhamos propriedades disposicionais, como desejos e crenças, que não se deixavam

[16] Se a condição P (Joana aceita o convite) fosse superior a 0.5, pode facilmente constatar-se que o agente teria escolhido a aposta 1, cuja utilidade seria a seguinte: 0,6.1+0,4.0=0.6. Já a aposta 2 teria a seguinte utilidade: 0,6.0+0,4.1=0.4. Não sei se é realista pensar na condição P como algo deste tipo, ainda mais porque a acção de ir jantar com a Joana teria de ser redefinida como, talvez, 'convidar a Joana para ir jantar'. Contudo, face à dificuldade em encontrar uma condição com probabilidade 0.5 que seja, de facto, psicologicamente realista, este tipo de raciocínio parece-nos ser útil, trazendo a teoria para um nível que todos podemos intuitivamente reconhecer como fazendo parte da nossa experiência comum. De facto, por que motivo tantas vezes na nossa vida nos abstivemos de convidar alguém para jantar, senão porque de facto não acreditávamos que o nosso convite seria aceite? Por outro lado, parece ser o caso que a maioria de nós necessita de um grau de confiança muitíssimo superior a 0.5 para arriscar uma recusa, tendo em conta que a utilidade do trauma subsequente teria um valor potencialmente infinito.

[17] Após construída a função de utilidade do agente, as suas probabilidades subjectivas, diferentes de ½, podem ser determinadas através da construção de fracções cujos numeradores são diferenças entre utilidades esperadas de apostas e utilidades de consequências, e cujos denominadores são diferenças entre utilidades de consequências. Os valores necessários a esta construção serão já conhecidos.

definir de um modo preciso e de forma independente umas das outras, agora temos o conceito de utilidade subjectiva que pode ser definido em termos de comportamentos de escolha. Se quisermos, a decisão de executar uma acção cuja utilidade subjectiva é a mais elevada pode ser definida de uma forma bastante semelhante à das propriedades disposicionais, como o comportamento que seria adoptado se determinadas condições empíricas, observáveis, se verificassem. Do mesmo modo, a atribuição subjectiva de probabilidades aos estados do mundo pode ser definida através da disponibilidade do agente para agir de determinada maneira, por exemplo, sair de casa com um chapéu-de-chuva, após ter escutado no rádio o boletim meteorológico.

A questão que finalmente se coloca é a seguinte: independentemente de se verificar experimentalmente, ou não, que os seres humanos agem realmente de modo a maximizar a sua utilidade subjectiva esperada, importa saber se uma teoria com os meios de análise da teoria bayesiana pode ter, ou não, um valor teórico ou explicativo real. Esta questão é pertinente não só quando se considera que a teoria descreve ou exige realmente algum tipo de cálculo – abordagem que me parece errada (§3.1) - ou quando a teoria apenas descreve e exige a satisfação, por parte do agente, de certos axiomas da preferência.

Uma das maneiras de elucidar o problema que se encontra na base desta questão pode passar por uma análise da evolução do conceito de utilidade. No caso dos primeiros utilitaristas, a utilidade era equiparada directamente àquilo que hoje se poderiam considerar estados mentais, dores e prazeres, ou atitudes proposicionais como ‘ser ou estar feliz’, e, nesse sentido, era perfeitamente concebível oferecer explicações do tipo das da Psicologia Popular. A Maria escolheu comer a laranja, em detrimento da banana e da maçã, porque comer a laranja lhe iria proporcionar maior prazer. O conceito de utilidade passou também a estar associado ao do valor moral de uma determinada consequência, o qual não teria de ser equivalente, por exemplo, ao valor monetário de uma aposta. Esta perspectiva permitia-nos compreender as razões pelas quais indivíduos diferentes tomariam decisões diferentes em circunstâncias semelhantes. Por exemplo, a razão pela qual um jogador que ganha o prémio mais baixo num concurso de raspadinhas opta por reclamá-lo, ao contrário de outro que opta por trocá-lo pelo seu valor em raspadinhas novas. Contudo, à medida que o conceito de utilidade vai assumindo um papel mais técnico em disciplinas como a economia, tornando-se necessário encontrar procedimentos para medir o valor da utilidade de maneira rigorosa, o conceito perde o seu substrato psicológico inicial.

O que rigorosamente se segue da análise comportamentalista é o seguinte: podemos afirmar que Maria prefere B a A porque Maria é observada a escolher B em vez de A. Com efeito, se o conceito de preferência é definido dessa forma, através das escolhas que Maria é observada a fazer, tudo o que a teoria mostra é que Maria pode ser consistente no seu comportamento e que este pode continuar a ser descrito como sendo o comportamento de alguém que age *como se* estivesse a maximizar a utilidade subjectiva esperada. Nesta medida, a conexão entre as preferências de Maria e aquela que é realmente a acção que melhor as satisfaz continua a ser mantida, mesmo de acordo com uma interpretação behaviorista da teoria. Contudo, o abandono de qualquer pressuposto psicológico na definição dos conceitos-chave parece conduzir ao abandono de uma explicação causal para o comportamento, em que a posse efectiva de uma ordem de preferências, ou de uma função de utilidade e uma distribuição de probabilidades, seria a causa de ela agir como age.

Não se verifica aqui exactamente uma petição de princípio, pois o que se pretende explicar através do modelo de acção racional - uma instância comportamental x, um evento x ou uma determinada escolha - não está incluído nos elementos ou, se quisermos, nos valores que constituem *inputs* para a computação. Ou seja, a acção levada a cabo, ou a escolha final, é conceptualmente diferente dos procedimentos utilizados para construir as funções em causa, as quais deveriam explicar por que motivo a acção foi efectuada. Acontece, todavia, que os eventos a explicar, certos comportamentos, são eventos do mesmo tipo dos que são utilizados para efectuar a explicação, mais precisamente, outros comportamentos. Tudo isto conduz à suspeita de que esses elementos poderão ter o mesmo tipo de causas, e o conhecimento das mesmas deveria, assim, constituir o substrato da explicação propriamente dita (ver Zilhão, 2010b: 112-115).

Uma maneira de expressar esta preocupação consiste em fazer notar, como no início, que a teoria parece ser neutra quanto à suposta relação de causa-efeito que normalmente se pretenderia inferir, das acções para os estados mentais, como explicação de uma qualquer acção x. Tudo o que uma perspectiva comportamentalista nos obriga a concluir é que parece existir uma relação de conjunção mais ou menos constante entre certas acções e o comportamento cuja observação é necessária para se construir a função de utilidade e a distribuição de probabilidades do agente. A capacidade preditiva daí resultante não é, contudo, sinónimo de verdadeiro poder explicativo. Explicação implica previsão, mas o inverso não se verifica. Uma determinada leitura do barómetro ocorre em conjunção

constante com um determinado estado do tempo, mas daí não se segue que a primeira seja a causa do segundo. Mais uma vez, suspeita-se que ambos sejam efeitos coocorrentes da mesma causa. Do mesmo modo, não é possível concluir que a posse de uma função de utilidade e de uma distribuição de probabilidades seja a causa de uma determinada acção. O que esta abordagem sugere é que tanto a acção final como os comportamentos de escolha observados têm ambos o mesmo tipo de causa.

Conclui-se, assim, que as teorias que assumem versões comportamentalistas para explicar a acção humana intencional – a teoria do silogismo prático e a teoria bayesiana da decisão - não possuem realmente um valor explicativo elevado, a menos que pretendam de algum modo determinar como podem factores internos ter uma eficácia causal na produção tanto daquelas acções tidas usualmente como racionais, como daquelas classificadas como irracionais.

3. Lógica das preferências

3.1. Completude

Um jogador tem de optar entre receber 50 Euros certos ou jogar um jogo que consiste em atirar ao ar uma moeda não-viciada duas vezes seguidas. Se sair duas vezes cara, ganha 200 Euros, se sair duas vezes coroa ganha 100 Euros, e se sair só uma vez cara (ou coroa) não ganha nada. A utilidade esperada do jogo (que supomos ser idêntica ao seu *valor monetário esperado*) calcula-se da seguinte maneira:

$p(0.25).u(200) + p(0.25).u(100) + p\ (0.50).u(0) = 75.$

Ele opta por jogar. Mais, ele acredita que, se jogar repetidamente este jogo um número *n* de vezes, a probabilidade com que ganhará cada um dos três valores vai variando – relativamente a 0.25, 0.25 e 0.50 - por uma certa margem ε até que ε convergirá para 0 à medida que *n* se vai aproximando de infinito. Esta convicção probabilística encontra-se na base da *lei dos grandes números*, também designada como *lei empírica do* acaso, e que se supõe ser verdadeira para todo o $\varepsilon > 0$, por mais pequeno que ε seja. Um argumento a favor da maximização da utilidade esperada que faça apelo à *lei dos grandes números*

será um argumento que estabelece o seguinte: a *longo prazo* ficaremos *quase de certeza* melhor maximizando a nossa utilidade subjectiva esperada do que ficaríamos se não o fizéssemos; ou seja, o jogador saberá que à medida que *n* se aproxima de infinito, a probabilidade com que a utilidade média e a utilidade esperada do jogo diferem irá convergir para 0.

Um número elevado de condições teria de ser satisfeito de modo a que cada jogo contasse como o mesmo evento e cada decisão de jogar como a mesma decisão: que a moeda fosse sempre a mesma, que fosse lançada sempre da mesma maneira, etc. Mesmo num casino, as condições em que se joga podem variar e é duvidoso que um jogador seja confrontado com o mesmo problema de decisão várias vezes seguidas. Contudo, para poder ser aplicada, a *lei dos grandes números* não necessita da suposição de que estamos sempre perante o mesmo problema de decisão, bastando supor que a probabilidade de cada consequência é a mesma ao longo do tempo. Isto vai, no entanto, contra o que resulta de certas concepções de probabilidade. Por exemplo, da concepção subjectiva não se segue necessariamente que a probabilidade de obter um determinado resultado num jogo de roleta não varia ao longo do tempo, embora seja claramente pouco sensato, no mínimo, atribuir uma probabilidade diferente de 1/36 a cada uma das consequências possíveis de um lançamento numa roleta não-manipulada.

Mas a maioria das decisões em causa nos problemas de decisão relevantes são únicas num sentido muito mais forte do que este: qual a probabilidade de sucesso, caso decida licenciar-me em Física em vez de Filosofia? qual a probabilidade de ser feliz, caso decida casar com a Maria? Estas decisões verdadeiramente *únicas* tornam irrelevante o apelo à *lei dos grandes números*.

Um outro argumento a favor da utilização do princípio da utilidade esperada faz uso de um método axiomático. Se for pedido ao agente que estabeleça preferências relativamente a um conjunto de consequências alternativas, a obtenção das quais envolve um certo grau de incerteza, e se esse conjunto de preferências satisfizer um certo número de restrições estruturais, ou axiomas, é então possível demonstrar que as decisões do agente podem ser descritas *como se* este estivesse a escolher atribuindo probabilidades numéricas aos estados do mundo e valores de utilidade às consequências das acções, ou seja, a maximizar a sua utilidade subjectiva esperada. Isto significa que é possível demonstrar o seguinte teorema: se certas relações de preferência entre itens satisfizerem certos axiomas, então existe uma distribuição de probabilidades *P* e uma função de utilidade *U*, tal que se

a utilidade subjectiva esperada (*USE*) das acções disponíveis for *calculada* com base nessa distribuição e nessa função, um item *X* será preferido a um item *Y* se, e somente se, a *USE* (*X*) > *USE* (*Y*).

A questão que neste ponto se coloca é a seguinte: dada esta demonstração, o que exige de nós a teoria no que diz respeito ao exercício da nossa racionalidade prática? A ideia de cálculo expressa acima poderá induzir em erro, podendo levar a pensar que a teoria exige que os agentes efectuem necessariamente qualquer tipo de cálculo antes de agirem. Contudo, se a teoria exigisse qualquer tipo de cálculo, então seria desde logo implausível, do ponto-de-vista psicológico, atribuir-lhe não só valor descritivo, como também considerá-la normativamente adequada. Além disso, é perfeitamente possível demonstrar que, em certos casos, a teoria exige o contrário: que se aja espontaneamente. Isto porque calcular a utilidade é, em si mesma, uma acção para a qual se podem determinar consequências sob cada estado do mundo relevante. Como tal, num qualquer problema de decisão, a utilidade esperada da acção que consiste em calcular a utilidade esperada das outras acções não tem necessariamente de maximizar a utilidade subjectiva esperada. Consideremos o seguinte exemplo (Maher 1993: 5-6), em que um agente tem de decidir, ao sair de casa numa manhã cinzenta, entre três acções possíveis:

G: Levar um guarda-chuva, sem calcular a utilidade esperada.
¬ *G*: Sair sem um guarda-chuva, sem calcular a utilidade esperada
C: Calcular a utilidade esperada (*C*) de *G* e de ∼ *G*, e de seguida escolher, de entre estas, aquela que maximiza a utilidade esperada.

Os estados do mundo relevantes são os seguintes:

E1: $c \wedge (C \rightarrow G)$ (chove e o cálculo da utilidade conduz à escolha de *G*)
E2: $\neg c \wedge (C \rightarrow G)$
E3: $c \wedge (C \rightarrow \neg G)$
E4: $\neg c \wedge (C \rightarrow \neg G)$

Suponhamos, também, que as probabilidades subjectivamente atribuídas aos quatro estados e as utilidades de cada consequência são as seguintes:

pr (E1) = *pr* (E4) = 0.4; *pr* (E2) = *pr* (E3) = 0.1
u (¬ *c* ∧ ¬ *G*) = 4; *u* = (¬ *c* ∧ *G*) = 3; *u* (*c* ∧ *G*) = 2; *u* (*c* ∧ ¬ *G*) = 0

Sendo *x* o custo (ou desutilidade) de efectuar o cálculo (o agente prefere não perder tempo a calcular),

USE (*G*) = 2.5; *USE* (¬ *G*) = 2; *USE* (*C*) = 2.7 – *x*.

Conclui-se, assim, que, dependendo do agente, se para ele *x* for suficientemente elevado, a acção de calcular a utilidade esperada das acções disponíveis pode ter uma utilidade esperada menor do que uma ou ambas as acções espontâneas. A própria teoria pode, assim, recomendar que não se calcule a utilidade esperada; ou, de modo a acentuar a aparência de paradoxo, recomendar que ela própria não seja aplicada. Pode-se afirmar que efectuar o cálculo é racionalmente exigido apenas quando os seus benefícios são efectivamente maiores dos que os custos, desde que estejamos perante casos em que o valor *x* é claramente irrelevante.

De acordo com estes resultados, o principal objectivo da teoria não consiste propriamente em oferecer ao agente, no momento da escolha, e em todas as ocasiões, os meios que lhe assegurem a racionalidade das suas decisões, mas apenas um critério para determinar *a posteriori*, ou do ponto-de-vista de um terceiro, que acções são racionais e que acções não o são.

O que se poderá dizer favoravelmente desta perspectiva? O que está aqui em causa não é apenas o valor descritivo da teoria, mas também a preservação do seu valor normativo. Se não existir uma maneira de se saber *a priori* que realmente compensa efectuar os cálculos, não veremos esse valor ser severamente reduzido? Existirão certamente casos em que é muito plausível, ou até mesmo evidente, que os benefícios do cálculo ultrapassam os custos. Se trabalharmos numa empresa de seguros, então é fácil imaginarmo-nos sentados a uma secretária, rodeados de dados estatísticos, a efectuar cálculos. Contudo, se pretendermos que a teoria tenha um poder normativo considerável, desejaremos vê-la aplicada a uma gama muito maior de comportamentos humanos.

Mas se as exigências ou restrições de racionalidade que a teoria impõe não estiverem exclusivamente relacionadas com a necessidade do cálculo, sobre o que actuariam, então, essas exigências? Na justificação do princípio da utilidade através do método axiomático,

a possibilidade de representar as preferências do agente através de uma função de utilidade encontra-se dependente da satisfação de certos axiomas por parte do agente. A posse e a respectiva ordenação destas preferências constitui propriamente a componente espontânea, não calculada, da decisão do agente comum e, portanto, são o objecto daquilo que a teoria pretende explicar através da sua grelha matemática, de onde se segue que a explicação racional da acção tem de ter como objecto a estrutura dessas preferências, assim como qualquer injunção prática tem de se fazer exercer sobre o modo através do qual o agente as ordena.

Estas restrições estruturais não se tratam, propriamente, de condições que determinam a racionalidade, ou falta dela, dessas mesmas preferências. Estas em si, e do ponto de vista da racionalidade instrumental, não são racionais nem deixam de o ser, constituindo apenas os dados do problema.[18] Estes axiomas determinam, se quisermos, qual deverá ser a estrutura racional das ordenações de preferências do agente. Nessa medida, eles podem ser considerados como uma espécie de pré-requisitos da racionalidade instrumental. A questão fundamental acerca da justificação da teoria - 'Por que motivo é racional maximizar a utilidade esperada?' - pode ser reformulada da seguinte maneira: 'Por que devemos aceitar os axiomas da preferência?'

Neste ponto, uma analogia com as condições de racionalidade do raciocínio dedutivo poderá ser esclarecedora. O exemplo é de Joyce (1999: 81-82): poderá talvez ser razoável aceitar-se que um agente é epistemicamente racional se, e somente se, existir uma interpretação tarskiana da sua linguagem, a qual torne todas as suas crenças verdadeiras. Esta definição encerraria aquilo que se poderia designar como uma condição necessária e suficiente para a *consistência global* das crenças do agente. Não seria, contudo, razoável exigir do agente epistémico que este, ao adquirir, rever e rejeitar as suas crenças, tentasse seguir este princípio. O que pode ser exigido do agente é que este aplique uma série de regras de *consistência local* – lei da não-contradição, *modus ponens*, etc. – as quais são condições necessárias, e conjuntamente suficientes, da *consistência* global. Isto não é apenas razoável de uma perspectiva prática, por razões de expediente, mas também porque não é obviamente possível existir consistência global sem existir consistência a

[18] Esta perspectiva é designada como 'bayesianismo incondicional'. Uma outra, que pode ser designada como 'bayesianismo condicional' defende que a teoria não estabelece condições suficientes para a racionalidade instrumental, mas apenas condições necessárias Existiriam assim outras condições relacionadas com a aquisição de crenças e a posse de certas preferências.

nível local, e vice-versa. Mais importante, essas regras de aplicação local podem ser justificadas independentemente, através do recurso aos axiomas do cálculo de predicados. Tornando a analogia explícita: os teoremas da coerência e completude do cálculo de predicados de primeira ordem permitem demonstrar que um conjunto crenças tem um modelo tarskiano se, e somente se, esse conjunto não violar nenhuma das regras locais incorporadas nos axiomas do cálculo. Por seu lado, o teorema da representação (das crenças e desejos do agente) demonstra que um agente é racional se, e somente se, as suas preferências não violarem um conjunto de restrições estruturais ou axiomas. Assim, tal como é possível justificar independentemente as regras do cálculo de predicados, sem recurso aos teoremas da coerência e da completude, também deverá ser possível justificar independentemente os axiomas da preferência, sem recurso ao teorema da representação. Mas, afinal, que tipo de argumentos podem justificar axiomas? Não deveria a sua verdade ser auto-evidente? Sem dúvida que grande parte do seu apelo reside na sua plausibilidade intuitiva. Contudo, se essa verdade for auto-evidente, como se poderá explicar a sua violação? Será a irracionalidade prática explicada apenas através da cegueira intuitiva dos agentes? Se formos capazes de mostrar que existe um raciocínio que justifique a sua aceitação, ou, na medida em que eles são análogos a regras, um raciocínio que mostre as consequências negativas do seu incumprimento, teremos, então, estabelecido uma ligação entre tais regras e o valor normativo da teoria. Mais, teremos uma forma de explicar a irracionalidade prática dos agentes, a qual se encontra directamente relacionada com a cognição de razões e não somente com a cegueira intuitiva.

Com efeito, a justificabilidade dos axiomas prende-se directamente com a questão do valor normativo da teoria. Os argumentos que oferecem razões para se aceitar os axiomas são argumentos pragmáticos, directamente relacionados com a satisfação dos interesses dos agentes. Mais precisamente, são argumentos que tentam mostrar de que modo um agente que não respeite os axiomas da preferência pode ser induzido a comportar-se de maneira a ir contra os seus próprios interesses.

Exemplos típicos de argumentos pragmáticos são os *dutch book arguments*: argumentos que mostram as consequências negativas que terá de enfrentar um agente que atribua probabilidades violando os axiomas do respectivo cálculo.[19]

[19] Suponha-se que aceitamos apostar 100 Euros numa hipótese de 9/10 em *p* (*Brigadier Gerard* ganha a corrida); e, simultaneamente, 100 Euros numa hipótese de 1/2 em não-*p*. Se *p* for o caso, ganhamos 10 Euros com a primeira aposta e perdemos 50 na segunda; se não-*p* for o caso, ganhamos 50 na segunda, mas perdemos 90 na primeira. Em qualquer caso, perdemos sempre 40 Euros.

Considerem-se os dois conceitos comparativos fundamentais: ser 'melhor do que' ou 'preferido a', designado como *preferência estrita*; e ser 'igual em valor a' ou 'tão bom quanto', designado como *indiferença*. Corresponda '$\succ$' ao primeiro, e '$\approx$' ao segundo destes conceitos. Temos, assim, três dos principais axiomas da preferência, utilizados em demonstrações, como a de Savage, de teoremas da representação:

1) Transitividade: se $A \succ B$ e $B \succ C$, então $A \succ C$
2) Completude: $A \succ B$ ou $B \succ A$ ou $A \approx B$
3) Independência: se $A \succ B$, então $Ap C \succ Bp C$, em que $Ap C$ é uma lotaria que oferece A com uma probabilidade p e C com uma probabilidade $(1- p)$.

O axioma da transitividade diz-nos o seguinte: se um agente *x* prefere A a B, e B a C, então prefere A a C. De acordo com o axioma da completude, entre dois itens A e B, um qualquer agente *x* ou prefere A a B, ou B a A, ou é indiferente entre ambos. Finalmente, de acordo com o axioma da independência, se um agente *x* prefere A a B, então prefere uma lotaria que lhe oferece A com uma probabilidade *p* e C com uma probabilidade (1 – *p*) a uma outra lotaria que lhe oferece B com uma probabilidade *p* e C com uma probabilidade (1 – *p*), em que *p* tem sempre o mesmo valor.[20]

Irei discutir cada um destes axiomas com algum detalhe, mas um argumento pragmático que pode ser utilizado a favor de cada um deles é conhecido por argumento *money-pump*, e é assim designado porque pretende mostrar como pode um agente ser explorado ao ser-lhe extraída (bombeada) uma quantia indefinida de dinheiro.[21] Por exemplo, se alguém tiver uma ordenação cíclica de preferências,

[20] O modo como os axiomas são apresentados poderá variar, mas, para já, a formulação acima é suficiente. Convém também notar que os conceitos de *preferência estrita* e de *indiferença* não são por vezes utilizados como primitivos, sendo antes construídos, por uma questão de rigor matemático, a partir do conceito de *preferência fraca*: '$A \succcurlyeq B$', um objecto A é *fracamente preferido* a B se, e somente se, um agente *x* prefere estritamente A a B ou é indiferente entre ambos. Assim, um agente *x* é indiferente entre A e B se, e somente se, *prefere fracamente* A a B e B a A. E um agente *x prefere estritamente* A a B se, e somente se, *preferir fracamente* A a B, mas não o contrário. Assim, tem-se não apenas a transitividade da *preferência estrita*, mas também da *preferência fraca*, da *indiferença*, e de combinações entre as três. Como o argumento pragmático que irei considerar pode aplicar-se a todas essas combinações, incluindo a preferências intransitivas não-cíclicas, utilizarei daqui em diante a *preferência estrita* por me parecer assim mais intuitiva a formulação dos axiomas.

[21] Este argumento surge pela primeira vez em D. Davidson, J. McKinsey e S. Siegel (1957), tendo-o os seus autores desenvolvido a partir de uma ideia de Frank Ramsey.

$A \succ B \succ C \succ A$,

pode achar-se numa situação em que acabará por perder todo o dinheiro que possui. Suponhamos que A, B e C são selos e um colecionador encontra-se na posse de A. É razoável assumir que existe uma certa quantia de dinheiro, digamos 1 cênt., que o coleccionador se encontra disposto a pagar para trocar A por C, C por B, ou B por A. O coleccionador entra numa loja de filatelia e o vendedor oferece-lhe a oportunidade de trocar A por C, na condição de pagar 1 cênt. O colecionador, de acordo com a sua ordenação de preferências, aceita o negócio. De seguida, o vendedor sugere-lhe trocar C por B, a troco de 1 cênt. Mais uma vez, o colecionador aceita. Por fim, é-lhe sugerido trocar B por A, pagando 1 cênt. Aqui, o colecionador poderá talvez aperceber-se de que está a ser bombeado e decide recusar a oferta. De qualquer modo, ele vê-se perante uma situação em que pode ou ficar com um selo de menor valor que o original, menos 2 cênts., ou aceitar a troca e ficar com o mesmo selo com que entrou, mas com menos 3 cênts. Neste caso, após 1 milhão de trocas, o colecionador perderá 10.000 Euros. Se quisermos dramatizar o argumento, podemos conceber uma situação em que a diferença de valor entre cada selo é muito maior e, em que, dessa maneira, serão necessárias muito menos trocas para levar o colecionador à falência. Poderíamos ainda espaçar as trocas no tempo para oferecer maior plausibilidade ao exemplo.

Haverá alguma maneira de mostrar que a transitividade não é uma condição de racionalidade das nossas preferências? Existe, pelo menos, uma situação em que é possível violar a transitividade e, no entanto, preservar a intuição de plausibilidade associada à nossa ordenação de preferências.

Quando $A \succ B$ e $B \succ C$, mas A e C são incomensuráveis (ou incomparáveis), não se segue que $A \succ C$. Por exemplo, se A for uma certa quantia de dinheiro e C corresponder a um estado de saúde ou bem-estar, é razoável acreditar que A e C possam ser, para certos agentes, incomensuráveis. O único problema com este ataque à transitividade reside no facto de negar outro dos axiomas necessários à racionalidade da acção, a completude. Ou seja, aparentemente podemos ter violações da transitividade, embora estas venham sempre acompanhadas de uma violação da completude.

A completude tem na sua base duas teses que o defensor da incomparabilidade tem de negar. Mais uma vez, a intuição da sua plausibilidade é variável:

1) Tricotomia: apenas uma de entre três relações comparativas se verifica entre duas opções ou itens, melhor do que', 'pior do que' e 'tão bom como ou idêntico a'.
2) Comparabilidade: duas opções ou itens são comparáveis se, e somente se, uma das relações expressas na tese da tricotomia se verificar.

A questão da incomparabilidade tem surgido não apenas no contexto da discussão das teorias da agência racional, mas também no domínio da ética (Joseph Raz 1986). É precisamente neste domínio que os supostos exemplos de incompletude se tornam mais dramáticos e verosímeis. Pode-se achar, por exemplo, que o valor de uma vida humana não é comparável a uma certa quantia de dinheiro. Mas, se aceitarmos o axioma, temos aparentemente de aceitar que existe um ponto exacto em que é preferível não prescindir de uma determinada quantia a salvar uma vida humana. Por exemplo, quando prescindir dessa quantia pode pôr em causa a salvação de outras vidas humanas. (Embora seja argumentável que aqui já não estamos a comparar dinheiro com vidas humanas, mas vidas humanas com vidas humanas).

Existirá, neste caso, algum argumento pragmático que sirva para justificar o axioma da completude? Tal depende do modo como entendemos que a incomensurabilidade se relaciona com as nossas atitudes de escolha e troca. Se acharmos que faz sentido, do ponto de vista do agente, efectuar trocas entre objectos incomensuráveis, então é obviamente possível conceber-se um argumento *money-pump* para a completude. Perante a seguinte ordenação de preferências, em que '*I*' representa a relação de incomparabilidade,

$$A \succ B,\ B \succ C,\ A\ I\ C,$$

se o agente estiver na posse de A e aceitar trocar A por C (se estes objectos são incomparáveis, supõe-se que qualquer pagamento extra não fará diferença), então estará disposto a pagar, de seguida, 1 cênt. para trocar C por B. Mas, como A é preferido a B, ele aceita trocar B por A, pagando mais um cênt. E assim sucessivamente.

Contudo, e face aos exemplos éticos mais dramáticos, não fará parte do conceito mesmo de incomparabilidade a recusa peremptória em efectuar qualquer troca entre um par de objectos assim classificados? Ou será isto uma ilusão proveniente do facto de considerarmos a vida humana, como no exemplo acima, muito mais valiosa do que

qualquer quantia de dinheiro, e não estarmos dispostos a trocar a primeira pela segunda? É difícil de dizer.

Convém, todavia, notar que a teoria da preferência que se encontra na base da ciência económica, e que tem vindo há décadas a ser utilizada para cálculos de utilidade em situações de escolha social, a chamada teoria da *preferência revelada* (Samuelson 1938), não aceita casos de incomparabilidade ou, como também se designa, de incompletude irresolúvel. Por razões que considerámos atrás, uma teoria da preferência que tenha por base pressupostos comportamentalistas, como a teoria de Samuelson, verá o seu valor explicativo bastante reduzido. Mas, para já, é mais importante considerarmos aquele que, a meu ver, é o argumento em redor da discussão do qual se encontra dependente a plausibilidade do axioma da completude, o argumento do *pequeno acréscimo*.

Considere-se uma situação em que um determinado agente nem prefere estritamente 10 Milhões de Euros a salvar uma vida humana, nem prefere estritamente salvar uma vida humana a 10 M Euros. De acordo com a lógica das preferências, se o agente for, de facto, indiferente entre duas alternativas – se considerar que elas têm igual valor – então uma ínfima melhoria numa delas implicará um ínfimo acréscimo de utilidade. Supõe-se, portanto, que o agente passará a preferir a alternativa que foi infimamente melhorada. Contudo, é razoável supor que um tal acréscimo não alterará as preferências do agente, seja ele quem for, no caso considerado. Logo, o agente não era realmente indiferente perante as duas alternativas. Mas, se não era indiferente, e se não preferia uma delas à outra, parece legítimo concluir-se que para ele as duas alternativas eram, de facto, incomparáveis. Por outras palavras, conclui-se que a tese da tricotomia é falsa.[22]

Espinosa (2008) apresentou a seguinte representação formal deste argumento, em que *'T*' significa 'tão bom quanto' e 'B+' significa 'B mais um pequeno acréscimo':

1. $\neg (A \succ B) \wedge \neg (B \succ A)$, esta premissa diz-nos que o agente nem prefere A a B, nem B a A.
2. $B+ \succ B$, esta premissa diz-nos que o agente prefere B mais um pequeno acréscimo a B sem esse pequeno acréscimo.

[22] Foi o próprio Savage (1954: 17) quem sugeriu este 'teste da indiferença': (If) the person really does regard f and g as equivalent, that is, if he is indifferent between them, then, if f or g were modified by attaching an arbitrarily small bonus to its consequences in every state, the person's decision would presumably be for whichever act was thus modified'.

3. $((A\ T\ B) \wedge (B+ \succ B)) \rightarrow (B+ \succ A)$, esta premissa expressa a hipótese de que um pequeno acréscimo numa das alternativas modifica a preferência do agente.
4. $\neg\ (B+ \succ A)$, esta premissa nega o consequente da hipótese expressa em 3.
5. $\neg\ ((A\ T\ B) \wedge (B+ \succ B))$ 3, 4 Modus Tollens
6. $\neg\ (A\ T\ B)$ 2,5 DeM, D. Neg. e Sil. Disjuntivo
7. $\neg\ (A \succ B) \wedge \neg\ (B \succ A) \wedge \neg\ (A\ T\ B)$ 1,6 Intr. Conjunção

Ora, 7 diz-nos aquilo que o argumento tenta estabelecer: que nenhuma das três relações valorativas se verifica entre A e B. Mas, se é certo que o argumento nega a tese da tricotomia, ele não estabelece que essa quarta opção deixada em aberto seja necessariamente a incomparabilidade. Embora haja uma forte sugestão no sentido de se aceitar a hipótese da incomparabilidade, dada a dificuldade em conceber outra relação valorativa para além das três já consideradas, existe uma outra possibilidade, segundo a qual, por motivos relacionados com a vagueza das relações comparativas, o agente não é capaz de aceitar *determinantemente* a verdade do terceiro conjunto de 7.

Espinosa pretende explorar esta possibilidade, concebendo precisamente um contra-argumento com o intuito de demonstrar a possibilidade da indeterminação ou vagueza nas relações comparativas. Para o efeito, ele introduz as noções de *verdade indeterminada* e *verdade determinada* (Fine 1975). Eu acrescento, apenas, que estas noções não nos obrigam a comprometer-nos com uma lógica não-bivalente, mas dizem apenas respeito a uma atitude proposicional do agente, caracterizada, se quisermos, pela dúvida ou recusa em classificar como *determinantemente* verdadeira ou falsa uma frase como 'o António é alto', quando o António tem 1m76. Por outro lado, o agente está disposto a classificar como *determinantemente* verdadeira uma frase como 'o José é alto', quando o José tem 2m. Ou seja, representando '*I*' a qualidade de algo ser *indeterminado* (e '*D*' a de algo ser *determinantemente* verdade), $I\varepsilon \leftrightarrow (\neg\ D\varepsilon \wedge \neg\ D\ \neg\varepsilon)$: ε é *indeterminado* se, e somente se, não for *determinantemente verdade* que ε, nem *determinantemente falso* que ε.

Utilizando os devidos prefixos, ou classificando as relações comparativas como *determinantemente verdadeiras* ou *falsas*, o argumento do *pequeno acréscimo* pode ser reformulado da seguinte maneira:

1'. $D \neg (A \succ B) \wedge D \neg (B \succ A)$

2'. $D (B+ \succ B)$

3'. $(D (A \; T \; B) \wedge D (B+ \succ B)) \rightarrow D (B+ \succ A)$

4'. $D \neg (B+ \succ A)$

5'. $D \neg (D (A \; T \; B) \wedge D (B+ \succ B)$

6'. $\neg D (A \; T \; B)$

7'. $D \neg (A \succ B) \wedge D \neg (B \succ A) \wedge \neg D (A \; T \; B)$ (Espinosa 2008: 131).

A conclusão 7' diz-nos o seguinte: é *determinantemente falso* que A é preferido a B, é *determinantemente falso* que B é preferido a A, e não é *determinantemente verdade* que A seja tão bom quanto B. O terceiro conjunto da conclusão, $\neg D (A \; T \; B)$, abre realmente a possibilidade de (A *T* B) ser *falso* para um determinado agente, mas não necessariamente. O contra-argumento de Espinosa depende, portanto, da seguinte verdade: $\neg D (A \; T \; B) \leftrightarrow (D \neg (A \; T \; B) \vee I (A \; T \; B))$.[23] Ou seja, não só é possível ser determinantemente falso que A seja tão bom quanto B, o que satisfaz a pretensão dos críticos da completude, como também pode ser verdade que, para um determinado agente, a relação de preferência entre *A* e *B* não esteja bem determinada. Isto significa que, dada a verdade de 7', uma das seguintes tem de ser verdadeira:

8. $D \neg (A \succ B) \wedge D \neg (B \succ A) \wedge D \neg (A \; T \; B)$
9. $D \neg (A \succ B) \wedge D \neg (B \succ A) \wedge I (A \; T \; B)$

A questão que deve ser aqui colocada consiste em saber onde se insere neste contra-argumento a noção de vagueza e como esta se encontra relacionada com a ideia da *indeterminação* da relação de 'ser tão bom como'. A conclusão que Espinosa procura estabelecer é a de que não só o argumento do *pequeno acréscimo* é incapaz de distinguir entre *incomparabilidade* e *indeterminação*, mas também que a vagueza da distinção qualitativa entre itens pode ser a causa da indeterminação da relação de *'ser tão bom*

[23] Se considerarmos as três seguintes relações comparativas como mutuamente exclusivas e conjuntamente exaustivas, $D\varepsilon$, $D \neg\varepsilon$, $I\varepsilon$, então uma extensão trivial da lei do terceiro excluído diz-nos que $\neg D\varepsilon$ se, e somente se, $D \neg\varepsilon \vee I\varepsilon$.

como'. Como existem várias possibilidades de explicação para a vagueza, a conclusão do seu contra-argumento encontraria uma sustentação sólida num fenómeno bem identificado. Em suma, *incomparabilidade* e *indeterminação* são conceitos diferentes. Enquanto o primeiro nega a possibilidade da indiferença, o segundo estabelece a sua possibilidade, na medida em que o agente é incapaz de fazer uma distinção entre alternativas quanto à sua característica relevante. Por seu lado, essa incapacidade encontraria explicação no fenómeno da vagueza. Repare-se que quando duas coisas são por natureza incomparáveis, sê-lo-ão para sempre, caso se mantenham inalteradas; por exemplo, duas vidas humanas em particular. Contudo, se duas coisas forem 'incomparáveis' por motivos de vagueza, se houver maneira de dissipar essa vagueza – por exemplo, aguçando a capacidade cognitiva ou perceptiva do agente – será possível ao agente determinar que é realmente indiferente ou que, afinal, prefere uma alternativa à outra.

Contudo, para que o contra-argumento de Espinosa seja producente, é necessário estabelecer que 9 pode ser verdadeira, i.e., que I (A T B) é compatível com as preferências iniciais do agente, $D \neg (A \succ B) \wedge D \neg (B \succ A)$. Para isso é necessário analisar o conceito expresso por I (A T B).

A meu ver, quando é verdade que I (A T B), a qualidade de uma relação de ser indeterminada aplica-se não apenas à relação de 'ser tão bom como', mas também à relação de 'ser preferido a', quando respeitante aos mesmos itens. A minha sugestão é a seguinte: da indeterminação da indiferença entre dois itens segue-se que pelo menos um dos conjuntos da premissa 1' é falso. Do facto de ser indeterminado para um agente que um objecto é tão bom quanto outro, segue-se que ele não pode ter para si simultaneamente como *determinantemente* falso que (A $\succ$ B) e que (B $\succ$ A); e também não se segue, claro, que pelo menos um tenha de ser melhor do que outro, pois poderá vir a saber-se que ambos têm o mesmo valor. Por outro lado, se um agente não tem a certeza se prefere A a B ou B a A, parece seguir-se que também não tem a certeza que A seja tão bom quanto B. A definição de I (A T B) passa a ser a seguinte:

$$10.\ I(A\ T\ B) \leftrightarrow ((D \neg (A \succ B) \wedge I (B \succ A)) \vee$$
$$\vee (D \neg (B \succ A) \wedge I (A \succ B)) \vee$$
$$\vee (I (A \succ B) \wedge I (B \succ A)))$$

Ou seja, se é indeterminado que A é tão bom quanto B, então pode verificar-se qualquer um dos três seguintes casos: 1) apesar de o agente ter a certeza de que A não é melhor do que B, ele não sabe se B é melhor do que A ou se B é tão bom como A; 2) apesar de o agente ter a certeza de que B não é melhor do que A, ele não sabe se A é melhor do que B ou se A é tão bom quanto B; 3) é possível que qualquer uma das três relações comparativas se verifique - (A ≻ B), (B ≻ A) ou (A *T* B) – ou que nenhuma delas se verifique (e as preferências do agente sejam incompletas).[24]

O problema encontra-se, agora, no facto de precisarmos de uma nova formulação do argumento que nos permita obter uma conclusão análoga a 7', mas com as devidas alterações do âmbito do operador *D*. Considere-se, por exemplo, que as seguintes são as preferências iniciais do agente:

11. $D \neg (A \succ B) \wedge \neg D (B \succ A)$

Uma das premissas do contra-argumento de Espinosa é que $D \neg (B+ \succ A)$. O problema torna-se imediatamente claro. Se, para um agente, não é *determinantemente verdadeiro* que B é melhor do que A, é-lhe ainda assim possível saber que é *determinantemente verdadeiro* que B+ não é melhor do que A? Não me parece que isso seja possível, pois afirmar $\neg D (B \succ A)$ é admitir a possibilidade de $I (B \succ A)$, o que, por sua vez, é admitir a possibilidade de $B \succ A$. E se $B \succ A$, então certamente $B+ \succ A$. Logo, se aceitarmos a verdade de $\neg D (B \succ A)$, temos de negar a verdade da premissa 4' do contra-argumento:

4''. $\neg D \neg (B+ \succ A)$

Dada 4'', temos que

4'''. $\neg D \neg (B+ \succ A) \rightarrow ((D (B+ \succ A) \vee I (B+ \succ A))$

[24] Se aceitarmos 10, e nos lembrarmos que $I (A \succ B) \rightarrow (\neg D (A \succ B) \wedge \neg D \neg (A \succ B))$ e que $I (B \succ A) \rightarrow (\neg D (B \succ A) \wedge \neg D \neg (B \succ A))$, verificamos que $\neg(D \neg (A \succ B) \wedge D \neg (B \succ A) \wedge I (A \, T \, B))$.

Podemos ver que, mesmo admitindo a verdade da premissa 3' – (D (A T B) ∧ D (B+ ≻ B)) → D (B+ ≻ A) -, nenhum dos disjuntos do consequente de 4''' é a negação do consequente de 3'. A negação de que é *determinantemente verdade* que B+ ≻ A é a afirmação de que é *determinantemente falso* que B+ ≻ A. Por outro lado, se ¬D (B+ ≻ A) fizer parte das preferências iniciais do agente, em vez de 4', tambem não é possível fazer *modus tollens* com 3', o que anula o contra-argumento de Espinosa.[25] Pela mesma ordem de razões, o contra-argumento também é anulado caso as preferências iniciais do agente sejam as seguintes:

12. ¬ D (A ≻ B) ∧ D ¬ (B ≻ A)

Resta, portanto, a terceira alternativa, ou modo de tornar as preferências do agente compatíveis com I (A T B):

13. I (A ≻ B) ∧ I (B ≻ A)

Se 13 substituir 1', então, dada a definição em 10, podemos constatar que estas preferências são uma condição suficiente de I (A T B) e que, portanto, estamos perante uma petição de princípio, i.e., I (A T B) é um dado de partida do contra-argumento de Espinosa.

Só parecem existir duas maneiras de defender a possibilidade da *indeterminação* de (A T B) e, nessa medida, defender o axioma da completude: assumir que esta constitui um dado

[25] A chave aqui é compreender uma importante distinção. Por um lado, podemos considerar a compatibilidade entre ¬ D (B+ ≻ A) e D ¬ (B+ ≻ A) *simpliciter*; ou seja, dado que ¬ D (B+ ≻ A) → (D ¬ (B+ ≻ A) ∨ I (B ≻ A)), a compatibilidade existe. Ora, esta compatibilidade não é do tipo epistémico. Apesar de o agente saber que, da sua crença de que não é *determinantemente verdade* que (B+ ≻ A), algo se segue - nomeadamente, a disjunção acima -, ele ainda não sabe, contudo, exactamente o quê. Para que ele o saiba, alguma indeterminação terá de se evaporar. Por exemplo, se, *para o agente*, ¬D (B+ ≻ A), então, caso seja eliminada alguma indeterminação, ele pode vir a concluir/descobrir que, *para si*, D ¬ (B+ ≻ A). É neste sentido que a crença em ¬ D (B+ ≻ A) não é compatível (não pode ser sustida simultaneamente) com a crença em D ¬ (B+ ≻ A). Os únicos casos em que não se verifica qualquer indeterminação das relações comparativas entre dois itens são aqueles em que uma relação de preferência é *determinantemente verdadeira* ou *falsa*. Contudo, do ponto de vista epistémico, a última é distinta da primeira. Se, quando a primeira se verifica – e.g. D (B ≻ A) -, nada mais há para saber, já quando a segunda é o caso - D ¬ (B ≻ A) -, ainda falta saber (para nós e talvez para o próprio agente) se A é melhor do que B, ou se os dois itens têm o mesmo valor.

primitivo que não necessita de demonstração ou defender que *I* (A *T* B) é perfeitamente compatível com as preferências iniciais do agente. A primeira destas opções não parece muito viável, dada a consistência e a convicção aparente das crenças de inúmeros agentes na incomparabilidade de certos itens, por exemplo, nos casos de duas vidas humanas ou nos casos de escolha social, tal como o da alocação de meios limitados para o tratamento de doenças igualmente graves, etc. A segunda opção é talvez a mais viável, embora, como foi visto acima, não me pareça muito plausível.

Importa notar que o contra-argumento de Espinosa tem como intuito demonstrar que o *argumento do pequeno acréscimo* não estabelece a falsidade da tese da tricotomia. Apesar disso, Espinosa encontra-se não só comprometido com a ideia de que todos os casos de indeterminação são casos genuínos de vagueza, mas também que, dadas as condições ideais, o agente poderia desfazer essa indeterminação e aplicar sempre o princípio da maximização da utilidade esperada. Não lhe basta, portanto, dizer que existe um pequeno número de casos em que não existe realmente genuína indeterminação, mas sim genuína incomparabilidade. Ou seja, se após a resolução da vagueza - e, *a fortiori*, da indeterminação - se segue que é *determinantemente* verdade que, afinal, não é o caso que A seja tão bom quanto B, então o contra-argumento de Espinosa não alcançou o seu objectivo.

De qualquer modo, talvez haja uma maneira de mostrar que a hipótese da indeterminação, e mais especificamente da vagueza, é mais plausível que a da incomparabilidade. Suponhamos que em vez de um pequeno acréscimo a uma das alternativas, efectuamos um grande acréscimo? Será possível efectuar, em todos os casos, um acréscimo tão grande quanto necessário para fazer com que o agente torne as suas preferências determinadas? Não é possível responder de uma forma definitiva a esta questão, mas em vários casos tal parece perfeitamente possível. Assim, quanto maior for a *área de penumbra* (onde se faz sentir a vagueza), maior terá de ser o acrescento para que o agente passe a preferir uma alternativa a outra. Se a área de penumbra for relativamente pequena, então um pequeno acréscimo talvez seja suficiente.

Resumindo, o argumento do *pequeno acréscimo*, face ao falhanço da reformulação de Espinosa, parece realmente manter a sua força. Mas, face à negação da tese da tricotomia, o que resulta daí para a força normativa da teoria? Richard Jeffrey (1964), e a maioria dos autores na sua peugada, sugeriu que a completude não é um requisito de racionalidade instrumental dos agentes, mas que essa racionalidade apenas requer destes que as suas

preferências sejam coerentemente extensíveis [*coherently extendable*]. Ou seja, mesmo que as preferências não sejam completas, deve ser possível completá-las sem que o agente viole quaisquer outros dos axiomas. De outra forma, mesmo que a relação de preferência não possa estabelecer uma ordenação completa do domínio das consequências possíveis, tem de existir pelo menos um ranking de preferências completo que satisfaça os axiomas da teoria.

Isto significa, portanto, que um agente pode ser tido como racional, mesmo que o valor relativo dos seus desejos não se preste sempre a ser representado numericamente por uma escala de intervalos. Resulta daí que a completude já não constitui uma condição necessária da racionalidade instrumental. A versão com que ficamos da teoria da utilidade esperada resulta, assim, mais fraca: se um agente racional prefere *x* a *y*, então $u(x) > u(y)$. Ou seja, é possível que a $u(x)$ seja maior que a $u(y)$ e que um agente racional não prefira *x* e *y*. Mais precisamente, é possível que o agente não tenha qualquer preferência entre estes dois itens.

3.2. Transitividade

O axioma da transitividade assume particular importância no contexto da discussão em redor das teorias normativas da decisão. Essa importância advém do facto de ser uma condição necessária para a existência de uma função ordinal de utilidade *u*, tal que, para todo o *x* e *y*, a $u(x) \geq u(y)$ se, e somente se, *x* for preferido a (ou tão bom como) *y*, e uma condição suficiente, caso o número de alternativas seja finito ou enumerável.[26]

Este princípio pode ser apresentado das duas seguintes maneiras:

$\forall x \forall y \forall z\ ((xPy \wedge yPz) \rightarrow xPz)$ (transitividade da preferência estrita),

$\forall x \forall y \forall z\ ((xPy \wedge yIz) \rightarrow xPz)$ (transitividade da indiferença).

O argumento pragmático favorável à transitividade mostra que alguém cuja ordenação de preferências entre *x*, *y* e *z* não respeita um ou outro dos princípios acima, nomeadamente no caso de *xPy*, *yPz* e *zPx*, pode vir a ser manipulado e transformar-se numa *money-pump*.

[26] Se as preferências não forem transitivas, então não podemos ter uma ordem. Se não temos uma ordem, as preferências não podem ser representadas pelo sistema de números reais, da menos desejável para a mais desejável.

Este argumento aplica-se quando as preferências do agente são cíclicas, ou seja, quando este se encontra racionalmente comprometido a trocar *x* por *y*, *y* por *z* e *z* por *x*. Contudo, o agente pode ter uma ordenação de preferências não-cíclica e, no entanto, violar na mesma um dos princípios acima. Para que o argumento *money-pump* fique completo, torna-se necessário mostrar que é possível converter violações não-cíclicas da transitividade da indiferença em ciclos de preferência estrita. As seguintes ordenações correspondem a dois tipos de violação não-cíclica:

$xPy \wedge yPz \wedge xIz$,
$xPy \wedge yIz \wedge xIz$.

Considera-se que o agente se encontra racionalmente *comprometido* a trocar um item por outro, caso se verifique entre ambos uma relação de preferência estrita. Quando o agente é indiferente entre dois itens, considera-se que é racionalmente *permitido* ao agente trocar um pelo outro. Ou seja, o agente não é racionalmente *obrigado*, nos dois exemplos acima, a trocar *x* por *z*, bloqueando, se assim o desejar, a possibilidade de ser transformado em *money-pump*.[27]

O argumento mais forte, a meu ver, contra a intransitividade acíclica de preferências é apresentado por Gustafsson (2010). Este argumento depende da verdade do princípio da dominação, o qual é bastante menos controverso que o da transitividade. Suponha-se que o agente possui uma ordenação de preferências como aquelas que são acima apresentadas; por exemplo, $aPb \wedge bIc \wedge aIc$. É, então, possível construir três lotarias diferentes, *L1*, *L2* e *L3*, em que o agente recebe, com uma determinada probabilidade, *a*, *b* ou *c*, de acordo com a verificação de um ou outro de três estados do mundo, *E1*, *E2* ou *E3* (os três formam uma partição), a obtenção dos quais é independente da escolha de lotaria:

	E1	*E2*	*E3*
L1	*a*	*b*	*c*
L2	*b*	*c*	*a*
L3	*c*	*a*	*b*

[27] Tal como no caso da indiferença, também quando se verifica a incompletude das preferências, o agente não se encontra racionalmente comprometido com a troca um item por outro. Daí que o argumento *money--pump* contra a incompletude não seja tão convincente quanto o *money-pump* contra a intransitividade.

O princípio da dominação diz-nos, portanto, que é racional preferir estritamente uma lotaria *L* a outra *L'*, caso exista pelo menos um estado em que *L* é estritamente preferida a *L´* e não exista qualquer estado em que *L* não seja pelos menos tão boa quanto *L'*. Será fácil de constatar que, seja qual for a distribuição de probabilidades ao longo da partição, *L1* é sempre preferida a *L2*; *L2* é sempre preferida a *L3*; e *L3* é sempre preferida a *L1*. Portanto, o agente pagará uma pequena soma para trocar *L3* por *L2*; outra pequena soma para trocar *L2* por *L1* e outra pequena soma para trocar *L1* por *L3*, voltando, assim, ao ponto de partida. Se considerarmos convincente o argumento *money-pump* contra a intransitividade, teremos, assim, uma boa base para defendermos a racionalidade do princípio (verificaremos, mais adiante, se tal é o caso). Contudo, existem casos de violação da transitividade, os quais, segundo certos autores, não revelam necessariamente irracionalidade por parte do agente.

Na formação de preferências, as comparações entre alternativas podem ser unidimensionais ou multidimensionais. Quando temos de optar, por exemplo, entre comer uma laranja, uma maçã ou uma banana, o que conta para a formação de preferências pode consistir apenas numa só característica destes frutos. A mais óbvia será o seu sabor, embora seja plausível que outras características possam isoladamente ter a sua influência; por exemplo, o seu contributo para uma exigência nutricional específica. Contudo, as relações de preferência podem envolver comparações que abrangem várias dimensões ou características das alternativas. Nestes casos, a lógica não garante a transitividade dessas relações de preferência e é, portanto, neles que ocorrem com maior facilidade violações da transitividade.

Tversky (1969) conduziu experiências que mostram precisamente como esses casos podem surgir, mesmo entre indivíduos sofisticados, tais como estudantes universitários ou executivos de empresas. Os sujeitos das experiências foram confrontados com diversos pares de alternativas e só apenas no final tomaram conhecimento dos resultados. Uma característica em comum aos vários casos foi que a maioria dos sujeitos violadores da transitividade reconheceu tal violação como um erro e desejou alterar as suas escolhas. Contudo, uma minoria manteve-se fiel às suas escolhas intransitivas, fazendo notar, precisamente, que o peso conferido a cada uma das dimensões pode variar consoante os diferentes pares comparados.

Uma das experiências de Tversky consistiu em seleccionar candidatos a um cargo de professor na universidade. Os candidatos foram avaliados de acordo com as seguintes três

características: competência intelectual, estabilidade emocional e facilidade no convívio social. A tabela abaixo é uma versão simplificada do ranking dos candidatos apresentado por Tversky (na versão original havia dez candidatos):

	Dimensões		
Candidatos	I	E	S
a	66	90	95
b	67	77	80
c	69	70	75
d	71	63	66
e	73	58	59

Os sujeitos foram informados de que a dimensão mais importante era, naturalmente, a competência intelectual dos candidatos, mas que as outras dimensões também seriam consideradas. Os perfis foram deliberadamente construídos de modo a existir uma correlação negativa entre os valores da dimensão I e os valores das dimensões E e S. O que sucedeu foi o seguinte: quando as comparações eram feitas entre dois candidatos adjacentes, havia uma tendência para escolher o candidato com uma pontuação menor na dimensão I. Os sujeitos consideraram que uma diferença pequena na dimensão I era compensada pelas diferenças maiores nas dimensões E e S. Contudo, quando as comparações eram feitas entre candidatos mais afastados entre si, ou seja, quando as diferenças relativas à dimensão I eram maiores, o candidato normalmente escolhido era o que tinha uma maior pontuação em I. A obtenção de intransitividades resulta facilmente destas tendências de escolha; por exemplo, $dPe \wedge cPd \wedge bPc \wedge aPb$, mas ePa, gerando-se, assim, um ciclo.

Mas será esta ordenação de preferência necessariamente irracional? Tversky acha que não. Segundo ele, quando a dificuldade (ou o custo) de avaliar as alternativas e a possibilidade de cometer erros são demasiado elevadas, pode ser útil utilizar-se um método de selecção (como aquele que foi utilizado pelos sujeitos) que simplifica essa avaliação, ainda que este possa conduzir a intransitividades. Estes métodos, segundo Tversky, poderão fazer com que nos aproximemos das nossas verdadeiras preferências, permitindo-nos efectuar escolhas de uma forma mais ponderada; na situação acima, por exemplo, poder-se-á

escolher um candidato que esteja a meio caminho entre os dois que se situam nos extremos que dão origem ao ciclo. Assim, partindo do pressuposto de que o mundo não conspira contra nós para se aproveitar das nossas violações da transitividade, poderemos minimizar os nossos erros. Por outro lado, na origem da intransitividade poderá estar uma alteração das nossas preferências à medida que vamos fazendo as nossas comparações; ou seja, os sujeitos da experiência poderão concluir, ao chegar ao último par, que foram dando uma excessiva importância às dimensões E e S, o que lhes permitirá determinar com maior precisão a importância relativa de cada uma das dimensões.

Neste ponto, a única justificação para a racionalidade da intransitividade parece resumir-se à dificuldade que os agentes têm de lidar com a informação, seja por esta surgir em demasiada quantidade, seja por ser demasiado complexa, ou devido aos custos do seu processamento. Mas se este é o caso, por que não avaliar as alternativas de acordo com apenas um único critério, em vez de três? O que parece estar em causa é a importância do problema de decisão; ou seja, se se tratar de uma decisão realmente importante, então talvez os agentes tivessem criado um algoritmo que lhes permitisse lidar com o valor relativo de cada uma das dimensões relevantes. O exemplo não mostraria, assim, que num problema de decisão suficientemente importante é por vezes racional ter preferências intransitivas.

Um outro tipo de violação da intransitividade no contexto multidimensional ocorre quando a aplicação da regra da maioria resulta no chamado Paradoxo de Condorcet. Suponha-se que pretendemos escolher um de entre tês automóveis, os quais são avaliados de acordo com três dimensões, e que aplicamos a regra da maioria para o fazer, ou seja, o automóvel que for superior num maior número de dimensões será o escolhido:

	Segurança	Velocidade	Estética
Volkswagen	2	1	3
Volvo	3	2	1
Mercedes	1	3	2

O Volvo é preferido ao Volkswagen porque é-lhe superior em segurança e em velocidade; o Mercedes é preferido ao Volvo porque é-lhe superior em velocidade e estética; e o Volkswagen é preferido ao Mercedes porque é-lhe superior em segurança e estética. Encontra-se, assim, gerado o ciclo intransitivo.

Existem, contudo, soluções para resolver este aparente paradoxo. Em primeiro lugar, um raciocínio deste tipo apenas faz sentido quando temos a certeza de que as várias dimensões têm o mesmo peso. Como sabemos, por exemplo, que a diferença entre o Volkswagen e o Volvo não é suficientemente grande para contrabalançar o peso das outras duas dimensões? Caso exista essa diferença, o problema torna-se idêntico ao dos candidatos e cabe ao decisor encontrar o algoritmo que corresponda a uma caracterização precisa das suas preferências. Caso o valor relativo das dimensões seja idêntico, por que não ser indiferente entre os três automóveis? Afinal, cada um deles é igualmente bom em dois dos aspectos desejados, os três contribuindo igualmente para a felicidade do agente, gerando-se um círculo de indiferença que não constitui qualquer ameaça à transitividade. Mais plausível é hipótese de estes três itens serem incomensuráveis, não tendo o agente qualquer ideia do valor relativo das características. Isto faz com que os três itens sejam incomparáveis entre si, indicando não uma relação de indiferença, mas sim uma ausência de preferência, ou seja, as preferências do agente são incompletas, o que constitui uma violação do axioma da completude e não da transitividade. Ele fará naturalmente a sua escolha, embora isso não signifique que ele prefira o item escolhido aos outros dois.

Por outro lado, um agente avalia alternativas unidimensionalmente quando estas são comparadas entre si de acordo com apenas um critério. No contexto da avaliação unidimensional, os casos mais graves de intransitividade dão-se quando as preferências são formadas de acordo com a aplicação de um predicado vago. Importa notar, contudo, que não são as próprias preferências que são vagas; trata-se sim de formar preferências relativamente a itens aos quais se aplicam predicados vagos. Considere-se o predicado ‘ser tarde’ e os seguintes pressupostos: a) um agente prefere levantar-se de manhã no minuto *x* exactamente anterior àquele em que já será tarde demais para se levantar do que num minuto *x-1* exactamente anterior ao minuto *x*; 2) uma diferença de um minuto *x* para um minuto *x+1* não faz com que *x+1* seja tarde, caso *x* não seja tarde. Portanto, se oito da manhã não é tarde, então oito e um minuto não é tarde; se oito e um minuto não é tarde, então oito e dois minutos não é tarde. Dando continuidade a este raciocínio, através da aplicação sucessiva de *modus ponens*, chegamos a um ponto em que, por exemplo, a proposição ‘oito e trinta minutos não é tarde’ é dada como verdadeira. Sabemos, contudo, que oito e trinta é tarde; se nos levantarmos a essa hora, temos a certeza de que chegaremos tarde ao trabalho. Dada esta contradição, é fácil constatar que a ordenação de

preferências do agente quanto à hora de se levantar chegará a um ponto em que se tornará intransitiva; por exemplo:

(8+1) *P* (8) ∧ (8+2) *P* (8+1) ∧ (8+3) *P (*8+2) … (8) *P* (8+ *n*).

Esta intransitividade tem a seguinte característica peculiar: a estrutura da ordenação de preferências do agente é paralela à estrutura do paradoxo *sorites*; ou seja, cada uma das condicionais do raciocínio que conduz ao paradoxo constitui a base para uma preferência do agente entre dois itens contíguos da série. O predicado *sorítico* 'ser tarde' (neste caso 'não ser tarde') preenche todas as condições necessárias à geração do paradoxo: aplica-se inequivocamente ao primeiro membro da série, pois o agente tem a certeza de que oito horas não é tarde; é inequivocamente falso para o último membro da série, pois o agente tem a certeza de que oito e meia é tarde; e cada par contíguo da série é indistinguível no que diz respeito à aplicação do predicado (e trata-se, claro, de uma série ordenada). A questão consiste em saber se esta violação do axioma da transitividade é irracional. Paralelamente ao que acontece com o paradoxo, em que a conclusão do raciocínio é claramente falsa, deixando o predicado de se aplicar ao último membro da série, também no caso da preferência, a qual depende dessa aplicação, parece natural que se verifique uma falha na relação de transitividade; ou seja, a ordenação intransitiva de preferências parece justificar-se racionalmente.

É importante compreender-se a estrutura do paradoxo, pois as teorias formais da vagueza tentam precisamente diagnosticar a inconsistência do raciocínio empregue, tentando mostrar, nomeadamente, que as premissas são mutuamente inconsistentes. Assim, tendo em conta que as preferências do agente assentam nessas premissas, se uma das teorias da vagueza tiver sucesso em mostrar a inconsistência dessas premissas, passamos a ter uma razão para classificar esta intransitividade como irracional.

Jonathan Aldred (2004) apresentou um exemplo com a mesma estrutura daquele que é acima mencionado e tentou mostrar que não há maneira de dar as preferências do agente como irracionais. Suponha-se que um certo terreno comunitário é trabalhado rotativamente por um grupo de agricultores. À medida que se vai dando essa rotação, o terreno vai perdendo as qualidades favoráveis à produção de uma boa colheita. Uma única rotação não faz com que o terreno passe a estrar estragado [*spoilt*], caso não o estivesse antes da última utilização. Portanto, se o terreno estiver bom no início, então estará bom

após a primeira rotação; se estiver bom após a primeira rotação, então estará bom após a segunda rotação, e assim sucessivamente. Contudo, após a centésima rotação, o terreno estará definitivamente estragado. As preferências do agente, entre estados de desenvolvimento do terreno, estruturam-se de acordo com estes dados: o agente prefere um estado a outro se, e somente se, o terreno estiver bom no primeiro e estragado no segundo; caso isto não suceda, o agente é indiferente entre estados. Assim, o agente é sucessivamente indiferente entre cada dois estados contíguos, rotação após rotação, mas prefere o primeiro estado ao último, mostrando preferências intransitivas, as quais, segundo Aldred, não revelam qualquer tipo de irracionalidade.

Argumentarei no sentido de mostrar que, neste caso e noutros semelhantes, a intransitividade se deve a uma espécie de erro cognitivo da parte do agente e também a uma inadequação da linguagem utilizada. Nos dois exemplos mencionados verifica-se o seguinte pressuposto: o agente não consegue distinguir entre dois estados contíguos da sequência sorítica; contudo, mesmo que isso suceda, tal não significa que não haja realmente uma diferença e que o agente não possa agir de acordo a mesma. Por exemplo, o predicado 'estar estragado' pode ser convertido num atributo quantitativo que serve para distinguir estados consoante o número de rotações envolvidas. Assim, torna-se perfeitamente racional existir uma preferência estrita entre estados contíguos da série, x e $x+1$, mesmo que x seja apenas infimamente (e imperceptivelmente) melhor do que $x+1$.[28] Não se trata de negar a natureza tolerante do predicado 'estar estragado', mas sim conceder que, apesar dessa tolerância, é possível, e racional, preferir estritamente um estado a outro estado contíguo; as duas coisas não são contraditórias. O que estamos a questionar é o pressuposto que se encontra na base das preferências intransitivas do agente, ou seja, que um estado deve ser preferido a outro se, e somente se, o predicado 'estar estragado' se aplique a apenas a um deles.[29] De acordo com a perspectiva do erro cognitivo, é racional que um agente prefira um de entre dois estados quando ambos podem ser qualificados como bons ou não-estragados, ou quando ambos estiverem estragados, caso 'estragado' admita ainda graus de deterioração.

[28] Creio que o seguinte princípio normativo não se encontra em disputa: $xMy \rightarrow xP_{r}y$; ou seja, se x é melhor do que y, então x *deve ser* preferido a y.

[29] Do mesmo modo, apesar de não se negar a tolerância de um predicado como 'calvo', é perfeitamente razoável um agente que não é calvo preferir ter $x+1$ cabelos do que x; e o mesmo acontece com quem é calvo. Isto aplica-se, a meu ver, mesmo quando o próprio agente não é capaz de detectar a ausência de um único cabelo.

Esta perspectiva é classificada por Alder como implausível, pois a mesma implicaria, segundo ele, tratar o agente como se ele fosse 'amnésico':

'(…) o conhecimento do agente do desenvolvimento cumulativo do terreno apresenta-se como evidência contraditória da sua percepção de que "nada muda" entre fases sucessivas" (…). Em vez disso, se o pressuposto da preferência for levado a sério, deixará de se verificar tal contradição' (Aldred 2004: 383).

Ou seja, de acordo com a hipótese do erro cognitivo, um agente apenas poderia apresentar as preferências intransitivas acima mencionadas, caso se esquecesse que se deram várias rotações do terreno, o que seria implausível. Contudo, esta implausibilidade apenas nos deve preocupar, caso aceitemos tal ordenação de preferências como racional, e aquilo que estamos precisamente a defender é que um agente que seja indiferente entre rotações é, senão irracional, pelos menos descuidado com a satisfação dos seus interesses. Mas não estaremos aqui a ir contra a ortodoxia da teoria da decisão, ao questionarmos as preferências dadas do agente? Não estaremos, assim, a deixar de ser neutros quanto a isso? Não. Se aceitarmos a ideia do erro cognitivo, podemos afirmar que o agente, ao ser indiferente entre estados contíguos do terreno, distintos quanto à sua qualidade, está a prejudicar-se a si mesmo.

O que apoia ainda a perspectiva do erro cognitivo, a meu ver, é o predicado 'estar estragado' não ser observacional. A aprendizagem do uso competente de um predicado observacional é feita ostensivamente, ou seja, através da experiência directa. Esse uso implica a capacidade de distinguir, com base apenas nos órgãos dos sentidos, entre os casos em que o predicado é correctamente aplicado e os outros em que isso não acontece. Os atributos que dizem respeito à cor são um exemplo claro deste tipo de predicados. É certo que duas pessoas diferentes podem ter diferentes sensibilidades à cor, embora isso não seja importante quando o que está em causa é a formação de preferências pessoais. Portanto, quando dois objectos são sensorialmente equivalentes, isso significa que a distinção entre ambos não pode ser feita através de predicados observacionais. Avaliar se as condições de um terreno são propícias, ou não, à produção agrícola, é algo que pode depender de uma avaliação de perito, através da qual, rotação após rotação, se avalia a qualidade do solo. É plausível que essa qualidade varie de x para $x+1$, e assim sucessivamente. Portanto, se esta distinção não-sensorial entre estados contíguos do

terreno pode ser feita através de uma análise minuciosa, parece pouco racional da parte do agente ignorá-la. Por outro lado, se entre duas utilizações sucessivas do terreno não existir, de facto, alteração qualitativa da qualidade do solo, então não existe sequer qualquer razão para que o agente sinta necessidade de estabelecer uma preferência entre dois objectos que, para todos os efeitos, não são distintos quanto às propriedades relevantes.

Embora a questão da ostensibilidade sirva, a meu ver, para contestar a relevância do exemplo de Aldred, existem outros exemplos de argumentos soríticos nos quais as preferências se baseiam em distinções puramente sensoriais entre as alternativas. Quando as diferenças entre alternativas são de tal modo irrisórias que se tornam sensorialmente indistinguíveis, torna-se plausível que o agente seja indiferente entre as mesmas. Considere-se o seguinte exemplo de Patrick Maher (1993: 57-60): um agente prefere tomar o seu café com apenas uma colher de açúcar. Tal como toda a gente, supõe-se, ele não consegue distinguir entre duas chávenas de café quando estas diferem apenas num grão de açúcar; logo, ele é indiferente entre as mesmas; sendo sucessivamente indiferente entre chávenas de café que diferem apenas num grão de açúcar, dar-se-á, num certo ponto, uma violação da transitividade. Neste exemplo temos dificuldade em insistir que é possível fazer uma distinção entre a indiferença e aquilo que é uma ligeira preferência, pois tudo o que interessa ao agente é o sabor do café; como duas chávenas de café que diferem apenas num grão de açúcar têm o mesmo sabor, o agente será indiferente entre as mesmas. Maher considera que este argumento é uma falácia. O que se segue é apenas um resumo do seu raciocínio. Considerem-se as seguintes verdades:

i) Não se consegue distinguir, apenas através do paladar, um café que não contém açúcar de um café que contém 10 grãos de açúcar
ii) Não se consegue distinguir, apenas através do paladar, um café que contém 10 grãos de açúcar de um café que contém 20 grãos de açúcar.
iii) Consegue-se distinguir através do paladar um café que não contém açúcar de um café que contém 20 grãos de açúcar.

De acordo com i) e ii), segue-se, aparentemente, que, através de uma comparação directa, o café com 10 grãos tem o mesmo sabor que o café sem açúcar. Igualmente, através de uma comparação directa, o café com 10 grãos parece ter o mesmo sabor que o café com

20 grãos. Mas iii) permite-nos concluir algo diferente: se, numa comparação com o café com 20 grãos, substituirmos o café sem açúcar por um café com 10 grãos, a diferença que se sentia deixar-se-á de se sentir. Ora, isto só pode ser verdade se o café com 10 grãos tiver um sabor diferente do sabor do café sem açúcar. Ou seja, mesmo que, numa comparação directa, um agente não consiga distinguir entre chávenas de café que diferem apenas em 10 grãos, tal não significa que esses dois cafés tenham o mesmo sabor; e se não têm o mesmo sabor, cuja diferença é tudo o que importa para a formação de preferências, o agente pode plausivelmente preferir aquele que tem apenas mais 10 grãos de açucar. Esta conclusão tem o aspecto de um oxímoro, mas generalizando o exemplo talvez consigamos tornar claro aquilo que está em causa. Utilize-se a notação '$a \gg b$' para denotar que *a* é detectavelmente melhor do que *b*; e '$a = b$' para denotar que *a* e *b* não são distinguíveis apenas com base no sabor. Temos, assim, três condições suficientes para podermos afirmar que *a* sabe melhor do que *b*:

. $a \gg b$;
. $a = b$, mas existe um *c*, tal que $a = c$ e $c \gg b$; é isto que se passa no exemplo dos grãos de açúcar: *a* corresponde ao café com 10 grãos, *b* ao café sem açúcar e *c* ao café com 20 grãos;
. $a = b$, mas existe um *d*, tal que $a \gg d$ e $d = b$; isto pode acontecer, plausivelmente, caso, por exemplo, *a* seja um café com 20 grãos, *b* um café com 10 grãos e *d* um café com 5 grãos (desde, claro, que uma diferença de 15 grãos seja detectável pelo paladar).

Estando garantida pelo menos uma destas três condições, um qualquer agente passa a ter uma razão para preferir *a* a *b*, dado o nosso pressuposto de que uma diferença de sabor, ainda que indetectável numa comparação directa, justifica uma preferência. A aparência de oxímoro desvanece-se, caso identifiquemos o café com 20 grãos como uma espécie de apurador do palato, o qual permite ao agente concluir que, afinal, uma diferença de 10 grãos pode ter influência no sabor do café. Acredito que esta estratégia de Maher pode ser adaptada para justificar qualquer preferência baseada numa ínfima, mas existente, diferença sensorial, seja qual for o órgão em causa.
Finalmente, quanto ao contributo das teorias da vagueza para a defesa da transitividade, a hipótese do erro cognitivo é compatível com a solução do paradoxo oferecida pela teoria

epistémica da vagueza.[30] Segundo esta teoria, existe nos casos de vagueza uma fronteira bem definida que determina a aplicabilidade do predicado em causa. A vagueza não passa, assim, de ignorância da nossa parte quanto à existência, ou não, de uma propriedade real das coisas. O paradoxo passa a ser visto como uma redução ao absurdo que demonstra a falsidade de pelo menos uma premissa. Neste caso, a falsidade da seguinte premissa:

Se um número *n* de rotações não estraga o terreno, então um número *n* +1 de rotações também não estraga o terreno.

Ou seja, existe um ponto exacto em que, após uma rotação, o terreno passa a estar estragado. Como estamos a interpretar uma teoria da vagueza de acordo com a maneira como esta serve de guia para a formação de preferências do agente, a teoria epistémica oferece ao agente uma justificação para ter uma preferência estrita no ponto exacto em que o predicado deixa de se aplicar, quebrando desse modo a cadeia de relações de indiferença.

Se neste exemplo podemos considerar que existe uma certa intuição de implausibilidade, ao ter de se negar, aparentemente, a natureza gradual do processo de transformação do terreno, existem outros casos em que a solução da teoria epistémica é bastante plausível. Isto sucede, por exemplo, quando se trata do predicado 'estar atrasado'; parece existir um minuto exacto em que o agente passa a estar atrasado, se por atrasado se estiver a falar de chegar um minuto depois da hora agendada para o seu compromisso, apesar de ser impossível determinar com exactidão que minuto é esse. Ou considere-se ainda este caso: se um iogurte não está estragado num dia *n*, então não está estragado num dia *n* +1. Contudo, se a fronteira entre não-estar e estar estragado for estabelecida através da má disposição do agente, parece existir uma hora exacta a partir da qual o agente ficará mal-disposto, caso coma o iogurte.

Aldred (2004: 389-90) argumenta contra a possibilidade de se estabelecerem preferências com base na solução para o paradoxo oferecida pela teoria epistémica. Ele faz notar que a ignorância do agente não pode ser caracterizada como um estado em que a informação disponível é imperfeita ou incompleta; de acordo com a teoria, o limite é, por definição, impossível de ser conhecido, ou seja, o agente não dispõe de qualquer informação. Segue-se daí que ele não poderá sequer formar uma distribuição de probabilidades relativamente

[30] Para uma defesa da teoria epistémica, ver Williamson (1994).

às alternativas, o que o deixa numa posição de paralisia quanto à decisão a tomar. A transitividade é salva às custas da capacidade de tomar uma decisão. Aldred dramatiza a situação, afirmando que é tão provável que o limite se estabeleça nas vinte rotações quanto nas quarenta rotações.

A ideia de que é impossível conhecer o limite significa que é impossível seleccioná-lo de uma forma cognitivamente informada, i.e., de uma forma que não seja através do mero acaso. Se for apresentada ao agente uma paleta de cores entre o vermelho e o laranja, ele não saberá apontar qual delas corresponde ao último vermelho e qual delas corresponde ao primeiro laranja, mas ele sabe quais delas são de certeza vermelhas e quais são de certeza laranja. Contudo, a meu ver, o agente não sabe tão pouco quanto isso. Ninguém contesta a necessidade de tomar decisões com base em predicados vagos, nomeadamente em questões éticas. Um governo que esteja a tentar definir um escalão para aplicar uma taxa especial de IRS aos mais ricos não irá certamente colocar o limite numa zona em que se levantem questões de justiça relativamente aos casos de fronteira.

O meu ponto é o seguinte: quando, nos casos de vagueza, estamos a lidar com o estabelecimento de preferências, ou com questões práticas em geral, sendo impossível estabelecer o limite de uma forma cognitivamente informada, é racional adoptar um método prático de solução para a vagueza inspirado na teoria supravaloracionista [*supervaluationism*] (ver Fine 1975), ainda que aceitemos como verdadeira a análise da teoria epistémica. Tal método consiste em estabelecer uma zona de vagueza ao longo da qual é aceitável qualquer selecção do limite que justifica a preferência estrita (cada uma dessas selecções é chamada 'um afinamento' [*sharpening*] do predicado. É isto que os defensores da moralidade do aborto tentam estabelecer. A localização dos extremos dessa zona de vagueza não será arbitrária, desde que respeite as percepções/conhecimento que determinam, com elevado grau de certeza, se o predicado se aplica ou não. Isto não implica que aceitemos o princípio da teoria supervaloracionista, de acordo com o qual uma frase que contenha um predicado vago é verdadeira se, e somente se, for verdadeira de acordo com todos os afinamentos possíveis. Se estivéssemos a tratar de questões semânticas, e supondo a correcção da teoria epistémica, o estabelecimento de um limite através de um decreto, como nas questões de direito, seria completamente arbitrário. Mas como lidamos com questões práticas, nas quais está em causa a tomada de decisões, é

racional estabelecer as nossas preferências de acordo com aproximações suficientemente precisas, as quais nos garantam que não sairemos prejudicados.[31]

3.3. Independência

O axioma da independência diz-nos o seguinte: caso duas acções tenham a mesma consequência num determinado estado do mundo, então as preferências do agente, relativamente a essas acções, devem ser independentes do que acontece nesse estado do mundo. Este axioma pode ser formulado de diferentes maneiras. Aqui irei adoptar uma formulação que se encontra directamente relacionada com o argumento anti--independência que iremos utilizar.[32] Seja X*p*Z uma lotaria em que o agente ganha X com probabilidade *p* e Z com probabilidade 1 – *p*; o mesmo se aplicando para Y*p*Z. Segundo o axioma,

$XpZ \succ YpZ$ se, e somente se, $X \succ Y$ (desde que $p > 0$).

Neste caso também é possível construir-se um argumento *money-pump*, mostrando que um agente cujas preferências não satisfaçam a independência, pode vir a ser explorado, i.e., transformado numa *money-pump*. Para esse efeito, considere-se o seguinte conjunto de alternativas:

X – receber 1,000,000 Euros
Y – aceitar a seguinte aposta: 5,000,000 Euros se, e somente se, chover amanhã (C), com probabilidade 10/11, e 0 Euros se não chover (¬C).
Z – receber 0 Euros
X + t – receber 1,000,000 Euros mais um cêntimo
Z + t – receber 0 Euros mais um cêntimo

[31] Quem permaneça céptico relativamente à racionalidade intrínseca do axioma da transitividade pode desfrutar, por exemplo, da bateria de exemplos engenhosos apresentadas mais recentemente por Larry Temkin (2012).
[32] Veja-se, por exemplo, a formulação de Savage (1954: 21), em que este identifica a ideia que está na base do axioma como 'o princípio da coisa certa' [*sure-thing principle*].

Considerem-se agora as seguintes preferências de um determinado agente perante pares de alternativas, entre as acima indicadas:

1) Aceitar X ou aceitar Y? É razoável esperar que um certo agente avesso ao risco (e a maioria das pessoas são avessas ao risco) prefira 1 milhão certos a uma aposta em que pode ganhar 5 milhões com probabilidade 10/11 ou nada com uma probabilidade 1/11.
2) Aceitar X*p*Z ou aceitar Y*p*Z? Duas lotarias: uma em que se ganha X com uma probabilidade 0.11 ou Z com uma probabilidade 0.89; e outra em que se obtém a oportunidade de ganhar Y com uma probabilidade 0.11 ou Z com uma probabilidade 0.89. Como só se ganha Y se chover amanhã, e sendo esta probabilidade de 10/11, então a probabilidade de Y*p*Z é de 0.11×10/11 = 0.10. É razoável, então, esperar que um certo agente prefira uma aposta em que pode ganhar 5 Milhões com uma probabilidade 0.10 ou nada com uma probabilidade 0.90, a uma aposta em que pode ganhar 1 Milhão com probabilidade 0.11 ou nada com uma probabilidade 0.89.

Daqui se conclui que o agente evidencia a seguinte ordenação de preferências:

$X \succ Y$ e $Yp Z \succ Xp Z$,

o que constitui uma violação do axioma da independência. Mais, é de supor que um cêntimo não modifique significativamente as preferências do agente, daí que possamos concluir com segurança que a seguinte ordenação também se verifica:

$Yp Z \succ Xp Z + t \succ Xp Z$.

Para se conceber um argumento pragmático que demonstre que um agente que exiba esta ordenação de preferências pode ser colocado numa situação em que, faça o que fizer, fica sempre a perder, basta pedir-lhe que resolva um certo problema de decisão sequencial e que, ao fazê-lo, tente observar o plano genérico que se espera de um agente racional, ou seja, que aja de modo a satisfazer as suas preferências. Verificar-se-á, neste caso, aquilo a que os estudiosos da teoria da decisão chamam de ‘inconsistência dinâmica’: o agente

embarcará num plano de acção que não será levado até ao fim, mesmo que as suas preferências não se alterem à medida que o plano escolhido vai sendo implementado. Para esse efeito, considere-se a seguinte árvore de decisão:[33]

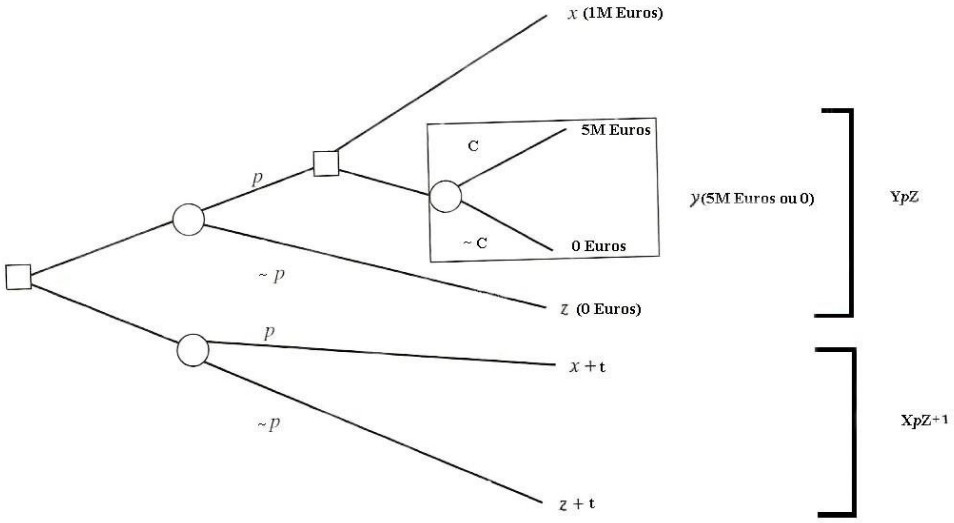

Fig. 1

Nesta árvore de decisão, os quadrados são *nódulos de escolha* e os círculos *nódulos de acaso*.[34] O agente em causa, procurando obter Y*p*Z, escolherá o ramo superior no primeiro nódulo. Aí, se *p* ocorrer (0.11), o agente chega ao segundo *nódulo de decisão*, em que tem de escolher entre ficar com o milhão que já tem ou tentar ganhar cinco milhões se C ocorrer (se chover amanhã, com probabilidade 10/11). O plano do agente deveria, neste ponto, fazê-lo escolher a segunda hipótese, mas como ele prefere X a Y, ele nunca escolherá a segunda hipótese, acabando por nunca obter Y*p*Z.

Contudo, se no primeiro *nódulo de decisão* ele tivesse escolhido o ramo inferior, então, ocorresse ou não *p* (0.11), ele ficaria sempre melhor do que se tivesse escolhido o ramo superior. Ou seja, obteria o mesmo, um milhão ou zero, mais 1 cêntimo para cada uma destas hipóteses, um pequeno *bónus*. Nessa medida, a implementação do plano inicial

[33] Esta pode ser encontrada em Rabinowicz (1995: 589), sem algumas pequenas alterações que introduzi para tornar mais claras as opções ou planos do agente.

[34] Um *plano* consiste numa especificação, prévia à própria acção, do modo como o agente irá agir em cada um dos nódulos de decisão de um dado problema sequencial. Nesta medida, o plano prepara o agente para todas as contingências possíveis (fruto do acaso) que poderá vir a enfrentar, menos aquelas que decorrem de um desvio relativamente ao próprio plano.

condu-lo sempre a um resultado *sub-óptimo*, o que vai contra um dos princípios de racionalidade mais fundamentais, o princípio da escolha dominante.

Este argumento parece depender de alguns pressupostos relacionados com as nossas intuições acerca da aversão ao risco. Para o argumento correr é necessário que aceitemos como plausíveis algumas preferências do agente, nomeadamente, a preferência por um Milhão garantido a uma chance elevada de ganhar 5 Milhões. É, por esse motivo, importante ter em conta que uma certa noção de aversão ao risco é compatível com o princípio da maximização da utilidade subjectiva esperada. O tipo de aversão ao risco a que me refiro foi caracterizado por Kenneth Arrow e é suficiente para o nosso propósito.[35] Segundo a definição de Arrow, alguém tem aversão ao risco quando, partindo de uma posição de certeza ou garantia, não está disposto a aceitar uma aposta justa [*fair*]:

3 maçãs garantidas $\succ$ uma lotaria em que ganha 6 maçãs com *p* 0.5 ou 0 maçãs.

Se a nossa utilidade por maçãs for marginalmente decrescente, então a aversão ao risco é compatível com o princípio da maximização da utilidade subjectiva. Isto acontece sempre que a função de utilidade do agente, relativamente a maçãs, é côncava; ou seja, quando a sua curva se vai tornando menos inclinada à medida que cresce. Assim, num determinado intervalo, quanto mais lentamente crescer a função de utilidade, mais avesso ao risco este será.

Neste ponto coloca-se a questão que põe em causa não apenas a defesa do axioma da independência, mas também a dos axiomas da completude e da transitividade. Parte importante da aceitação destes axiomas depende do quão persuasivo se considera ser o argumento pragmático *money-pump*. Um agente que incorra em inconsistência dinâmica não mostrará, afinal, uma certa miopia [*short-sightedness*]? Não será ele capaz de antever que a sua ordenação de preferências se presta a ser explorada? Afinal, uma certa capacidade de antever o futuro [*foresight*] parece ser suficiente para se evitar cair numa situação em que as preferências do agente o compelem a efectuar uma troca ou a tomar uma decisão que o deixará prejudicado. Um agente que possui esta capacidade é designado na literatura da teoria da decisão como um agente *sofisticado* (Rabinowicz 1995). As características fundamentais de uma escolha sofisticada são as seguintes: em

[35] Para uma breve discussão de várias noções de aversão ao risco, ver, por exemplo, Peterson (2009: 179-183).

primeiro lugar o agente tem de estar plenamente informado acerca da natureza do problema de decisão sequencial, ou seja, ele tem de ser capaz de visualizar a totalidade da árvore de decisão que caracteriza o problema; em segundo lugar, o agente tem de ter uma dose considerável de segurança na sua racionalidade futura, pois a racionalidade das acções que realizará *antes* dependerá da racionalidade daquelas que realizará *depois*; finalmente, estando estas duas condições satisfeitas, o agente poderá raciocinar e solucionar o seu problema de decisão através do método designado por *indução retroactiva* [*backwards induction*].

Este método consiste em resolver um problema de decisão 'de trás para a frente', ou seja, começando cronologicamente pelo fim, ou pela última decisão que se terá de tomar, e acabando na decisão inicial. Primeiro, à luz das suas preferências, o agente identifica qual é a decisão que é racional tomar no último nódulo de escolha da árvore de decisão; depois, com esta decisão em mente, determina qual é a decisão racional no penúltimo nódulo; e assim sucessivamente até chegar à decisão inicial. É como se o agente enfrentasse em cada nódulo uma árvore de decisão diferente que começa nesse mesmo nódulo e se estende por aí em diante; dessa maneira, o agente vai acrescentando, 'de trás para a frente', um novo braço à sua árvore de decisão, até chegar à decisão inicial. Este procedimento é tornado claro, e facilmente ilustrado, pelo caso da ordenação cíclica de preferências:

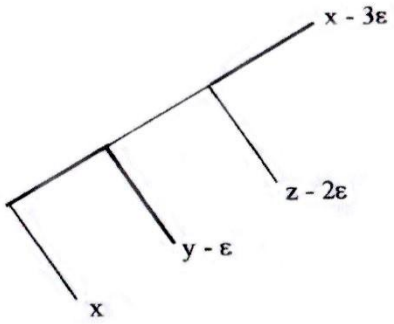

Fig. 2

Cada bifurcação nesta árvore corresponde a um nódulo de escolha, em que o agente pode escolher, virando para baixo, ficar com o item que obteve através da troca anterior ou, continuando para cima, efectuar uma nova troca; ε é a pequena quantia que o agente está disposto a perder para trocar um item por outro; e os traços escurecidos correspondem

aos movimentos autorizados pela indução retroactiva.[36] O ciclo é gerado pela seguinte ordenação de preferências: $z \succ y \succ x \succ z$.

No último nódulo o agente aceita a troca, pois prefere estritamente $x - 3\varepsilon$ a $z - 2\varepsilon$; no segundo nódulo, sabendo que irá aceitar a troca no nódulo seguinte, o agente recusar-se-á a trocar, optando por ficar com $y - \varepsilon$, resultado que ele prefere estritamente a $x - 3\varepsilon$. Mas como o agente prefere estritamente $y - \varepsilon$ a x, ele aceitará trocar no primeiro nódulo. Portanto, o agente sofisticado optará por fazer apenas uma troca, bloqueando, dessa maneira, a possibilidade de ser explorado.

Na árvore da fig. 1, o agente sofisticado, ao prever que a tentativa de obter Y*p*Z o vai conduzir a X, opta por virar para baixo no primeiro nódulo de escolha, pois este garante-lhe um resultado, X*p*Z + t, que é estritamente melhor do que X, evitando desse modo a violação do princípio da escolha dominante. O agente sofisticado escolhe, entre os planos executáveis, aquele que tem um melhor resultado esperado. Um plano executável caracteriza-se da seguinte maneira: o agente sabe que, ao embarcar nele, irá conseguir levá-lo até ao fim, não encontrando no seu caminho qualquer razão para se desviar; ou, de outra maneira, um plano executável é tal que todos os seus movimentos são sancionados pelo método da indução retroactiva. Nesta medida, o plano que consiste em tentar obter Y*p*Z não é executável, e o agente sofisticado dirá, simplesmente, que um plano que não é executável não é realmente uma opção.

Existem vários argumentos que tentam mostrar que a indução retroactiva não é um método fiável de raciocínio e que, como tal, a escolha sofisticada não é realmente uma opção para o agente maximizador da utilidade.[37] Contudo, o argumento que me parece mais forte contra a possibilidade da escolha sofisticada é um argumento que mostra que apesar da nossa capacidade de antever as nossas escolhas futuras, um explorador arguto

[36] Esta árvore pode ser encontrada em Rabinowicz (2000: 137), não tendo sofrido aqui qualquer alteração.

[37] A principal objecção pode, a meu ver, resumir-se da seguinte maneira (Rabinowicz 2000: 139): como pode o agente pressupor que irá agir racionalmente quando chegar ao último nódulo de escolha se, para chegar até aí, teve presumivelmente de fazer várias escolhas que não foram sancionadas pela indução retroactiva? De outra maneira: como pode o agente ter uma confiança robusta na sua racionalidade futura se, de acordo com a experiência acumulada, adquiriu evidência da sua irracionalidade? Este é um problema que se coloca principalmente em árvores grandes, com mais nódulos de escolha do que aquelas que temos vindo a considerar. A meu ver, este não é um argumento forte. O agente, ao determinar o que fará no último nódulo, não teve realmente que passar por quaisquer outros nódulos de escolha, nem tem sequer que pressupor que o fará; ou seja, basta saber o que faríamos se nos víssemos confrontados com essa escolha, para determinar que não queremos estar numa circunstância em que essa escolha tem de ser feita. Portanto, a acusação de inconsistência (ver Pettit e Sugden (1989), para o que estes chamam de *paradox of backwards induction*) não me parece particularmente forte.

consegue ainda assim divisar uma maneira de nos explorar. Considere-se, para o efeito, a seguinte árvore de decisão:[38]

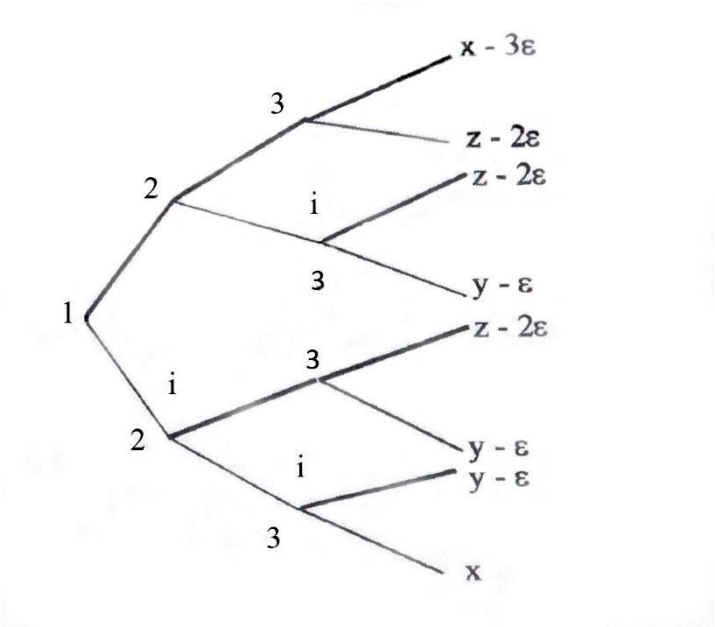

Fig. 3

Supõe-se, mais uma vez, que o agente possui a mesma ordenação cíclica de preferências. As linhas escurecidas representam os passos sancionados pela indução retroactiva. Os números correspondem aos nódulos de escolha, em sequência, de todos os planos de acção possíveis. E os *i*'s correspondem a momentos em que o *explorador* insiste em propor uma troca. A característica distintiva deste novo problema de decisão consiste, portanto, na teimosia do *explorador*, o qual, após a decisão do agente em não efectuar uma troca, volta a insistir com outra troca possível.

Aplicando o método da indução retroactiva, constata-se que o agente irá aceitar trocar em cada um dos últimos nódulos de escolha (correspondentes aos vários 3) de cada plano, pois essa troca garante-lhe um resultado esperado estritamente melhor do que aquele que resultaria da recusa em trocar. Se o agente escolher ir para cima no nódulo 1, então escolherá também ir para cima no nódulo 2, pois ele sabe que, caso siga para baixo em 2, acabará com $z - 2\mathcal{E}$, resultado que é estritamente pior do que $x - 3\mathcal{E}$. Se escolher ir para baixo no nódulo 1, então escolherá ir para cima no nódulo 2, pois ele sabe que caso siga para baixo em 2, acabará com $y - \mathcal{E}$, resultado que é estritamente pior do que $z - 2\mathcal{E}$.

[38] Esta árvore, sem algumas alterações que acrescentei para a tornar mais clara, encontra-se em Rabinowicz (2000: 141).

Contudo, $z - 2\mathcal{E}$ é estritamente pior do que $x - 3\mathcal{E}$, logo ele nunca escolherá ir para baixo no nódulo 1. Em suma, a indução retroactiva oferece ao agente, como único plano executável, o caminho que consiste em seguir para cima nos três nódulos de escolha e acabar com $x - 3\mathcal{E}$. O problema deste resultado torna-se, assim, claro: o agente prefere naturalmente ficar com *x* a acabar na mesma com *x* menos qualquer valor que seja, por mais pequeno que este possa ser. Mas é isto mesmo que acontece se o agente aceitar as três trocas, em vez de recusar trocar três vezes e ficar com *x*. O agente sofisticado com uma ordenação cíclica de preferências encontra-se sujeito a ser transformado em *money--pump* por um *explorador* insistente, tal como acontece com o agente *míope*, o qual não possui a capacidade de antever o futuro.[39]

Apesar de este argumento mostrar que a indução retroactiva não impede o agente sofisticado de incorrer numa perda certa, tal acontece nos casos em que está em causa a violação do axioma da transitividade. Na fig. 1, o agente sofisticado optará sempre por ir para baixo no primeiro nódulo de escolha, acabando por ficar com X + *t* (ou Z + *t*) em vez de X (ou Z), evitando essa perda. Contudo, para o agente sofisticado, o plano que consiste em procurar obter a sua opção preferida, Y*p*Z, continua a não ser executável, o que nos permite afirmar, em certa medida, que as preferências que violam a independência prejudicam o agente. Isto acontece porque uma das consequências da aplicação da indução retroactiva consiste em reduzir os planos executáveis a apenas um; seja qual for a árvore de decisão, o caminho que o agente faz para alcançar o seu objectivo tem de ser todo ele composto por linhas escurecidas, como se pode facilmente constatar nas figs. 2 e 3. A questão que se coloca neste ponto é a de saber se existe algum outro método de escolha que permita ao agente ter à sua disposição um maior número de planos executáveis e, como tal, optar pelo plano *óptimo*, i.e., aquele que tem um melhor resultado esperado. Um tal método seria certamente superior à *escolha sofisticada*.

McClennen (1990) apresentou um novo método para fazer face ao problema da tomada de decisão dinâmica e designou-o por *escolha resoluta*. Tal como o nome indica, um agente *resoluto* é aquele que escolhe um determinado plano e permanece decidido a leva-

[39] Como já se referiu, parece-me que uma outra maneira de tentar explorar um agente sofisticado, embora não tão eficaz, poderá consistir em espaçar no tempo as ofertas de troca, de tal modo que apenas um agente cuja memória lhe permita lembrar-se das trocas passadas consiga perceber que está a ser vítima de uma tentativa de exploração. Por exemplo, no problema da fig. 2, após a oferta da primeira troca, o explorador deixa passar muito tempo até oferecer a troca seguinte. Se o agente não se lembrar que começou com *x*, então a indução retroactiva não o impedirá de, a partir de $y - \mathcal{E}$, tentar obter a alternativa estritamente melhor, $x - 3\mathcal{E}$.

-lo até ao fim, independentemente das tentações que encontra no seu caminho para dele se desviar. A resolução em manter-se fiel ao plano inicial motiva-o de maneira a que, face a cada novo par de escolhas, as suas preferências se alterem de modo a seguir tal plano. Verifica-se, assim, uma distinção quanto às preferências do agente. Por um lado, este possui preferências *estáticas*, aquelas que ditam a sua escolha num determinado nódulo, supondo que esse nódulo é o começo de uma árvore de decisão independente; por exemplo, no segundo nódulo de decisão da fig. 1, as suas preferências *estáticas* ditam que ele escolherá o milhão em vez de optar por Y*p*Z.[40] O que caracteriza estas preferências é o facto de elas ignorarem o caminho feito para se chegar a um determinado nódulo escolha. Por outro lado, o agente possui o que poderíamos designar por preferências *dinâmicas*, aquelas que dependem da escolha de um plano inicial e que, por esse motivo, são capazes de reverter as *estáticas* quando necessário, fazendo com que, por exemplo, o agente opte por Y*p*Z em vez do milhão certo no segundo nódulo de escolha da fig. 1. O que caracteriza as preferências *dinâmicas* é a atenção à história do percurso realizado, incluindo também, poder-se-á dizer, a consideração de hipóteses contrafactuais: quando chegado ao segundo nódulo da fig. 1, o agente considera que, caso tivesse optado por X*p*Z + 1 no primeiro nódulo, então estaria melhor do que optando por X, e isto fá-lo-á reverter a sua preferência por X, levando-o a optar por Y*p*Z e a levar o plano até ao fim.

O método da escolha resoluta abre novas possibilidades ao agente: planos que eram apenas teoricamente possíveis passam a ser todos eles executáveis, pois a resolução do agente evita quaisquer distracções, eliminando qualquer perigo de inconsistência dinâmica. O agente resoluto encara, portanto, o seu problema de decisão sequencial como se este se tratasse de um problema em que apenas uma escolha tem de ser feita, neste caso uma escolha entre vários planos disponíveis. De seguida, o plano é implementado de uma forma automática ou, se quisermos, de uma forma robótica, de olhos no chão. Será fácil de constatar que este método contorna os problemas que a *escolha sofisticada* enfrenta, permitindo ao agente alcançar uma satisfação *óptima* das suas preferências, ou seja, obter Y*p*Z.

Falta, contudo, na análise da robustez, de McClennen, uma explicação do modo como esta funciona aos níveis cognitivo e psicológico. Sem essa explicação, acerca de como podem os indivíduos agir de forma robusta, a pretensão dos seus proponentes

[40] Estas preferências correspondem àquelas que foram apresentadas antes de ser revelado o problema de decisão em causa.

permanecerá vazia. Por outr lado, o tipo de raciocínio que se encontra na base da *escolha sofisticada* é perfeitamente claro e acessível. A questão consiste, portanto, em saber de que maneira o nosso eu futuro se manterá fiel a resoluções tomadas no passado, pois só é racional enveredar por um determinado plano, caso tenhamos um grau elevado de confiança em como conseguiremos levá-lo até ao fim. A meu ver, a análise mais plausível consiste em classificar a *resolução* como uma forma de adesão a princípios normativos do tipo: 'Nunca comer sobremesa' ou 'Não fumar mais do que um cigarro após as refeições'. Robert Nozick (1993) apresentou com algum detalhe as várias funções destes princípios normativos: intelectual, interpessoal e intrapessoal. Para nós é suficiente considerar a última.

A principal função intrapessoal dos princípios, para além de nos conferirem identidade e nos permitirem ser a pessoa que desejamos ser – por exemplo, alguém com determinados princípios - consiste em fazer-nos ignorar as tentações ou obstáculos que encontramos quando queremos alcançar um determinado bem futuro. A tendência para nos deixarmos facilmente tentar por bens mais imediatos deve-se a uma característica nossa bem conhecida, que os dados recolhidos no decurso de investigações em áreas como a economia e a psicologia tornaram evidente: atribuímos agora uma menor utilidade a uma recompensa futura, do que quando chega o momento de a reclamar; e quanto mais distante estiver essa recompensa, menor utilidade lhe atribuímos agora. Uma outra maneira de colocar a questão consiste em afirmar que 'descontamos o futuro'. Este mecanismo ou tendência pode muito bem ter sido evolutivamente selecionado, consistindo numa maneira de nos precavermos contra a natural incerteza relacionada com a obtenção de um bem futuro que circunstâncias inesperadas podem vir a impossibilitar. O problema surge quando este mecanismo se torna redundante, na medida em que possuímos algumas ferramentas para nos precavermos de uma forma racional, e já não instintiva, contra essa incerteza, ferramentas essas que surgem na forma de dados estatísticos, das modernas teorias da decisão ou na capacidade cada vez maior que possuímos de prever o futuro. Afinal não nos encontramos em perigo constante de sermos atacados por leões ou de sucumbirmos à fome.

Uma maneira de melhor compreender o funcionamento deste mecanismo consiste em visualizar num gráfico uma curva que representa o 'desconto do futuro', ou o aumento da utilidade da recompensa (medida no eixo *y*) por período de tempo (medido no eixo *x*), e constatar que essa curva cresce à medida que o tempo passa, descrevendo uma hipérbole.

Podemos, assim, ter duas curvas, representado a utilidade de dois bens futuros, sendo um deles mais imediato do que o outro; por um curto período de tempo, o bem mais imediato apresentará uma maior utilidade, i.e., a sua curva será mais pronunciada de início, até que a curva do bem menos imediato acaba por ultrapassá-la. Coloca-se, portanto, a seguinte questão: 'Por que razão devemos evitar a tentação de ficar com o bem com menor utilidade, embora mais imediato, e esperarmos até alcançar o bem com maior utilidade, embora mais longínquo'? Nozick (1993: 16) apresenta dois motivos a favor da racionalidade deste procedimento, os quais me parecem bastante válidos. Em primeiro lugar, o curto período de tempo em que o bem mais imediato possui maior utilidade não é representativo das preferências do agente, ou seja, o bem com maior utilidade é preferido pelo agente por um período de tempo bastante maior do aquele em que o bem mais imediato é preferido; se quisermos, a preferência mais estável do agente é aquela em que o bem mais longínquo tem maior utilidade. Em segundo lugar, quando o agente decide ceder à tentação, e ficar com o bem mais imediato, é inevitável que ele venha a sentir-se arrependido quando o curto intervalo de tempo em que esse bem tem maior utilidade se esgota. É o que acontece após termos comido um gelado à sobremesa ou após termos fumado mais um cigarro do que o estabelecido previamente.

Existem várias maneiras bem conhecidas de ultrapassar um período intermédio de tentação. Por exemplo, encontrando uma maneira de condicionar a nossa liberdade de escolha antes de chegar o momento crítico: o caso de Ulisses atado ao mastro do navio, antes de começar o canto das sereias, ou o de Bruce Banner fechado numa cela de betão, antes de se transformar no Hulk, são ilustrações claras desse método. Outra maneira consiste em investir tempo e custos na preparação de um determinado plano, de modo a que mais tarde nos venhamos a sentir obrigados a honrá-lo; por exemplo, pagar um semestre adiantado no ginásio. Mas a estratégia que nos interessa, aquela que se adequa às situações de escolha dinâmica, é o estabelecimento de princípios normativos, de carácter geral, para a regulação do comportamento.

Estes princípios funcionam através de um mecanismo simbólico de agregação de acções do mesmo tipo. Através do princípio, a decisão de fazer, ou não, uma determinada acção em particular – comer *esta* sobremesa, fumar *este* cigarro – encontra-se conectada com uma classe de acções do mesmo tipo, passando cada uma delas a *representar* a totalidade da classe. Assim, fazer qualquer uma das acções particulares consiste em fazer, simbolicamente, *todas* as acções que fazem parte dessa mesma classe. A utilidade, ou

desutilidade, associada a cada uma das acções representa a utilidade, ou desutilidade, da soma de todas as acções juntas; e, claro, a utilidade associada ao prazer resultante de se ceder à tentação nunca será suficientemente grande para compensar toda essa desutilidade. Pela mesma ordem de ideias, um princípio não admite qualquer excepção, pois falhar em satisfazê-lo numa ocasião em particular *representa* falhar em satisfazê-lo sempre. A lei psicológica do reforço positivo ajuda a manter-nos fiéis aos nossos princípios e explica de que modo estes funcionam como condutores de probabilidade: cada vez que satisfazemos um princípio, maior confiança temos em como no futuro iremos continuar a fazê-lo; e quanto mais tivermos no passado investido na satisfação de um princípio, maior será o custo de uma violação; por outro lado, sempre que existe uma violação, as nossas estimativas quanto à probabilidade de futuras violações irão igualmente aumentar. Tendo em conta esta análise, compreender-se-á quando Nozick (1993: 26-35) fala do valor simbólico das acções, casos em que a utilidade da acção é associada não apenas às consequências que esta produz, mas também àquilo que ela *representa*. Isto significa que as acções *expressam* coisas como crenças, valores ou emoções, a utilidade das quais não se encontra ligada, pelo menos de modo imediato, às consequências das acções. Só assim se explica que as pessoas façam coisas como reciclar o lixo ou votar em grandes eleições. É claro que de um ponto de vista estritamente consequencialista, o valor simbólico de uma acção, na medida em que influencia a opção por uma ou outra acção, tem as suas próprias consequências do tipo causal, podendo contribuir para que uma acção seja considerada causalmente *óptima*. A necessidade de incorporar, ou não, a utilidade simbólica das acções no modo de funcionamento das teorias da decisão fica, assim, por esclarecer.

Aquilo que nos importa, para averiguar acerca da possibilidade da escolha resoluta, é compreender que a adesão a um plano de acção em particular, que consideramos racional, pode ser uma instância de um princípio geral que nos diz o seguinte: 'Devemos sempre cumprir, ou levar até ao fim, um plano de acção que nos permite obter um resultado *óptimo*'. Acresce a isto a possibilidade de a adesão a planos poder ter igualmente uma utilidade simbólica, a qual pode ter uma influência não apenas intrapessoal, mas também interpessoal: por exemplo, desejamos ser vistos como pessoas fiáveis que cumprem os planos pelos quais decidem enveredar.

Finalmente, para que possamos aplicar o método da *escolha resoluta* ao problema de decisão da fig. 1, é necessário averiguar se o problema da adesão a planos de acção, em

problemas de decisão consequencial, é estruturalmente idêntico ao problema que consiste em evitar ceder a tentações. Para que isto seja o caso, certas características têm de se verificar no problema da fig.1, nomeadamente, a existência de um bem futuro com uma maior utilidade do que um bem imediato, e a justificação do agente em sentir-se arrependido caso se desvie do plano original.

Um argumento que me parece demonstrar essa identidade estrutural consiste no seguinte: afirmar que o agente, quando chega ao segundo nódulo de decisão da fig. 1, é confrontado com uma escolha entre um Milhão, que corresponde ao bem imediato com menor utilidade, e Y*p*Z, o bem mais longínquo com maior utilidade. Poder-se-á, contudo, contestar esta identificação da seguinte maneira: um Milhão tem maior utilidade do que Y*p*Z sempre, e não apenas no segundo nódulo, sendo este um dado que nos é oferecido na formulação do problema, sem o qual não existiria sequer um problema de decisão para resolver. Além disso, quando chegamos ao segundo nódulo, o problema de decisão chega ao seu término, não existindo mais decisões a ser tomadas no futuro, para além daquela que se nos apresenta. Logo, o problema de decisão da fig.1 não tem a mesma estrutura do problema que consiste em evitar tentações e alcançar o bem futuro.

A meu ver, estas razões podem ser contestadas com eficácia. Em primeiro lugar, quando se opta por um plano de acção, o resultado óptimo, aquele que tem maior utilidade, é a recompensa final do plano. Quando os defensores da escolha resoluta afirmam que numa situação idêntica à do segundo nódulo o agente modifica a suas preferências estáticas, encontra-se implícita a ideia de que o agente é tentado por um par de preferências que não se encontrava contemplando aquando da opção por um dos planos; ou seja, o agente tem de ignorar as suas preferências extra-(problema de decisão), quando se lhe apresenta a oportunidade de as satisfazer. Portanto, quando se afirma que a escolha resoluta implica uma consideração da história passada, ou dos passos anteriores da respectiva árvore de decisão, tal significa que é a resolução de se manter fiel ao plano que o faz ignorar as preferências no segundo nódulo. Ou seja, a recompensa final do plano escolhido é o bem ao qual o agente atribui uma maior utilidade, pois, aquando da avaliação dos vários planos à sua disposição, o que estava em causa era a sua preferência relativamente ao par de opções {X*p*Z, Y*p*Z}, e não a sua preferência relativamente ao par {X, Y*p*Z}. À pergunta acerca de como é a resolução possível, os defensores da escolha resoluta afirmarão que estão apenas a satisfazer o princípio geral que nos diz para levarmos os nossos planos até à sua conclusão. Não existe aqui qualquer perigo de confundir o plano com o princípio

que motiva o agente a concluí-lo: o primeiro consiste em tomar os devidos passos para obter Y*p*Z e o segundo é a base da resolução de cumprir o primeiro.

Quanto à segunda parte da objecção, poder-se-á responder que a árvore de decisão só chega ao fim após todos os nódulos estarem ultrapassados, sejam estes nódulos de decisão ou nódulos de acaso. Como após o segundo nódulo de decisão, no ramo superior da árvore, ainda se encontra um nódulo de acaso, é perfeitamente razoável aceitar a ideia de que a recompensa final apenas chegará no futuro, após o segundo nódulo de decisão. Este aspecto vem confirmar a ideia de que é racional esperar pela obtenção Y*p*Z, segundo um dos critérios de Nozick, pois a preferência por este bem é aquela que ocorre durante mais tempo a partir do momento em que o agente escolhe o seu plano: se a duração da implementação do plano for medida pelo número de nódulos que o agente tem de atravessar, então pode-se constatar que a preferência por Y*p*Z ocorre no primeiro nódulo de decisão, mantém-se no primeiro nódulo e volta a ocorrer no segundo nódulo de acaso, enquanto a preferência por X apenas se faz sentir, de modo temporário, no segundo nódulo de escolha.

Conclui-se, assim, que Y*p*Z é realmente um bem futuro que apresenta maior utilidade, durante um maior período de tempo, do que o bem mais imediato X. Dada a racionalidade do mecanismo que consiste em evitar tentações através da formulação de princípios, a viabilidade e a racionalidade do método de tomada de decisão através da escolha resoluta sai, assim, reforçada e, consequentemente, sai também reforçada a ideia de que é racional violar o axioma da independência. Ficamos a saber, pelo menos, que escolher de forma resoluta é uma característica francamente desejável, a qual nos devemos esforçar por adquirir. Fica, contudo, por estabelecer se se trata de um requisito ou exigência de racionalidade. Se o for, temos, então, de aceitar que escolher de forma resoluta é sempre a coisa indicada a fazer. Mas não será esta uma conclusão demasiado forte?

Algo que o banal senso comum permite reconhecer é que, por vezes, ceder às tentações é a coisa mais racional a fazer. Não se podem negar os perigos que as decisões baseadas em princípios e na utilidade simbólica das acções podem acarretar, nomeadamente, o perigo de se incorrer em comportamentos neuróticos. Isto pode acontecer quando as consequências da acção são negativas, mas a sua utilidade simbólica é demasiado elevada, tal como acontece, por exemplo, nos casos de comportamento obsessivo-compulsivo; por outro lado, o apego exagerado a um princípio pode conduzir a uma certa inflexibilidade de carácter e à incapacidade de distinguir situações em que o princípio se aplica de outras

situações que exigem uma atitude de ‘menor seriedade’ ou de adaptação às circunstâncias. Parece ser possível formular um princípio que tenha como objectivo regular a aplicação de todos os outros, respondendo à necessidade de estabelecer condições para a sua violação; contudo, um tal princípio ser-nos-ia de pouca utilidade, pois as condições que regulam as violações terão necessariamente de variar de princípio para princípio. Além disso, um tal princípio de segunda-ordem poderá ser redundante, pois cada princípio de primeira-ordem poderá já incluir uma especificação da forma como será satisfeito; nesta medida, a formulação de princípios é suficientemente liberal, ou flexível, para permitir acomodar qualquer possibilidade de violação, por mais complexamente detalhada que esta possa ser. Isto apenas vem confirmar o que já foi visto acima: os princípios, dada a sua natureza simbólica, não admitem verdadeiras violações. A questão da racionalidade das violações não apresenta, assim, um problema sério para os defensores da escolha resoluta, pois apenas são irracionais as violações não-autorizadas pelos princípios ou, de outra maneira, todas as ‘violações’ autorizadas pelos princípios são consideradas racionais.

Patrick Maher (1993) apresenta algumas considerações e exemplos bastante persuasivos, no sentido de mostrar que a *resolução* não pode ser um requisito de racionalidade. Em primeiro lugar, as nossas preferências podem mudar à medida que o progresso numa árvore de decisão nos oferece nova informação que, à partida, desconhecíamos. Alguém pode decidir licenciar-se em Física e mais tarde vir a saber que as suas capacidades não lhe permitem levar esse plano até ao fim; ou seja, após um nódulo de acaso – em que a natureza nos concedeu, ou não, essas faculdades - as preferências do agente ver-se-ão alteradas. Além disso, estas podem simplesmente alterar-se ao longo do tempo, seja qual for o motivo. Maher oferece o exemplo de um rapaz de oito anos que, não querendo ter nada que ver com raparigas, verá as suas preferências alterarem-se no futuro, o que conduzirá a uma violação da sua *resolução*, sem que, por isso, o tenhamos de considerar irracional. Neste caso, poder-se-ia tentar defender um critério de identidade pessoal que nos permitisse afirmar que o rapaz, ao atingir a puberdade, se torna alguém diferente, não se verificando, desse modo, uma violação da *resolução*. A partir do momento em que existe uma alteração de identidade, estamos na presença de uma outra pessoa e o problema de decisão inicial deixa de existir. Seria, no entanto, extremamente complicado defender esse critério como algo de plausível.

Ainda segundo Maher, o mesmo fenómeno verificar-se-ia no caso das dependências. Um agente prefere inicialmente consumir heroína e parar depois de uma única dose, evitando, assim, qualquer dano permanente; contudo, quando lhe é oferecida uma segunda dose ele já se tornou dependente, o que o levará a aceitá-la, verificando-se, assim, uma violação da *resolução*, devida à alteração de preferências. Tal como no caso do adolescente, também neste teríamos de aplicar um critério de identidade pessoal bastante implausível que nos permitisse afirmar que o agente que decide consumir a segunda dose é essencialmente distinto do agente que decidiu consumir a primeira e depois parar.

A possibilidade de alteração de preferências é, contudo, um problema que parece aplicar-se a vários dos axiomas da teoria e não parece estar relacionado com a natureza específica de cada um deles. Ou seja, é sempre possível que uma alteração das preferências do agente lhe possa causar um prejuízo certo, embora nem sempre tão dramático ao ponto de o transformar numa *money-pump* e o conduzir à falência. Não é difícil conceber casos em que uma alteração de preferências conduz, por exemplo, ao arrependimento devido a uma ou várias trocas anteriores. Nessa medida, este seria um problema relacionado com a escolha em geral e não especificamente com a teoria da utilidade esperada ou com qualquer método de escolha em problemas de decisão sequenciais. Parece-me que a questão da alteração das preferências, ao invés de colocar um problema sério à teoria da decisão, permite-lhe antes limpar as mãos de algo que não lhe diz propriamente respeito. Ou seja, a teoria não foi de todo concebida com o objectivo de resolver as dificuldades para o agente que essa alteração de preferências acarreta.

De modo a defender-se a escolha resoluta, poder-se-ia, por outro lado, estipular que as árvores de decisão não devem ultrapassar um determinado tamanho, devendo as mesmas ser percorridas num período de tempo que não permita a alteração de preferências. Mas isto não só parece uma restrição *ad hoc*, como também dificilmente desejaremos fazer depender o ataque aos axiomas da independência e da transitividade de coisas tão contingentes como a maturidade do agente, a rapidez com que cada um resolve um problema de decisão sequencial ou o tamanho das árvores de decisão que ilustram os argumentos.

Parece-me que, da mesma forma que a alteração de preferências, após tomada uma decisão, não constitui um problema para a teoria da decisão em geral, também o estatuto da resolução como requisito de racionalidade pode não sair afectado mediante a constatação dessas alterações, após a escolha de um plano de acção. Para termos uma

ideia clara sobre a questão é necessário revermos os exemplos oferecidos por Maher. A resolução tem como objectivo evitar as situações de inconsistência dinâmica, ou seja, evitar as falhas no cumprimento de planos, sendo esses planos estipulados com o objectivo de resolver problemas de decisão. O nosso jovem de oito anos tem de decidir se vale a pena manter interacção com o sexo feminino, mais exactamente, se as eventuais vantagens dessa interacção compensam os aborrecimentos que daí podem advir, e chega à conclusão que deve manter-se afastado, estabelecendo um princípio que fortalecerá a sua decisão e o impedirá de ceder a tentações. Mas assim que chega à adolescência, adquirirá nova informação e as suas preferências alterar-se-ão dramaticamente, o que tornará o princípio obsoleto. De um ponto de vista técnico, poder-se-á dizer que a sua função de utilidade se altera, que os estados do mundo relevantes passam a ser diferentes, e que, consequentemente, ele poderá vir a enfrentar um problema de decisão distinto daquele que motivou a resolução da sua escolha. Estas mesmas considerações podem igualmente ser aplicadas ao caso do toxicodependente e, estou em crer, à esmagadora maioria dos casos de alteração de preferências.

Uma maneira de compreendermos melhor aquilo que está em causa quando nos referimos a requisitos de racionalidade passa pela consideração de ciclos resultantes da violação da transitividade da relação de indiferença, e já não da preferência estrita:

$$A \approx B \approx C \succ A$$

Um agente que apresente esta ordenação de preferências pode ser transformado numa *money-pump*. Se o agente começar com um selo A e lhe for solicitada uma quantia adequada para trocar A por C, então ele encontra-se racionalmente obrigado a aceitar essa troca. Além disso, como ele é indiferente entre C e B, e entre B e A, ele aceita trocar C por B e, novamente, B por A, embora, nestes dois últimos casos, sem qualquer custo. Está, assim, completado o ciclo e o agente encontra-se no mesmo estado em que se encontrava inicialmente, menos uma certa quantia em dinheiro. Repetindo-se este procedimento um certo número de vezes, o agente é levado à falência.

Contudo, este não é um resultado inevitável. O agente pode ser transformado numa *money-pump*, mas não necessariamente. Embora ele se encontre racionalmente comprometido com a aceitação da primeira troca, o mesmo já não acontece com as duas trocas seguintes. O requisito de racionalidade está, portanto, associado à satisfação das

preferências do agente; se existe um meio que permita essa satisfação, baseada na sua ordenação de preferências, então o agente deve enveredar por ele.

A vantagem de se ter associado a *escolha resoluta* ao mecanismo de adesão a princípios torna-se, assim, mais evidente. A adesão a princípios normativos intrapessoais é um procedimento claramente racional, pois permite uma plena satisfação dos interesses do agente em problemas de decisão sequencial como aquele que foi acima analisado. As razões por que tal acontece foram já avançadas aquando da análise do referido método: o bem mais distante, Y*p*Z, é mais representativo das preferências do agente do que o bem mais imediato X, que assume o aspecto de uma tentação temporária, pois Y*p*Z é o bem com maior utilidade durante a maior parte do período de tempo que decorre entre a selecção do plano e o seu cumprimento; por outro lado, qualquer princípio de decisão tem de satisfazer um requisito mínimo que consiste em evitar situações de arrependimento. Neste caso, a *escolha resoluta* evita que as decisões do agente resultem em tais situações, pois uma consequência inevitável da violação de princípios, que resulta da natureza simbólica do mecanismo que lhes dá origem, consiste, precisamente, na geração de arrependimento: 'Eu não devia ter comido sobremesa'! Uma outra maneira de nos referirmos a este importante requisito dos princípios consiste em apontar que não devemos agir antes de avaliarmos toda a evidência disponível que nos é necessária para avaliar as consequências das nossas acções, inclusive a evidência que a própria decisão considerada nos oferece. Portanto, ao procedermos ao típico raciocínio deliberativo condicional, devemos avaliar não só as consequências objectivas da acção, mas também considerar se ficaremos satisfeitos por sermos a pessoa que somos após termos realizado a acção, nomeadamente, alguém que não é capaz de vencer tentações.

Conclui-se, assim, que o defensor da *escolha resoluta* pode encarar como natural a violação do axioma da independência no caso acima analisado, dado que é possível, a meu ver, estabelecer este método de escolha como um requisito de racionalidade.[41] Ainda

[41] Rabinowicz (1995; 2000) apresentou o seu próprio método de escolha para problemas de decisão sequencial, o qual designou como 'wise choice'. Segundo Rabinowicz (2000: 150): 'According to this suggestion, an agent in a sequential decision problem considers various theoretically possible action plans and treats as feasible only those plans he expects he would implement upon adoption. (…) This kind of *wise choice* makes room for resoluteness (resolute agents will have more plans that are feasible), but does not demand it. This *self-predictive* element closely reminds of the backward-induction procedure used by a sophisticated chooser. Still, there is a difference: The wise chooser makes self-predictions about his future behavior taking into consideration his dispositions to act on adopted plans. Remember that he makes them *conditionally* on various hypotheses about which plan he is going to adopt. This means that he does not necessarily see his future choices as «separable» from (i.e., uninfluenced by) his earlier resolutions: Such separability may obtain in some cases but it need not be a rule. As we have seen, some wise choosers may

que seja difícil estabelecer conclusões definitivas sobre este tema, a justificação do axioma da independência, enquanto condição de racionalidade a ser imposta às nossas ordenações de preferências, tem de ser oferecida por outros exemplos que não os de inconsistência dinâmica (através, quem sabe, de uma investigação empírica baseada em decisões de agentes reais, tal como a de Tversky para o caso da transitividade).[42]

Dá-se, assim, por encerrada a discussão acerca dos fundamentos da teoria bayesiana da decisão. Tal discussão permitiu-nos adquirir alguns conceitos fundamentais, o de função de utilidade e o de distribuição de probabilidades do agente, necessários a uma plena compreensão das problemáticas tratadas adiante. Um esclarecimento acerca do estatuto científico da teoria, nomeadamente a rejeição de uma abordagem comportamentalista, foi também considerado necessário, tendo em conta alguns contextos em que aspectos essenciais da teoria são frequentemente discutidos. Finalmente, não seria possível obter uma visão panorâmica da teoria sem abordar a questão da sua justificação.

Na parte 2 discutir-se-á o problema que motivou a actual controvérsia acerca de qual é a interpretação correcta - evidencialista ou causalista - do princípio da maximização da utilidade esperada.

well be less than resolute (…). Under such circumstances, wise choice can coincide with sophisticated choice'. À falta de exemplos, tenho dificuldade em interpretar esta sugestão, mas o que me parece estar em causa é o seguinte: a resolução é uma coisa boa para se ter, mas que não pode ser exigida aos agentes enquanto requisito de racionalidade. Rabinowicz pretende assim introduzir procedimentos que fortaleçam o método da escolha sofisticada. Como aqui argumentei que a escolha resoluta deve ser entendida como um requisito de racionalidade, escuso-me a discutir em detalhe a proposta deRabinowicz.

[42] A versão da teoria da utilidade esperada de Jeffrey (1964), com o seu teorema da representação (demonstrado por Ethan Bolker em 1966), não necessita do axioma da independência. A independência requer que as probabilidades dos estados do mundo sejam independentes das acções escolhidas, tal como acontece na teoria de Savage. Mas, na teoria de Jeffrey, as probabilidades dos estados do mundo, como se verá adiante, consistem em probabilidades condicionais, em que a condição é a acção escolhida. Assim, mesmo aqueles estados em que as consequências são idênticas podem ter influência para a escolha do agente, pois as probabilidades desses estados são as probabilidades condicionais dos mesmos, dada a escolha de uma ou outra acção.

Parte 2 – O Problema

4. Conflito de princípios e batalha de intuições

4.1. A teoria de Jeffrey

O Problema de Newcomb é frequentemente incluído nas antologias de paradoxos. O termo 'paradoxo' tem sido utilizado de formas demasiado distintas para que se preste a uma definição geral, embora em todos esses usos se verifique sempre a ideia de dificuldade insuperável ou de beco sem saída.

Existe, contudo, uma definição de 'antinomia', um termo muitas vezes utilizado como sinónimo de paradoxo ou descrito como sendo o caso mais extremo de paradoxo. Segundo esta definição, verifica-se a existência de uma antinomia quando nos encontramos perante duas proposições contrárias ou contraditórias, derivadas conjuntamente a partir de argumentos que não se revelaram incorrectos fora do contexto particular que gera o paradoxo. Ou seja, partindo de premissas que são geralmente aceites como verdadeiras, podem-se inferir duas proposições contrárias ou contraditórias. No caso do Problema de Newcomb, o que temos são duas proposições que se contradizem acerca da racionalidade de duas acções possíveis. Assim sendo, apenas uma delas poderá ser verdadeira.

A formulação que irei utilizar é, no essencial, equivalente à de Nozick (1969):[43]

Temos perante nós duas caixas: uma delas é transparente e contém 1,000 Euros; a outra é opaca e contém ou 1,000,000 Euros ou nada. É-nos dada a escolha entre ficarmos com as duas caixas ou apenas com a caixa opaca. O conteúdo da caixa opaca é determinado da seguinte maneira: existe um previsor cujas previsões têm uma elevada taxa de sucesso. Até agora ele previu correctamente todas as nossas decisões passadas. E, tanto quanto sabemos, ele tem vindo a prever correctamente as decisões de outros indivíduos mais ou menos semelhantes a nós. Se o previsor previu que iremos escolher as duas caixas, ele deixou caixa opaca vazia. Se o previsor previu que iremos escolher apenas a caixa opaca, ele colocou 1,000,000 Euros na caixa opaca. Primeiro, o previsor coloca, ou não, o milhão na caixa opaca, depois efectuamos a nossa escolha. O que devemos fazer para satisfazer

[43] Segundo Nozick, o problema foi inventado pelo Dr. William Newcomb, professor e físico teórico no Laboratório Lawrence Livermore da Universidade da Califórnia. Tanto quanto se sabe, o próprio Newcomb nada publicou acerca do problema.

a nossa preferência, sendo que esta consiste em adquirir a maior quantidade possível de dinheiro?

O aspecto paradoxal do problema resulta da existência de dois argumentos distintos, que recorrem a princípios de racionalidade igualmente distintos - embora bem-estabelecidos - mas que oferecem soluções incompatíveis entre si. Chamemos ao argumento favorável à escolha da caixa opaca, argumento monocaixista; e bicaixista ao argumento favorável à escolha das duas caixas. Ambos são bastante persuasivos e cada um deles tem a sua quota de notáveis defensores, o que se tornará por demais evidente daqui em diante.

Um dos dados do problema é que o previsor tem uma elevada taxa de sucesso nas suas previsões. Portanto, se escolhermos ficar com as duas caixas, é muito provável que ele o tenha previsto e, como tal, tenha deixado a caixa opaca vazia. Ou seja, temos quase a certeza de que as nossas escolhas corresponderão às suas previsões. Logo, se queremos ficar milionários, devemos escolher ficar apenas com a caixa opaca.

Por outro lado, outro dos dados do problema é que a nossa escolha é feita depois de o previsor ter colocado ou não o milhão na caixa opaca, não podendo depois disso a nossa escolha afectar o seu conteúdo. Logo, façamos o que fizermos, parece ser sempre melhor escolher as duas caixas: se o previsor errou, ficamos com 1,001,000 em vez de apenas 1,000,000; se o previsor acertou ficamos com 1,000 em vez de zero.

Será útil formular claramente as premissas de cada um destes argumentos, pois uma certa análise irá basear-se na disputa acerca da verdade das suas premissas. Esta formulação pode ser encontrada, no seu essencial, em Terence Horgan (1981).

Argumento monocaixista:

1. Se eu escolhesse as duas caixas, então o previsor prevê-lo-ia.
2. Se eu escolhesse as duas caixas e o previsor o previsse, então ganharia 1,000.
3. Se eu escolhesse as duas caixas, então ganharia 1,000.
4. Se eu escolhesse a caixa opaca, então o previsor prevê-lo-ia.
5. Se eu escolhesse a caixa opaca e o previsor o previsse, então ganharia 1 milhão.
6. Se eu escolhesse a caixa opaca, então ganharia 1 milhão.
7. Se 3 e 6 forem verdadeiras, então devo escolher a caixa opaca.

Argumento bicaixista:

1'. A caixa opaca contém 1 milhão ou está vazia.

2'. Se contiver 1 milhão, então eu receberia 1,001,000 se escolhesse as duas caixas.

3'. Se contiver 1 milhão, então eu receberia 1 milhão se escolhesse apenas a caixa opaca.

4'. Se estiver vazia, então eu receberia 1000 se escolhesse as duas caixas.

5'. Se estiver vazia, então eu não receberia nada se escolhesse apenas a caixa opaca.

6'. Das duas uma: receberia 1,001,000 se escolhesse as duas caixas e 1 milhão se escolhesse apenas a caixa opaca *ou* receberia 1000 se escolhesse as duas caixas e nada se escolhesse apenas a caixa opaca.

7′. Se 6' for verdadeira, então devo escolher as duas caixas.

Desde Nozick, a abordagem tradicional do problema consiste em encará-lo como um caso de conflito entre dois princípios da racionalidade instrumental, cada um dos argumentos acima invocando um princípio em detrimento do outro. O argumento monocaixista pode ser defendido através da aplicação da teoria bayesiana da decisão, ou seja, invocando o princípio da maximização da utilidade subjectiva esperada.

O argumento bicaixista, por seu lado, invoca o princípio da dominação: se, num problema de decisão, uma determinada acção *a* dominar todas as outras acções possíveis, então *a* deve ser executada. Assim, entre duas acções *a* e *b*, é racional executar *a* em vez de *b*, caso as duas seguintes condições se encontrem satisfeitas: 1) aconteça o que acontecer, fazer *a* nunca faz com que se fique pior do que fazendo *b*, e 2) existe, pelo menos, uma consequência possível de fazer *a* que faz com que se fique melhor do que fazendo *b*. Substituindo na matriz abaixo, que representa o problema de Newcomb, *a* pela acção de escolher as duas caixas e *b* pela acção de escolher a caixa opaca, obtemos uma ilustração aparentemente clara de uma situação em que o princípio da dominação deve ser aplicado:

	Previsor prevê que serão escolhidas duas caixas	Previsor prevê que será escolhida caixa opaca
Escolher duas caixas	1,000	1,000 + 1,000,000
Escolher caixa opaca	0	1,000,000

Tendo em conta que o previsor efectua a sua escolha em *t*0 e a acção é escolhida em *t*1, e desde que não estejamos dispostos a defender que a causalidade pode funcionar para trás no tempo, então, seja qual for o estado do mundo que se verifique (esteja ou não correcta a previsão), escolher as duas caixas faz sempre com que se fique melhor do que optando pela acção contrária.

O conceito de utilidade esperada que se encontra na base do argumento monocaixista é o mesmo que é adoptado pela teoria de Jeffrey (1964), ou seja, não se trata de calcular simplesmente a utilidade esperada de uma acção, mas sim a utilidade condicional esperada da mesma. Na medida em que os estados do mundo relevantes para o problema não são probabilisticamente independentes das acções empreendidas, não se utilizarão, para efeitos de cálculo, as probabilidades incondicionais dos estados, mas sim as probabilidades desses estados dadas as acções. Assim, especifique-se o conjunto *A* de proposições que descrevem as acções disponíveis, o conjunto *E* de proposições que descrevem os estados do mundo relevantes e o conjunto *C* de proposições que descrevem as consequências das acções sob cada um dos estados do mundo. O conjunto *C* é obtido a partir dos outros dois na medida em que $C = \{a \wedge e \mid a \in A, e \in E\}$. Cada $c \in C$ corresponde a uma célula na matriz apresentada. Assim, a utilidade condicional esperada de uma acção é definida da seguinte maneira:

$$\mathrm{UCE}\,(a) = \sum_{(i=1)}^{n} pr(ei|a) \times u(a \wedge ei),$$

em que *pr* é uma função de probabilidade subjectiva, definida numa álgebra – Ω de proposições incluindo *C*, *A* e *E*, enumerável e fechada sob negação e disjunção; *pr*(x|y) é a probabilidade condicional de x dado y; e *u* uma função de utilidade definida em *C*.[44]

Seja *C1* a acção de escolher a caixa opaca, *C2* a acção de escolher ambas as caixas, *PC1* o estado em que o previsor prevê correctamente a acção de escolher a caixa opaca e *PC2* o estado em que o previsor prevê correctamente a acção de escolher ambas as caixas. Considere-se também que a utilidade de cada uma das consequências é linear com o valor monetário encontrado nas caixas. Assim,

[44] Com estes dados temos também a seguinte definição formal de dominação: uma acção $a \in A$ domina (fracamente) uma acção $b \in A$ se, e somente se, para todos os estados $e \in E$, $u(a \wedge e) \geq u(b \wedge e)$, e existe pelo menos um estado *e* tal que a $u(a \wedge e) > u(b \wedge e)$, em que *u* é uma função de utilidade definida em *C*.

$$\text{UCE (C2)} = pr(PC2|C2) \times u(C2 \wedge PC2) + pr(PC1|C2) \times u(C2 \wedge PC1)$$
$$= pr(PC2|C2) \times 1{,}000 + (1 - pr(PC2|C2)) \times 1{,}001{,}000$$

$$\text{UCE (C1)} = pr(PC2|C1) \times u(C1 \wedge PC2) + pr(PC1|C1) \times u(C1 \wedge PC1)$$
$$= pr(PC2|C1) \times 0 + pr(PC1|C1) \times 1{,}000{,}000.$$

Supondo que $pr(PC2|C2) = pr(PC1|C1)$, pois não existe qualquer razão para acreditarmos que o Previsor é mais fiável a prever uma acção do que outra, segue-se que a UCE (C1) > UCE (C2) desde que a o grau de fiabilidade do previsor seja superior a 0,5005, valor ligeiramente superior ao de uma previsão feita através do lançamento de uma moeda não-viciado. Ou seja, $pr(PC2|C2) = pr(PC1|C1) > 0{,}5005$. Como sabemos que o Previsor é extremamente fiável, segue-se que a UCE (C1) > UCE (C2).

Tendo em conta que foi o próprio Nozick que, em primeiro lugar, abordou o problema sob o ponto de vista de um conflito entre princípios, é justo incluir aqui as suas reflexões e conclusões. Existem infindáveis exemplos de problemas de decisão em que é recomendável colocar de lado o princípio da dominação. Trata-se de casos em que os estados do mundo não são probabilisticamente independentes das acções. Um exemplo claro é oferecido pelo próprio Jeffrey (1983: 8-9): dois países com potencial nuclear têm de decidir se devem ou não desarmar. O defensor do desarmamento apresenta a seguinte matriz de decisão e argumenta usando o princípio da dominação:

	Armar	Desarmar
Armar	−100	0
Desarmar	−50	50

É preferível viver em condições deploráveis sob o domínio do inimigo a enfrentar a aniquilação total através de uma guerra nuclear (−50 > − 100), e é preferível viver em paz a viver na presente situação de hostilidade permanente (50 > 0). Contudo, um defensor da manutenção de um arsenal nuclear para efeitos preventivos pode defender o seu caso e aceitar na mesma os *payoffs* da matriz apresentada. Se houver razões para acreditar que desarmar aumenta significativamente a probabilidade de uma guerra, então armar pode

tornar-se na acção com maior utilidade esperada. Isto verifica-se quando existe uma probabilidade de 0.8 de haver uma guerra em caso de desarmamento e apenas uma probabilidade de 0.2 de haver uma guerra em caso de armamento.

O que distingue, então, este caso do Problema de Newcomb, tendo em conta que, em ambos, os estados do mundo não são probabilisticamente independentes das acções? A resposta de Nozick é aquela que tem vindo a ser defendida pelos defensores do bicaixismo sob formas cada vez mais sofisticadas: embora, no Problema de Newcomb, a escolha das acções influencie de forma legítima a estimativa das probabilidades dos estados, essa escolha não irá influenciar a obtenção de qualquer um deles, ao contrário do que acontece no caso da guerra nuclear. Nozick coloca as coisas em termos de um engano quanto à 'ordem da explicação' dos eventos cruciais, a qual deve seguir a ordem causal desses mesmos eventos. A escolha da acção não explica (nem causa) a previsão efectuada, embora tenha sido um determinado estado em que o agente se encontrava (presumivelmente), no momento da previsão, que explica (e causa) a previsão. Estaríamos, assim, perante um caso em que existe apenas uma ilusão de que a acção pode controlar ou influenciar a probabilidade de obtenção dos estados. A conclusão de Nozick é a de que, em casos semelhantes, o princípio da maximização da utilidade esperada deve ser posto de lado em detrimento do princípio da dominação.

4.2. Notas preliminares

Penso que é importante, para já, apresentar alguns argumentos em prol de cada uma das escolhas sem entrar em detalhes técnicos e mantendo esses argumentos apenas a um nível intuitivo. A intuição por detrás da opção representada pela análise de Nozick é clara e não precisa de ser muito mais adensada. A ideia é a de que os monocaixistas sofrem de uma espécie de *wishful thinking*, tomando uma acção que constitui apenas um sinal ou um augúrio de uma consequência desejada por uma acção que é realmente eficaz na produção dessa consequência.

Mas existem também duas maneiras de tentar desarmar a intuição que se encontra na base do bicaixismo. A primeira consiste em imaginar uma situação em que o Previsor, ao invés de ser descrito como extremamente fiável, é tido como infalível. (No sentido em que temos a certeza de que ele sabe sempre qual será a nossa escolha e não infalível num

sentido frequentista, em que até então todas as suas previsões se mostraram acertadas). Com um previsor infalível duas consequências parecem ser eliminadas: 1) escolher duas caixas e ganhar 1,001,000 Euros, e 2) escolher uma caixa e não ganhar nada. Restam, assim, apenas duas consequências, dependendo de nós qual delas se verificará: 3) escolher duas caixas e ganhar 1,000 Euros, e 4) escolher a caixa opaca e ganhar 1,000,000 Euros. Nesta situação parece difícil sustentar que é racional agirmos com base em duas situações hipotéticas cuja ocorrência é impossível; logo, devemos escolher a caixa opaca e garantir o milhão. Portanto, se é racional escolher a caixa opaca quando o previsor é infalível, por que deixará de ser racional escolher a caixa opaca quando o Previsor tem uma taxa de sucesso, por exemplo, de 0.9? Se não encontrarmos nenhuma razão forte para mudarmos a nossa atitude quando o previsor passa de infalível a extremamente fiável, veremos a plausibilidade do bicaixismo ser severamente diminuída.

A intuição de irracionalidade associada ao bicaixismo pode ainda ser fortalecida se constatarmos um aspecto do problema que, no caso limite em que se considera a infalibilidade do previsor, se torna saliente. Se tivermos a certeza de que apenas 3) e 4) são possíveis, então o conteúdo da caixa opaca encontra-se numa relação de dependência lógica relativamente ao conjunto das minhas crenças acerca da infalibilidade do previsor e das consequências possíveis das acções. Da crença nas premissas 3 e 6 do argumento monocaixista, que têm a mesma forma lógica das condicionais contrafactuais, pode-se deduzir a crença nas duas seguintes condicionais materiais: 3* Escolho as duas caixas → Recebo 1,000 e 6* Escolho a caixa opaca → Recebo 1,000,000. Resta-me, com a minha acção, adicionar a premissa que permite o *modus ponens*, dada a certeza da minha crença quanto ao conteúdo da caixa opaca. É plausível que na base das nossas intuições de racionalidade associadas ao caso limite, em que o previsor é infalível, se encontre a consciência que temos da verificação desta dependência lógica.

Da mesma forma que a mudança da nossa atitude não se encontra justificada aquando da passagem da infalibilidade do previsor para a sua extrema fiabilidade, também não parece encontrar-se justificada uma mudança da nossa atitude quando deste raciocínio passamos para uma inferência do mesmo tipo, embora formulada em termos probabilísticos: 3** Eu escolho as duas caixas → Tenho uma probabilidade de 0,9 de receber 1,000, e 6** Eu escolho a caixa opaca → Tenho uma probabilidade de 0,9 de receber 1,000,000. Estes dois tipos de dependência, lógica e probabilística, servem para contrabalançar o facto

inegável, que se encontra na base da intuição a favor do bicaixismo, de que o conteúdo da caixa opaca é causalmente independente das acções empreendidas.

Outro dos argumentos mais fortes a favor da posição monocaixista recorre ao nosso entendimento comum daquilo em que consiste a racionalidade instrumental e o sucesso dos nossos empreendimentos: escolher os meios mais adequados para satisfazermos os nossos principais desejos; uma outra maneira de o dizer é a seguinte: se mais dinheiro é melhor do que menos dinheito, então a acção que me garante o milhão, e não a que oferece apenas mil, é certamente a acção racional. Este argumento pode ser resumido a uma simples provocação, lançada pelo monocaixista: 'A tua suposta racionalidade manteve-te na pobreza, enquanto a minha irracionalidade me converteu em milionário'! Terá o bicaixista uma resposta à altura para oferecer? Apesar de ter de conceder [*bite the bullet*] que ser irracional pode algumas vezes ser vantajoso, ele pode também manter que este facto não faz com que seja racional escolher a caixa opaca. Imaginemos o seguinte cenário: existem dois tipos de personalidade, bicaixista e monocaixista, e o previsor é um genial avaliador de personalidades, nunca tendo falhado uma avaliação feita através de uma observação atenta dos agentes (como foi dito acima, o estado em que se encontram os agentes no momento da previsão é a causa da previsão do previsor). Ou imagine-se, por exemplo, que o previsor é um neurocientista que é capaz de identificar o tipo de personalidade dos agentes mediante a observação de diferenças consistentes nos padrões das ondas cerebrais dos agentes. Em ambas as situações, parece que alguém com personalidade bicaixista está condenado a nunca receber mais de 1,000 e alguém com personalidade monocaixista está condenado a ser milionário. Ou seja, o bicaixista pode responder que, apesar de ter permanecido pobre, as opções que o monocaixista teve nunca estiveram ao seu alcance. Dada a certeza que ele tem de que a caixa opaca está vazia, resta-lhe somente a possibilidade de não deixar escapar os mil Euros. Em suma, o bicaixista não parece ter razões para lamentar a sua escolha, apesar de ter boas razões para invejar o monocaixista, principalmente se o seu desejo por dinheiro for particularmente intenso.

Face a isto o monocaixista voltará a insistir: 'Então confessas! Gostarias de ser como eu! Isso quer dizer que, afinal, eu sou mais inteligente do que tu'! Se quanto a este último ponto as águas continuam demasiado turvas para podermos chegar a uma conclusão definitiva, quanto ao primeiro parece não existirem dúvidas. O bicaixista que deseja ser milionário teria certamente desejado possuir a personalidade monocaixista, antes de lhe

ter sido apresentada a escolha. Se lhe tivesse sido oferecida uma 'pílula da irracionalidade' (segundo a sua perspectiva), ele tê-la-ia certamente tomado. Mais, se o mundo em que vivemos for do tipo em que os Problemas de Newcomb são uma presença constante, penso que todos teríamos boas razões para tomarmos 'pílulas da irracionalidade'. Mas, apesar de confessar a sua inveja, o bicaixista coerente tem de manter que escolher as duas caixas é a acção racional, seja qual for a nossa personalidade, bicaixista ou monocaixista. Uma maneira de fortalecer esta posição e responder com a mesma moeda à ironia do monocaixista passa por colocar a este último uma questão que, de tão óbvia, parece escapar aos potenciais milionários: 'Por que motivo recusaste aumentar a tua fortuna em mil Euros'? Ou seja, mesmo que o monocaixista esteja seguro acerca da sua personalidade monocaixista, e de que o milhão está à sua disposição dentro da caixa opaca, os mil Euros continuam à vista de todos na caixa transparente. É neste sentido que o bicaixista gostaria de ter tido as opções que o monocaixista teve - juntar mil Euros à sua fortuna ou recusá-los - pois as suas escolhas foram muito mais pobres – ganhar mil Euros ou ficar de mãos vazias. Portanto, a pergunta do bicaixista faz todo o sentido e obriga a que o monocaixista ofereça uma resposta firme e clara, a qual faça justiça à sua convicção. Mas que resposta poderá ser essa? Voltar a invocar o seu estatuto de milionário não parece ser realmente uma resposta, mas apenas fugir ao assunto. Será este raciocínio bicaixista suficientemente persuasivo ou temos de continuar a admitir que a imagem dos monocaixistas a celebrar, abrindo garrafas de Dom Pérignon, é demasiado forte quando comparada com a desolação geral vinda do campo bicaixista, apesar destes últimos continuarem a erguer a sua bandeira com orgulho?

Em suma, o que cada um dos defensores dos diferentes argumentos terá de fazer é tentar convencer-nos de que existe uma assimetria em favor da intuição ou intuições associadas à sua posição. Nesta altura, podemos ter uma opinião segura acerca da racionalidade de cada uma das escolhas, seja pelo facto de encontrarmos alguma das intuições acima apresentadas absolutamente inescapável, seja por razões teóricas mais sofisticadas. Todavia, a força das nossas convicções pode ser submetida a um duro teste: a manipulação do valor contido na caixa transparente. Poder-se-á reduzir esse valor até se tornar insignificante ou aumentá-lo até quase igualar o valor da caixa opaca. No primeiro caso, se reduzirmos o valor da caixa transparente para 1 Euro, será que o defensor mais acérrimo do bicaixismo manterá a sua fé na intuição de racionalidade da sua posição? Se aumentarmos, no segundo caso, o valor da caixa transparente para 900,000 Euros,

manterá o defensor do monocaixismo a fé na sua convicção? Se este último acreditar que a adopção do princípio da maximização da utilidade esperada, na interpretação de Jeffrey, constitui uma condição necessária e suficiente para determinar a racionalidade da acção em qualquer circunstância, então ele deverá em princípio optar sempre pela caixa opaca, mesmo quando a taxa de sucesso do previsor é apenas de 0,5005 ou quando o valor da caixa é menor do que 980,000 Euros.

Devido ao facto de a alteração dos valores das caixas poder ter influência nas intuições dos agentes acerca da racionalidade das acções, pode haver quem atribua à teoria da decisão um estatuto idêntico aos princípios da lógica indutiva em geral, admitindo o risco inerente às decisões tomadas com base nas conclusões obtidas por intermédio desses princípios, e desejando atribuir à teoria uma plasticidade que lhe permita adaptar-se às circunstâncias do mundo, admitindo que, por vezes, podemos esbarrar com situações paradoxais que, embora raras, parecem estabelecer limites à racionalidade humana. Ou seja, por razões prudenciais, a confiança que depositamos nos princípios da decisão racional, e nas escolhas e argumentos a estes associados, pode ser variável, não apenas de indivíduo para indivíduo, mas também para um único indivíduo, de acordo com as circunstâncias.

Por outro lado, outros existirão que não desejam ter em conta razões relacionadas com a adequação pragmática dos princípios de uma teoria da decisão, encarando o exemplo da manipulação da caixa transparente apenas como algo que pode afectar e iludir a psicologia dos agentes e defendendo que a interpretação adoptada do princípio da utilidade esperada é válida para todos os problemas de decisão, independentemente da variação subjectiva das nossas intuições de racionalidade. Resta saber se essa consistência teórica pode ser obtida a um preço que não seja demasiado elevado. Mais adiante, ao serem introduzidos novos instrumentos teóricos de análise, e convertendo-se a discussão num confronto entre diferentes teorias da decisão, será mais fácil compreender as razões por detrás de cada posição e o desejo de manter a consistência teórica, tanto por parte de bicaixistas como de monocaixistas.

5. Argumentos falaciosos?

5.1. Maximização

Antes de se procurarem soluções para o Problema de Newcomb (daqui em diante PN), soluções que poderão passar pela eleição de um ou outro dos argumentos de Nozick, aduzindo-se razões que favoreçam um deles, poder-se-á questionar o carácter genuinamente paradoxal do problema e a justificação do emprego dos princípios de racionalidade em causa. Poder-se-á, por exemplo, questionar se estão reunidas todas as condições necessárias para a aplicação desses princípios, na justificação de uma ou outra das soluções.

Isaac Levi (1975) argumentou, precisamente, no sentido de mostrar que as condições que especificam o PN não são suficientemente detalhadas para a aplicação do princípio da maximização da utilidade condicional esperada; ou seja, que a aplicação desse princípio seria falaciosa. A posição favorável ao monocaixismo não estaria, portanto, justificada. Considere-se novamente a seguinte matriz:

	PC1	PC2
C1	$M	$0
C2	$M + $1000	$1000

C1 representa a acção monocaixista, C2 a acção bicaixista, PC1 o estado do mundo em que a previsor prevê C1 e PC2 o estado do mundo em que o previsor prevê C2. A quase-infalibilidade do previsor pode, então, ser representada pela ideia de que a $pr((C1 \wedge PC1) \vee (C2 \wedge PC2))$ é bastante elevada. Isto significa que as previsões do previsor e as escolhas dos agentes coincidem na grande maioria dos casos. Segundo Levi, o que Nozick pretende veicular é uma interpretação da ideia de infalibilidade, de acordo com a qual tanto a $pr(C1|PC1)$, como a $pr(C2|PC2)$ são bastante elevadas, do que se segue a também elevada probabilidade da disjunção acima. Estas últimas são as probabilidades de o agente escolher esta ou aquela acção dada a previsão do previsor. A questão essencial para o argumento de Levi é a de que não se devem confundir as probabilidades condicionais acima com estas outras: a $pr(PC1|C1)$ e a $pr(PC2|C2)$, a probabilidade de o previsor prever correctamente uma ou outra acção, dada a escolha do agente. Estas são aquelas

que se utilizam no cálculo da utilidade condicional esperada. Ora, mesmo que as primeiras sejam ambas elevadas, não se segue daí que as segundas também o sejam. Seria, portanto, a confusão entre estes pares de probabilidades que, segundo Levi, resultaria numa aplicação falaciosa do princípio da maximização da utilidade condicional esperada. Para ilustrar o seu ponto, ele apresenta três casos distintos em que a infalibilidade do previsor não estaria em causa, mas em que as recomendações do princípio seriam diferentes (1975: 164). Suponha-se que o PN foi apresentado a três grupos de mais de um milhão de indivíduos e que os resultados foram os seguintes:

Grupo 1.	Previsão escolha caixa opaca	/	Previsão escolha duas caixas
	Escolhem caixa opaca – 900 000		Escolhem caixa opaca -10
	Escolhem duas caixas - 100 000		Escolhem duas caixas – 90
Grupo 2.	Previsão escolha caixa opaca	/	Previsão escolha duas caixas
	Escolhem caixa opaca – 495 045		Escolhem caixa opaca - 55 005
	Escolhem duas caixas – 55 005		Escolhem duas caixas - 495 045
Grupo 3.	Previsão escolha caixa opaca	/	Previsão escolha duas caixas
	Escolhem caixa opaca – 90		Escolhem caixa opaca – 100 000
	Escolhem duas caixas – 10		Escolhem duas caixas – 900 000

A meu ver, a primeira coisa que devemos procurar fazer, na análise deste quadro de resultados, é determinar onde se devem ler os pares distintos de probabilidades condicionais. A probabilidade de os agentes escolherem uma ou outra acção, dada a previsão do previsor, é lida na vertical. Constata-se, assim, que em todas as colunas, em todos os três casos, estas probabilidades são elevadas: sempre 0.9. Por exemplo, no primeiro caso, de 900.000 em 1.000.000 para *pr*(C1|PC1), e de 90 em 100 para *pr*(C2|PC2).

Já a probabilidade de o previsor acertar na sua previsão, dada a escolha de uma ou outra acção, é lida na horizontal. Por exemplo, no caso 1 o previsor é muito bom a prever C1, a *pr*(PC1|C1) = 900.000/900.010 = 0.99998888, mas bastante fraco a prever C2, a *pr*(PC2|C2) = 90/100.090 = 0.0008991. No caso 2, o previsor é muito bom tanto a prever C1, como C2, o que faz com que estas probabilidades sejam ambas elevadas. E no caso

3, o previsor é muito bom a prever C2, mas bastante fraco a prever C1. Isto terá obviamente repercussões no cálculo da utilidade nos três casos. Considere-se a utilidade linear com os valores em dinheiro. Nos casos 1 e 3 o cálculo da utilidade condicional esperada favorece a decisão de escolher as duas caixas. E apenas no caso 2, em que a *pr*(C1) = *pr*(C2), quando o previsor prevê o monocaixismo e o bicaixismo com a mesma probabilidade, é que C1 é recomendado. Consideremos para efeito de ilustração:

Caso 1:

UCE (C1) = *pr*(PC1|C1) × *u*(PC1 & C1) + *pr*(PC2|C1) × *u*(PC2 & C1) =
= 0,9999888 × 1 000 000 + 0,0000112 × 0 =
= 999 988,8

UCE (C2) = *pr*(PC1|C2) × *u*(PC1 & C2) + *pr*(PC2|C2) × *u*(PC2 & C2) =
= 0,9991009 × 1 001 000 + 0,0008991 × 1000 =
= 1 000 100 + 0,8999 ≅ 1 000 101

Caso 2:

UCE (C1) = *pr*(PC1|C1) × *u*(PC1 & C1) + *pr*(PC2|C1) × *u*(PC2 & C1) =
= 0,9999888 × 1 000 000 + 0,0000112 × 0 =
= 999 988,8

UCE (C2) = *pr*(PC1|C2) × *u*(PC1 & C2) + *pr*(PC2|C2) × *u*(PC2 & C2) =
= 0,0000112 × 1 001 000 + 0,9999888 × 1000 =
= 1012

Segundo Levi, as possibilidades envolvidas nestes três casos tornam injustificada a utilização do princípio da maximização da utilidade esperada numa situação como a do PN, pois sem as estatísticas relevantes, como aquelas que definem os três casos acima, o agente não tem dados suficientes para tomar a sua decisão, mesmo sabendo que o previsor é quase infalível. Ou seja, de modo a aplicar o princípio de forma justificada, e mesmo sabendo que, de acordo com a interpretação lata de infalibilidade - *pr*((C1∧PC1) ∨ *pr*(C2∧PC2)) ≅ 1 - o agente tem igualmente de saber qual o valor de cada uma das probabilidades condicionais relevantes, a *pr*(PC1|C1) e a *pr*(PC2|C2). Uma outra maneira

de colocar a questão é afirmar que não sabemos de que caso é o PN uma variante: do caso 1, 2 ou 3.

Admitindo que no PN as probabilidades condicionais relevantes são indeterminadas, e que, mesmo assim, estamos perante um problema de decisão que merece ser levado a sério, Levi aconselha ao agente o uso da regra *maximin* como princípio de racionalidade adequado para resolver este problema, ou seja, dever-se-ia maximizar o valor mínimo (de utilidade) que é possível obter, seja qual for o estado do mundo que vier a ser o caso. Conclui, assim, o argumento: C2 é a acção que é racional empreender no PN, sendo este um problema em que as probabilidades condicionais utilizadas para o cálculo da utilidade são indeterminadas. Deixaríamos, assim, de ter um problema de decisão em condições de incerteza e passamos a ter um problema de decisão sob ignorância.

Mas não será verdade que a infalibilidade do previsor pode também ser representada pelas probabilidades condicionais utilizadas no cálculo da utilidade esperada e não pelas suas inversas? Pode, assim, colocar-se a seguinte questão: por que motivo associa Levi a infalibilidade do previsor às primeiras, *pr*(C1|PC1) e (C2|PC2), e não às segundas, *pr*(PC1|C1) e *pr*(PC2|C2)? Embora ele não ofereça uma resposta, podemos nós ensaiar uma: sendo que, de acordo com os dados do problema, o previsor faz primeiro a sua previsão e só depois é dado ao agente escolher, esta é uma ordem de eventos que parece favorecer a associação de Levi. Segundo a expressão de Nozick, esta interpretação estaria de acordo com a 'ordem da explicação'. É perfeitamente normal que, ao falarmos de dois eventos, o evento condicionante seja anterior ao evento condicionado, podendo considerar-se mais correcto falar da escolha dada a previsão dessa escolha, do que da previsão dada a escolha. Contudo, a apresentação de Nozick torna suficientemente clara a inexistência de qualquer relação de causalidade entre a previsão, ou o conteúdo da caixa opaca, e a escolha do agente. Fazer anteceder a previsão da escolha serve apenas para acentuar a inexistência dessa relação de causalidade. É, portanto, irrelevante que a escolha do agente seja feita antes, durante ou depois da previsão, desde que a mesma não seja do conhecimento do previsor. Dado isto, torna-se, a meu ver, indiferente falarmos da probabilidade da escolha dada a previsão ou da probabilidade da previsão dada a escolha, passando a não existir qualquer motivo para interpretar a infalibilidade do previsor em associação com a *pr*(C1|PC1), e não em associação com a *pr*(PC1|C1). Além disso, uma característica comum aos três casos apresentados é que a fiabilidade do previsor, medida, de acordo com Levi, pela segunda destas probabilidades, depende contingentemente da

distribuição de casos de C1 e C2. Mas isto não parece ser o que Nozick tinha em mente. A infalibilidade do previsor, no sentido que Nozick associa ao termo, consiste no facto de ele ser muito bom a prever qualquer um dos casos, C1 ou C2.

Independentemente da maneira como se interpreta a fiabilidade do previsor, Ellery Eells (1984a) veio mostrar que, para se verificar uma recomendação de C1, não é necessário que ambas as probabilidades relevantes para o cálculo da utilidade esperada sejam altas. Mais ainda, é até possível que o facto de serem ambas baixas seja compatível com a recomendação de C1.

O que Eells demonstra é o seguinte: se *pr*(PC1|C1) > *pr*(PC1|C2), então a UCE (C1) > UCE (C2). É isto que sucede no caso 2, mas não nos casos 1 e 3. Ou seja, a utilidade condicional esperada do monocaixismo é maior do que a do bicaixismo, se a probabilidade de o previsor prever correctamente decisões monocaixistas (C1∧PC1) for maior do que a de ele se enganar e prever incorrectamente decisões monocaixistas (C2∧PC1); de outra maneira, se a probabilidade do monocaixista ser milionário for maior do que a do bicaixista 'enganar' o previsor. Isto faz, desde logo, algum sentido intuitivo: quanto menos fiável for a taxa de sucesso do previsor na previsão de decisões monocaixistas, mais sentido fará o agente decidir 'arriscar' e escolher as duas caixas. Consideremos a demonstração de Eells (1984a: 66-67):

UCE(C1) > UCE(C2), se, e somente se,

$$pr(\text{PC1}|\text{C1}).u(\text{PC1\&C1}) + pr(\text{PC2}|\text{C1}).u(\text{PC2\&C1}) > pr(\text{PC1}|\text{C2}).u(\text{PC1\&C2}) + pr(\text{PC2}|\text{C2}).u(\text{PC2}|\text{C2}).$$

Como a disjunção entre as probabilidades condicionais acima constituem partições,

$$pr(\text{PC1}|\text{C1}).u(\text{PC1\&C1}) + (1 - pr(\text{PC1}|\text{C1}).u(\text{PC2\&C1})) > pr(\text{PC1}|\text{C2}).u(\text{PC1\&C2}) + (1 - pr(\text{PC1}|\text{C2}).u(\text{PC2\&C2})).$$

Daqui segue-se (algebricamente, ver apêndice 1) que

$$pr(\text{PC1}|\text{C1}) > pr(\text{PC1}|\text{C2}) \times \left(\frac{u(PC1\&C2) - u(PC2\&C2)}{u(PC1\&C1) - u(PC2\&C1)}\right) + \left(\frac{u(\text{PC2\&C2}) - u(PC2\&C1)}{u(PC1\&C1) - u(PC2\&C1)}\right).$$

Os valores para a utilidade são os tradicionalmente apresentados no PN e são lineares com os valores monetários. Chamemos M ao valor que pode ser encontrado na caixa opaca, 1.000.000. Portanto, $u(PC1 \wedge C1) = M$; $u(PC1 \wedge C2) = M + 1000$; $u(PC2 \wedge C1) = 0$; e $u(PC2 \wedge C2) = 1000$. Substituindo os valores, temos que a desigualdade acima é equivalente a:

$$pr(PC1|C1) - pr(PC1|C2) > 1000/M$$

Gostaria apenas de complementar a prova com uma ilustração do resultado. Basta para isso lembrar que a taxa de sucesso que o previsor tem de exibir, de modo a que o princípio da maximização recomende C1, tem de ser superior a 0,5005. Considere-se a seguinte tabela com as respectivas probabilidades de sucesso do previsor:

	PC1	PC2
C1	0,5005	0,4995
C2	0,4995	0,5005

Podemos constatar que $pr(PC1|C1) - pr(PC1|C2)$, ou seja, 0,5005 – 0,4995 é igual a 0,001. Este valor corresponde, portanto ao lado direito da desigualdade acima, quando M é igual a 1,000,000. Para se obter uma recomendação de monocaixismo é, portanto, suficiente que a $pr(PC1|C1)$ seja maior do que $pr(PC1|C2)$. Além disso, por simetria da relevância probabilística, as seguintes desigualdades equivalentes (que podem ser facilmente observadas na tabela acima) também se verificam: $pr(PC1|C1) > pr(PC2|C1)$, $pr(PC2|C2) > (PC2|C1)$ e $pr(PC2|C2) > pr(PC1|C2)$.

Conclui-se que o argumento a partir da maximização não é falacioso e que depende da infalibilidade do previsor, desde que estejamos preparados para considerar 'infalível' uma percentagem de sucesso superior a 0,5005. O carácter paradoxal do problema mantém-se obviamente inalterado, seja a taxa de sucesso do previsor de 0,5005 ou 0,9. Ou seja, é possível considerar que este resultado é consequência de uma certa noção de infalibilidade – ainda que esta possa não ser superior a 0,5005 - pois a probabilidade das duas previsões correctas é sempre maior do que a probabilidade das duas previsões incorrectas.

5.2. Dominação

Até agora analisou-se o argumento favorável ao monocaixismo, a partir do princípio da maximização da utilidade condicional esperada, e verificou-se que o mesmo não é falacioso. Mas o que dizer da aplicação do princípio da dominação? Será o modo como Nozick coloca a questão, como um confronto entre estes dois princípios, realmente genuína?

Sabemos, desde logo, que o confronto entre dominação e maximização pode não se verificar quando os estados do mundo são probabilística e causalmente dependentes das acções, tal como no caso acima do desarmamento nuclear. Mas é precisamente para dar conta destas situações que o princípio condicional de Jeffrey é utilizado; ou seja, o princípio sobrepõe-se à dominação quando se verifica essa dependência. Não é, portanto, este tipo de casos que nos interessa.

Levi (1975: 168) apresentou também uma versão revista do PN, com o intuito de retirar do cenário a plausibilidade do apelo ao princípio da dominação. A única diferença no exemplo de Levi é a seguinte: quando o previsor se engana, o bicaixista ganha o milhão e mil, mas tem de indemnizar o previsor em 1500 Euros:

	PC1	PC2
C1	$M	$0
C2	$M – 500	$1000

Tal como na versão original, os estados continuam a ser probabilisticamente dependentes das acções e a aplicação do princípio condicional continua a recomendar C1. Mas, ao contrário da versão original, não existe aqui dominação. Contudo, nesta versão, os estados também se encontram desde o ínicio ‘fixos e determinados’, o que é a maneira de Nozick dizer que são causalmente independentes das acções; e como não há dependência causal, então o princípio de Jeffrey não se deveria aplicar, de acordo com bicaixistas. Disto se concluiria que, na versão revista, mantém-se o conflito de intuições, e o carácter paradoxal do problema, sem que se verifique um confronto entre princípios de racionalidade: os monocaixistas desejarão aplicar o princípio evidencial, os bicaixistas continuarão a defender o contrário, mas a possibilidade de recorrer à dominação já não se encontra ao seu dispor.

A meu ver, esta conclusão é altamente questionável. Ou seja, é extremamente duvidoso que a versão revista constitua o mesmo tipo de problema que o PN. Nozick (1969) constrói uma matriz com todos os problemas de decisão possíveis, colocando, de um lado, a existência ou não de dominação, e no outro o tipo de relação probabilística existente entre estados e acções:

	Existe dominação	Não existe dominação
Prob. condicionais diferentes. Há causação	(1) Maximizar Utilidade Esperada	(2) Maximizar Utilidade Esperada
Prob. condicionais idênticas. Não há causação.	(3) Escolher acção dominante ou maximizar	(4) Maximizar utilidade esperada
Prob. condicionais diferentes. Não há causação.	(5) Escolher acção dominante	(6) ?

Nos quatro primeiros casos não se verifica qualquer problema. Em (1) e (2), dado que as acções contribuem para causar os estados, essa causação deve reflectir-se no uso do princípio condicional. Em (3) e (4), sendo as probabilidades condicionais dos estados idênticas, não existem reservas em maximizar a utilidade, cujo resultado, neste caso, coincidirá com a acção dominante, se esta estiver disponível. Em (5) temos o PN e, de acordo com o argumento de Nozick, é racional escolher a acção dominante. O engenho de Levi foi precisamente gerar um exemplo que corresponde ao caso (6) e, relativamente ao caso (6), a solução de Nozick é bem mais elaborada do que aplicar de uma forma directa o princípio da maximização da utilidade condicional esperada.[45]

[45] Acerca de (6), Nozick diz o seguinte: 'Let *p1*,…, *pn* be the conditional probability distribution of action *A* over the *n* states; let *q1*,…, *qn* be the conditional probability distribution of *B* over the *n* states. A probability distribution *r1*,… *rn*, summing to one, is between *p1*,…, *pn* and *q1*,…, *qn* if, and only if, for each *i*, *ri*, is in the closed interval (*pi, qi*) or (*qi*,*pi*). (…) Now for a recommendation: if relative to each probability distribution between *p1*,…, *pn* and *qi,…,qn*, action A has a higher expected utility do *A* (…). If,

Assim sendo, não me parece indicado afirmar que (6), ou a versão revista de Levi, possa ser considerado o mesmo problema que o PN. E, se não for o mesmo problema, não me parece que ponha em causa a análise clássica de Nozick. Conclui-se, assim, que ambos os argumentos - maximização e dominação - não são falaciosos, e que o confronto entre os princípios empregues em cada um deles constitui uma caracterização adequada do PN. Existe, contudo, aparentemente, uma maneira de, no PN, retirar de cena a questão da dominação, bastando para isso tornar os estados independentes das acções. Que isto pode ser feito é uma certeza. Pode ser provado que qualquer partição finita de estados do mundo, que não são probabilísticamente independentes das acções, pode ser transformada numa outra em que se tornam incondicionados.[46] Para que seja legítimo adoptar este procedimento num problema de decisão, há que garantir que a nova partição faz tanto sentido quanto a anterior, ou seja, que a natureza do problema se mantém inalterada, o que nem sempre é possível.

No caso do PN, tal parece ser possível (Bar-Hillel e Margalit 1972: 299). Considere-se uma tabela de decisão em que E1 e E2 representam as mesmas acções que anteriormente, escolher a caixa opaca e escolher as duas caixas, e considerem-se os estados C e ¬C como representado, respectivamente, 'o previsor prevê correctamente' e 'o previsor prevê incorrectamente':

	C	¬C
E1	p	$1 - p$
	1 000 000	0
E2	p	$1 - p$
	1000	1 001 000

on the other hand, it is not the case that relative to each probability distribution between *p1,…,pn* and *q1,…, qn*, A has a higher expected utility than *B* (and it is not the case that relative to each, *B* has a higher expected utility than *A*), then we are faced with a problem of decision under constrained uncertainty (…) on which kind of problem there is not, so far as I know, agreement in the literature' (Nozick, 1975, pp. 124-125). E mais adiante: 'It may seem strange that for case (6) we bring in the probabilities in some way (even though they do not indicate an influence) whereas in case (5) we do not. This difference is only apparent, since we could bring in the probabilities in case (5) exactly the same way. The reason why we need not do this, and need only note that *A* dominates *B*, is that if *A* dominates *B*, then relative to each probability distribution (and therefore for each between the conditional ones established by the two actions) *A* has a higher utility than *B*' (1969: 125).

[46] Mencionado em Bar-Hillel e Margalit (1972: 295), podendo-se encontrar a prova em Krantz, Luce, Suppes e Tversky (1971).

p não é mais do que a taxa de sucesso de previsor, entre 0,5005 e 0,9. Neste caso, a correcção do previsor é probabilisticamente independente de qual a escolha feita; ou seja, a correcção da previsão não depende de o agente escolher uma ou outra acção. Para que isso acontecesse, teria de constar nos dados do problema que a probabilidade de o previsor adivinhar uma das acções seria maior ou menor do que a probabilidade de adivinhar a outra. Portanto, com *p* entre estes dois valores, o princípio da utilidade esperada recomenda, como sabemos, E1. Estaria, assim, encontrada a única estratégia racional para resolver o problema: maximizar a utilidade esperada e escolher E1.

Ora se é verdade que o problema, assim reformulado, deixaria de evidenciar uma aparente contradição entre dois princípios de decisão, devemos, contudo, questionar-nos se a sua natureza paradoxal deixa completamente de estar presente. Portanto, a natureza paradoxal depende do quão pervasivo continua a ser o argumento da dominação a favor do bicaixismo, mesmo após a reformulação acima. Será que esse argumento se torna realmente falacioso? Neste ponto, parece que temos de admitir que sim. Tendo em conta que os novos estados continuam a ser causalmente independentes das acções, o seguinte raciocínio deveria continuar a ser legítimo: esteja ou não a previsão correcta, escolher as duas caixas faz com que se fique sempre melhor do que escolhendo a caixa opaca. Contudo, esta resposta está errada, pois se a previsão estiver correcta, escolher a caixa opaca é certamente melhor. O monocaixista poderá dar-se por contente, pois o tipo de raciocínio que está na base da aplicação do princípio da dominação tem como resultado uma resposta que lhe é favorável: dada a verificação de qualquer um dos dois estados, existe agora alguma acção que seja sempre melhor do que a sua alternativa? Não só a resposta é negativa, como o elemento paradoxal do problema parece ter sido eliminado.

Contudo, mesmo aceitando-se que a reformulação de Bar-Hillel e Margalit não altera as características essenciais do problema, a formulação inicial continua a ser adequada, e a aplicação da dominação não pode depender contingentemente deste tipo de alterações. Afinal, se o princípio da dominação oferece a recomendação correcta numa representação adequada do problema, então a sua recomendação permanecerá correcta, independentemente da representação adoptada. Ou seja, os bicaixistas continuarão sempre a afirmar que, a partir do momento em que as acções não têm qualquer eficácia causal sobre os estados, o seu raciocínio permanece impecável.

Ainda assim, um primeiro argumento, que não se baseia apenas na mera proclamação de intuições, foi apresentado a favor de uma das soluções possíveis, nomeadamente, a

monocaixista. É certo que este não é um argumento particularmente forte; contudo, na contabilidade final dos prós e contras, não poderemos esquecer que ele existe. Mas os bicaixistas podem recolher algo bastante proveitoso deste argumento monocaixista. Mesmo que a reformulação do problema não permita eliminar do cenário a dominação, talvez seja útil e necessário encontrar uma outra interpretação do princípio da maximização da utilidade esperada, interpretação essa que concorra com a de Jeffrey, e que se possa aplicar a estes casos em que existe uma dependência probabilística dos estados do mundo relativamente às acções, mas em que não existe uma dependência causal. Esse novo princípio será introduzido de seguida.

6. A Solução de Stalnaker – Contrafactuais

6.1. Utilidade causal esperada

É para todos evidente que a capacidade de pensar de forma condicional faz parte do nosso equipamento mental mais básico, e mais evidente ainda no caso das nossas deliberações acerca do que fazer: 'Se eu agir de maneira X, quais serão as consequências da minha acção'? Não será, portanto, de estranhar que qualquer investigação acerca daquilo em que consiste agir racionalmente inclua uma discussão acerca do modo como os princípios da acção racional representam ou modelam a nossa forma de raciocinar condicionalmente.
Consideremos um exemplo bastante comum do raciocínio de tipo condicional. Um estudante pondera entre duas acções alternativas: estudar para o exame ou não estudar para o exame. Uma forma falaciosa de raciocínio seria esta: 'se passar no exame, estudar é tempo perdido; se não passar no exame, estudar é também tempo perdido; faça o que fizer, estudar é tempo perdido; logo, o melhor é não estudar'. O raciocínio é errado porque estudar aumenta a probabilidade de passar no exame. Ou seja, tal como no exemplo do armamento nuclear, as nossas deliberações devem ter em conta a influência das acções nas probabilidades dos estados do mundo que determinam as possíveis consequências dessas acções. Neste caso, a acção de estudar determina uma consequência cuja desejabilidade, para o agente, é composta, digamos, pelo esforço despendido e pelo benefício de passar.

A questão que se nos coloca é, então, a seguinte: de que modo deve a teoria da decisão interpretar a probabilidade das proposições condicionais utilizadas para calcular a utilidade esperada das acções? Ou seja, a probabilidade de um estado E, no caso de A ser realizada, *pr*(E, se A). O cálculo de probabilidades, ao dar conta de um tipo específico de probabilidades, as de tipo condicional, oferece, desde logo, uma sugestão:

$pr(\text{E, se A}) = pr(\text{E}|\text{A}) = pr(\text{E\&A})/P(\text{A})$, quando $pr(\text{A}) \neq 0$.

Como já foi visto, a teoria da decisão adopta esta concepção de probabilidade condicional para interpretar a probabilidade de um estado E, caso uma determinada acção A seja realizada. A fórmula que usa probabilidades condicionais consiste, no modelo de Jeffrey (1964), numa generalização de uma fórmula mais simples para a utilidade esperada com probabilidades incondicionais (ver § 5.2).

A segunda questão que se nos coloca é, então, a seguinte: será esta interpretação adequada para resolver todo e qualquer problema de decisão em que estejam envolvidas proposições de tipo condicional acerca da probabilidade de ocorrência de estados do mundo, dada a execução de uma ou outra acção? A análise do PN ajudar-nos-á a encontrar uma resposta.

Podemos identificar da seguinte maneira o traço específico do problema: existe uma correlação entre uma acção aparentemente inferior (de acordo com o princípio da dominação, a escolha da caixa opaca) e um estado do mundo bom (a previsão da escolha da caixa opaca), tal que essa acção não promove, ou não causa, esse estado. Por outras palavras, trata-se de uma acção que é auspiciosa, i.e., que augura uma consequência desejável, mas que não contribui de todo para a produzir.

Consideremos um exemplo que nos permite compreender melhor a estrutura do problema. A ideia fundamental é esta: um determinado evento pode constituir evidência para uma consequência desejável, sem que, no entanto, contribua para produzir essa consequência. Suponhamos que atiramos ao ar uma moeda que sabemos estar viciada (embora não saibamos para quê) e nos sai caras. Este evento constitui evidência que nos permite esperar que saia caras no próximo lançamento. Contudo, o primeiro evento, estando correlacionado com o segundo, não contribui em nada para o resultado deste. Os dois lançamentos constituem um exemplo de eventos probabilisticamente independentes. Assim, num problema de decisão, o que se pretende, quando calculamos a utilidade

esperada de uma acção, é saber o quão eficaz é uma acção na produção de uma consequência desejável, e não determinar apenas o quão auspiciosa é essa acção. Ou seja, pretende-se que uma acção produza uma consequência desejável, e não apenas o sinal, ou a evidência, de uma consequência desejável.

Por vezes estes dois aspectos coincidem. Será fácil de constatar que, no caso do estudante, a acção que consiste em estudar para o exame não só produz, com um determinado grau de probabilidade, a consequência pretendida, como é também auspiciosa. Contudo, o mesmo não parece acontecer no PN.

Robert Stalnaker (1980), numa carta breve a David Lewis (de 1972), propôs uma maneira de reconciliar os dois princípios de tomada de decisão no PN. A sua sugestão consiste em negar que a interpretação das proposições condicionais envolvidas em problemas de decisão possa ser dada, em todos os casos, através de probabilidades condicionais, como no caso da teoria de Jeffrey. Como se viu, a ideia de probabilidade subjectiva condicional não permite discriminar entre os casos em que se verifica uma mera correlação estatística entre a acção e um estado do mundo, e os casos em que, para além desta correlação, se verifica uma causação genuína. Ou seja, o princípio de Jeffrey não é sensível à ausência de causação.

A sugestão de Stalnaker consiste no seguinte: em certos casos, para efeitos do cálculo da utilidade de uma acção, deve-se utilizar não a probabilidade condicional de um estado, dada uma acção - $pr(E|A)$ - mas sim a probabilidade de uma proposição do tipo 'Se A fosse realizada, então seria o caso que E' - $pr(A \gg E)$. Este tipo de proposições presta-se a uma análise semelhante àquela que é utilizada para determinar o valor de verdade de contrafactuais.

As condicionais contrafactuais, ou conjuntivas, apresentam, de acordo com Stalnaker (1968), condições de verdade diferentes das condições de verdade das condicionais indicativas. Estamos habituados a oferecer como exemplos de contrafactuais proposições condicionais em que o antecedente é falso no mundo actual:

(c) Se Oswald não tivesse matado Kennedy, então outra pessoa o teria feito.

Sendo uma contrafactual verdadeira se, e somente se, no mundo possível mais próximo do actual (que torna a antecedente verdadeira) a consequente for verdadeira, então (c) é verdadeira se, e somente se, no mundo possível em que Oswald não matou Kennedy,

outra pessoa o fez, tudo o resto 'permanecendo igual'. Se partirmos do princípio de que Oswald matou realmente Kennedy, e que não existiu qualquer conspiração, então (c) é falsa.

Mas podemos também incluir no conjunto das contrafactuais proposições condicionais que expressam expectativas não realizadas em que o antecedente consiste numa acção: 'A ≫ E', a qual deve ser lida como 'Se eu fizesse A, então seria o caso que E'. Nestes casos, 'A ≫ E' pode ser verdadeira independentemente de eu acabar, ou não, por fazer A. De uma forma mais precisa (Gibbard e Harper 1978): seja A uma acção que eu posso decidir fazer, ou não, em *t*; seja o mundo-A um mundo possível idêntico ao actual antes de *t*, no qual eu decido fazer A em *t*, e que continua a obedecer às leis da física a partir de *t*; seja M(A) o mundo-A que, em *t*, é mais semelhante ao mundo actual em *t*. Assim, M(A) é um mundo possível que se desenrola após *t* de acordo com as leis da física e cujas condições iniciais em *t* são minimamente diferentes das condições no mundo actual em *t*, tal que A ≫ E é verdadeira em M(A). Temos, assim, definidas com maior precisão, as condições de verdade das condicionais relevantes empregues nos argumentos em §4.1.

As questões que se devem agora colocar são as seguintes: em que consiste, então, calcular a probabilidade de (A ≫ E), de modo a que isso seja diferente de calcular a probabilidade condicional (E|A)? Não será, afinal, a seguinte identidade uma verdade lógica:

(1) $pr(A \gg E) = pr(E|A)$?

Ao contrário do que efectivamente a nossa intuição nos pode levar a crer, David Lewis (1976) mostrou que, sob pressupostos bastantes plausíveis, (1) não é verdadeira acerca de quaisquer duas proposições arbitrárias (ver Apêndice 2). Contudo, dada uma certa condição, e em determinados contextos decisórios, pode ser demonstrado que (1) é verdadeira. O que nos importa é compreender a natureza dessa mesma condição e determinar o que acontece quando essa condição não se verifica. Isto permitir-nos-á distinguir os casos em que a interpretação condicional de Jeffrey se aplica, dos casos em que, de acordo com a sugestão de Stalnaker, apenas a interpretação causal oferece os resultados correctos.

Condição *X*. Para uma acção A e uma consequência C, a contrafactual A ≫ C é epistemicamente independente da acção A.

Ou seja,

$$pr(A \gg C|A) = pr(A \gg C).$$

Isto significa o seguinte: saber que estamos prestes a efectuar a acção A não altera a probabilidade que atribuímos à proposição segundo a qual se estivéssemos prestes a efectuar a acção A, C seguir-se-ia de A. Ou seja, a contrafactual é epistemicamente independente da acção.
É necessário, então, demonstrar que em certos casos, (1) é verdadeira. Para o efeito, considere-se:

Axioma 1. $(A \wedge (A \gg E)) \rightarrow E$.

Este axioma oferece-nos a garantia de aplicação de *modus ponens* às contrafactuais.[47]

Axioma 2. $(A \gg \neg E) \leftrightarrow \neg (A \gg E)$,

Este axioma resume as condições de verdade de Stalnaker: $A \gg E$ é falsa se, e somente se, E não for o caso em M(A). Em suma, $A \gg E$ é verdadeira se, e somente se, E for o caso em M(A) e $A \gg \neg E$ é verdadeira se, e somente se, $\neg E$ for o caso em M(A). Destes axiomas pode-se derivar a seguinte consequência:

Consequência 1. $A \rightarrow [(A \gg E) \leftrightarrow E]$

Portanto, dado que a Consequência 1 é uma verdade lógica, para quaisquer proposições P e Q (relacionadas por meio dela), a sua probabilidade será igual a 1,

$$pr(P \rightarrow [(P \gg Q) \leftrightarrow Q]) = 1.$$

Se *pr*(P) for maior do que 0, então

[47] Os axiomas, demonstrações e condições que se seguem encontram-se na sua totalidade expostos em Gibbard e Harper (1978: 155-159), embora por uma ordem ligeiramente diferente.

$pr([(P \gg Q) \leftrightarrow Q]|P) = 1.$

Logo,

$pr(P \gg Q|P) = pr(Q|P).$

A partir desta verdade geral temos que

$pr(A \gg E|A) = pr(E|A).$

Aplicando a Condição *X*, temos garantida verdade de

(1) $pr(A \gg E) = pr(E|A).$

De que nos serve, então, a verdade de (1)? Vejamos: para a acção *A* e o estado do mundo *E*, temos duas maneiras de calcular a utilidade esperada de A. Uma delas utiliza a probabilidade condicional,

$$\text{UCE}\ (A) = \sum_{(i=1)}^{n} pr(Ei|A) \times u(A \wedge Ei),$$

A outra utiliza a probabilidade da contrafactual. A esta utilidade passamos a designar como utilidade causal esperada (UCaE), pois, como veremos, será esta que permitirá seguir o nexo causal, ou a sua ausência, entre a realização da acção e a probabilidade do estado do mundo relevante:

$$\text{UCE}\ (A) = \sum_{(i=1)}^{n} pr(A \gg Ei) \times u(A \wedge Ei).$$

Podemos, assim, responder à questão. Quando a condição *X* se verifica - quando (1) é verdadeira - a UCE (A) = UCaE (A). Ou seja, se a probabilidade de uma contrafactual (A $\gg$ E) for idêntica à probabilidade da sua correspondente condicional (E|A), então

correlação e causação coincidem, e calcular a utilidade condicional esperada de A equivale a calcular a utilidade causal esperada de A.

Consideremos o seguinte exemplo de Jeffrey (1983: 3): sou convidado para jantar em casa de uns amigos e tenho de decidir se levo uma garrafa de vinho tinto ou de vinho branco, mas não sei se os meus amigos vão servir carne ou peixe. Eu prefiro vinho tinto com carne a vinho branco com peixe e, no caso de não acertar no vinho, prefiro vinho tinto com peixe a vinho branco com carne. Logo, a acção de comprar vinho tinto domina a acção de comprar vinho branco. Podemos também constatar que a condição *X* se aplica: saber que se efectuará a acção de comprar tinto não aumenta a probabilidade da condicional contrafactual 'Se eu comprasse tinto, então iria comer vinho tinto com carne'. Ou seja, a utilidade condicional esperada de comprar tinto e a utilidade causal esperada de comprar tinto são idênticas. De outra maneira, causação e correlação coincidem neste caso, mais precisamente, nenhuma delas se verifica.

Contudo, o nosso propósito é encontrar uma situação em que esta igualdade não se verifique, ou seja, em que a utilidade causal esperada de uma acção seja maior ou menor do que a utilidade condicional esperada dessa mesma acção. A estrutura de um tal problema de decisão terá, portanto, de ser idêntica à estrutura do PN, de modo a permitir-nos aplicar os resultados do primeiro à tentativa de resolução do segundo. Essa foi precisamente a proposta de Stalnaker: no PN, aquilo que devemos calcular é a utilidade causal esperada das acções disponíveis e não a sua utilidade condicional esperada.

Consideremos o seguinte exemplo (Gibbard e Harper 1978: 163). O Rei Salomão deseja a mulher de Uriah, Bathsheba. Contudo, a sua consciência diz-lhe que roubar Batsheba ao marido é injusto. Além disso, Salomão possui também certos conhecimentos de psicologia e ciência política, de acordo com os quais ele sabe que os reis podem ter um de dois tipos de personalidade: ou são carismáticos ou não são carismáticos, sendo que o grau de carisma de um rei é determinado pelos genes e pelas experiências de infância, não podendo ser alterado depois de atingida a idade adulta. Enquanto os reis carismáticos tendem a agir de forma justa, os não carismáticos tendem a agir de forma injusta. Enquanto as revoltas bem-sucedidas contra reis justos são raras, contra reis injustos elas são frequentes. Ora, por si só, acções injustas não provocam revoltas; os reis não carismáticos tendem a provocar revoltas devido à sua falta de carisma. Em suma, Salomão não sabe se é carismático, embora saiba que é injusto roubar a mulher de outro homem. Que deve, então, Salomão fazer de maneira a melhor satisfazer os seus interesses?

Seja B a acção de roubar Batsheba e R a irrupção de uma revolta. A maneira adequada de encarar o problema, de acordo com a análise acima, consiste em saber qual é a relação entre a *pr*(B ≫ R) e a *pr*(R|B). O que se pode constatar neste caso é que a Condição *X* falha, ou seja, a condicional contrafactual B ≫ R não é epistemicamente independente da acção B. Ou seja,

$pr(B \gg R|B) \neq pr(B \gg R)$.

Mais precisamente, a primeira é maior do que a segunda. Na medida em que Salomão sabe que roubar Bathseba não contribui para provocar uma revolta, ele atribui a mesma probabilidade a (B ≫ R) e a R, portanto a *pr*(B ≫ R) = *pr*(R). Logo, a *pr*(B ≫ R|B) = *pr*(R|B). Como já foi visto, se Salomão soubesse que B (que roubaria Batshbeba), então obteria evidência de que não seria um rei carismático e, como tal, de que seria propenso a gerar revoltas bem-sucedidas. Logo, a *pr*(R|B) > *pr*(R). Juntando tudo, temos o seguinte resultado:

$$pr(B \gg R|B) = pr\,(R|B) > pr(R) = pr(B \gg R).$$

Ou seja, neste caso (1) é falsa, pois

$$pr(B \gg R) < pr(R|B).$$

Se a Condição *X* não se verifica, podemos então concluir que, ao contrário do que sucedia acima, a UCE (B) ≠ UCaE (B). No caso da UCE temos o seguinte: Se Salomão souber que ¬B é o caso (que se abstém de roubar Bathseba), então terá uma razão para pensar que é carismático e, como tal, não-propenso a gerar revoltas. Assim, a prob (R|B) é maior do que a prob (R|¬B). Com efeito, para uma dada probabilidade, que não precisa de ser muito elevada (de haver uma revolta dado o roubo de Bathsheba), a UCE (¬B) será maior do que a UCE (B). A acção recomendada a Salomão pelo cálculo da utilidade condicional esperada consiste em abster-se de roubar Bathsheba.

O que recomenda, então, neste caso, o cálculo da UCaE? Salomão sabe que roubar Bathsheba não provocará uma revolta, apesar de saber que, caso o venha a fazer, tal lhe dará uma razão para acreditar que não é carismático. Assim, abster-se de roubar Bathsheba é uma acção auspiciosa, na medida em que lhe dará uma boa notícia quanto ao

seu carisma, i.e., uma evidência de um resultado desejável, sem que, no entanto, essa acção contribua para produzir esse mesmo resultado. Neste caso, a UCaE (B) é maior do que UCaE (¬B).

Em suma, verificou-se neste caso, em que a acção que consiste em roubar Bathseba não é epistemicamente independente da contrafactual (B ≫ R), que a probabilidade da contrafactual não é idêntica à probabilidade condicional R|B. Verificou-se, contudo, que a acção que consiste em roubar Bathsheba (B) é causalmente independente do estado do mundo que consiste numa revolta bem-sucedida (R) contra Salomão. Concluiu-se, então, que a coisa mais racional para Salomão fazer, com o intuito de maximizar a sua utilidade causal esperada, ou de melhor satisfazer os seus interesses, consiste em roubar Bathsheba. Se aceitarmos este resultado neste caso, e se a sua estrutura for idêntica à do PN, então teremos de aceitar o resultado análogo no caso do PN.

Seja novamente C1 a acção que consiste em escolher a caixa opaca; C2 a acção que consiste em escolher ambas as caixas; PC1 o estado em que o previsor prevê correctamente a acção de escolher a caixa opaca e PC2 o estado em que o previsor prevê correctamente a acção de escolher as duas caixas. Dada a elevada fiabilidade do previsor, se eu souber que estou prestes a escolher a caixa opaca, então a probabilidade que atribuo à contrafactual 'Se eu escolhesse a caixa opaca, então o previsor prevê-lo-ia', condicionada à escolha caixa opaca, é maior do que a probabilidade incondicional dessa mesma contrafactual:

$$pr(\text{C1} \gg \text{PC1}|\text{C1}) > pr(\text{C1} \gg \text{PC1}).$$

Ou seja, a condição *X* não se verifica. Logo, tal como no exemplo anterior, a acção de escolher a caixa opaca não é epistemicamente independente do estado do mundo que consiste em o previsor prever correctamente a minha escolha. Do que se conclui que $pr(\text{C1} \gg \text{PC1}) \neq pr(\text{PC1}|\text{C1})$. E, portanto, sabemos também que no caso do PN, UCE (C1) ≠ UCaE (C1). Também como no exemplo anterior, dado que a acção de escolher a caixa opaca é causalmente independente da previsão, na medida em que não é eficaz na sua produção, a $pr(\text{C1} \gg \text{PC1}) = pr(\text{PC1})$. Em §4.1 calculámos a UCE de C1 e C2. Vamos agora calcular a UCaE de ambas. Suponhamos que ganhar mil tem uma utilidade de valor 10, ganhar um milhão valor 1000, ganhar um milhão e mil valor 101, e que não ganhar nada tem valor 0.

Como PC1 é causalmente independente das acções, a *pr*(C1 ≫ PC1) = *pr*(C2 ≫ PC1) = μ.

UCaE (C1) = *pr*(C1 ≫ PC1) × 1,000,000 + *pr*(C1 ≫ PC2) × 0 =
= μ × 100 + (1 - μ) × 0 = 100μ
UCaE (C2) = *pr*(C2 ≫ PC1) × 1,001,000 + *pr*(C2 ≫ PC2) × 1000 =
= μ × 101 + (1 - μ) × 10 = 91μ + 10

Assim, UCaE (C2) – UCaE (C1) = 10 – 9μ, e como μ ≤ 1, esta diferença é sempre positiva. Logo, para qualquer valor de μ, UCaE (C2) é sempre maior do que a UCaE (C1). Ou seja, o cálculo da UCaE recomenda que se escolha ficar com ambas as caixas.

Podemos, agora, reformular, à luz do que vem a ser dito, aquilo que se designou como o cerne do problema. Trata-se não apenas de um conflito entre o princípio da dominação e o princípio da utilidade esperada, mas sim de um conflito entre dois tipos de utilidade esperada, ou entre duas maneiras de calcular a utilidade de uma acção, sendo que uma delas se encontra, de facto, em conflito com o princípio da dominação. O que se constatou foi que, em determinados casos, apenas uma dessas maneiras de calcular a utilidade de uma acção consegue seguir o rasto, digamos, da relação de causalidade, ou da sua ausência, entre as acções disponíveis e os estados do mundo que determinam as consequências dessas acções. Vimos que o PN era um desses casos.

Concluindo, se a nossa perpectiva acerca do que é racional fazer em casos estruturalmente idênticos ao PN for a de que devemos aplicar o princípio da maximização da utilidade causal esperada (UCaE), em que se utiliza a probabilidade de contrafactuais em vez de probabilidades condicionais, então o mesmo terá de valer para o caso do PN. A solução de Stalnaker apresenta-se, deste modo, a uma luz extremamente apelativa, sugerindo, talvez, que os desafios à consistência da teoria bayesiana da decisão, colocados pelo PN, poderão ser ultrapassados mediante a elaboração sistemática de uma teoria da utilidade causal esperada.

6.2. A teoria de Savage

A conclusão delineada em §6.1 diz respeito à comparação que se estabeleceu entre a aplicação ao PN da teoria da decisão segundo o modelo de Jeffrey, por um lado, e um segundo modelo a que se chamou 'teoria causal da decisão', de Gibbard e Harper, baseado na sugestão de Stalnaker, por outro. Como os estados do mundo no PN não são probabilisticamente independentes das acções, a aplicação do cálculo da utilidade condicional esperada foi tida ao início como natural e imediata.

Mas temos ainda um outro modelo à nossa disposição, o de Savage (1954). A principal diferença deste relativamente ao modelo de Jeffrey é a seguinte: apenas se pode aplicá-lo quando os estados do mundo são probabilisticamente independentes das acções. Contudo, após a discussão acima, sabemos agora que existem duas maneiras de interpretar a noção de independência. A primeira pode ser denominada como independência evidencial e verifica-se quando a escolha de uma acção não constitui sequer indício ou evidência para se acreditar que um ou outro estado do mundo virá a verificar-se. A segunda pode ser denominada como independência causal e verifica-se quando a escolha de uma acção não contribui causalmente para a obtenção de qualquer um dos estados do mundo relevantes. Estas duas noções não são coextensionais, pois como vimos, no caso do PN, pode existir independência causal sem que exista independência evidencial.

A relevância destas considerações torna-se notória quando descobrimos que existe um método à nossa disposição para tornar os estados do mundo num problema de decisão probabilisticamente independentes das acções, de modo a podermos aplicar o modelo de Savage a esse mesmo problema. Esse método, baseado numa proposta de Itzhak Gilboa (2009), e desenvolvido por Bruno Jacinto (2011), consiste em transformar os estados - eventos expressos através de proposições categóricas - em conjunções de proposições condicionais do tipo contrafactual. Uma versão deste método surge pela primeira vez em Gibbard e Harper (1978), segundo os mesmos de acordo com uma sugestão do próprio Jeffrey. Contudo, a variante de Gilboa é utilizada para defender a posição monocaixista no PN. Assim, o objectivo desta secção consiste em explorar o referido método e verificar se as conclusões de Gilboa são suficientes para contrabalançar, ou pelo menos suspender, as conclusões favoráveis à solução bicaixista que a aplicação da teoria causal da decisão alcançou.

O modelo de Savage caracteriza-se principalmente pelo modo como define formalmente as acções enquanto funções de estados do mundo *E* para consequências *C*, f_A: $E \rightarrow C$. Ou seja, cada $a \in A$ estabelece uma correspondência biunívoca entre uma consequência *C* e um estado *E*, *A(E) = C* $[A, E]$. Savage é menos claro no que respeita a uma definição dos elementos dos conjuntos *E* e *C*, embora não se possa afirmar que estes permanecem como noções primitivas, dado que é possível dizer algo de relevante acerca deles. Um estado seria 'uma descrição do mundo que não deixa qualquer aspecto relevante por descrever' (Savage 1972: 13), nomeadamente, tendo em conta o requisito da independência probabilística, 'uma descrição das condições do mundo sobre as quais o decisor não exerce controlo directo' (Joyce 1999: 57). Uma consequência seria 'qualquer coisa que pode acontecer a uma pessoa' (Savage 1972: 13), nomeadamente, uma descrição suficientemente detalhada de 'todas as coisas desejáveis e indesejáveis que uma combinação de *A* e *E* pode trazer' (Joyce 1999: 52).

O cálculo da utilidade esperada de uma acção no modelo de Savage (UES, utilidade esperada de Savage) é feito utilizando-se probabilidades incondicionais de estados do mundo, e as utilidades correspondem aos valores de uma função f_A que toma como argumentos os elementos de *E*:

$$\text{UES}\ (a) = \sum_{(i=1)}^{n} \quad pr(ei) \times u(a(ei)).$$

Uma das consequências do modelo de Savage é a seguinte: se uma acção *a* domina uma acção *b*, então, dada a independência probabilística dos estados relativamente às acções, a UES $(a) >$ UES(b). Ou seja, ao contrário do que acontecia no modelo de Jeffrey, não existe aqui qualquer restrição sobre a aplicação do princípio da maximização da utilidade esperada quando, num problema de decisão, é possível raciocinar a partir da dominação. Considere-se, por exemplo, um problema de decisão com apenas dois estados do mundo, com probabilidades *p* e $(1 - p)$. Se uma acção dominar a outra, então, independentemente do valor de *p*, a acção dominante terá sempre uma maior utilidade incondicional esperada. Se, por outro lado, nenhum das acções dominar a outra, então o valor de *p* é determinante para o cálculo da utilidade. No modelo de Jeffrey, se a acção escolhida contribuir para determinar a probabilidade de cada um dos estados, então, mesmo que exista uma acção

dominante, a sua utilidade condicional esperada pode ser menor do que a da acção dominada.

De modo a aplicarmos a teoria de Savage, consideremos o exemplo do desarmamento nuclear de §4.1. As acções à disposição do governo de um determinado país são as seguintes: manter ou aumentar o seu arsenal nuclear (*A*), e desarmar (*D*). Os estados do mundo relevantes são: a verificação de um estado de guerra (*G*) e a verificação de um estado de paz ($\neg G$). Os estados não são probabilisticamente independentes das acções e, portanto, o raciocínio a partir da dominação é falacioso. Os *payoffs* encontram-se entre parêntesis em cada uma das células da matriz:

	G	$\neg G$
A	$G,\neg D$ (0)	$\neg G,\neg D$ (9)
D	G,D (1)	$\neg G,D$ (10)

O método referido, para tornar os estados probabilisticamente independentes das acções, consiste em substituir os dois estados do mundo da matriz inicial pelos quatro novos estados seguintes:

(0,0) – Faça *A* ou *D*, não haverá guerra.
(0,1) – *A* não provocará uma guerra, mas *D* sim.
(1,0) – *A* provocará uma guerra, mas *D* não.
(1,1) – Faça *A* ou *D*, haverá uma guerra.

Cada novo estado passa, assim, a ser expresso através de uma conjunção de contrafactuais. Por exemplo,

(0,0) – $[(A \gg \neg G) \wedge (D \gg \neg G)]$, deve ser lido como 'se eu fizesse *A*, então não haveria guerra e se eu fizesse *D*, então não haveria guerra'. Ou então,

(1,0) – $[(A \gg \neg G) \wedge (D \gg G)]$, deve ser lido como 'se eu fizesse *A*, então não haveria guerra e se eu fizesse D, haveria guerra'.

Com estes novos estados pode ser construída uma nova matriz para representar o problema de decisão:

	(0,0)	(0,1)	(1,0)	(1,1)
A	$\neg G,\neg D$ (9)	$\neg G,\neg D$ (9)	$G,\neg D$ (0)	$G,\neg D$ (0)
D	$\neg G,D$ (10)	G,D (1)	$\neg G,D$ (10)	G,D (1)

Aplicando a fórmula de Savage, temos, por exemplo,

UES $(A) = pr(0,0) \times u(A(0,0)) + pr(0,1) \times u(A(0,1)) + pr(1,0) \times u(A(1,0)) + pr(1,1) \times u(A(1,1))$.

Os lados direitos de cada um dos produtos especificam uma consequência de *A* e correspondem a cada uma das células da matriz na fila de cima. E, como se pode constatar, na matriz, as entradas de (0,0) e (0,1) por um lado, e de (1,0) e (1,1) por outro, são idênticas entre si. Logo,

$(A, (0,0)) = (A, (0,1)) = \neg G,\neg D$
$(A, (1,0)) = (A, (1,1)) = G,\neg D$

Como (0,0) e (0,1) são mutuamente exclusivos, assim como (1,0) e (1,1), temos que

$$\begin{aligned}
&\text{UES}(A) = pr((0,0) \vee (0,1)) \times u(\neg G,\neg D) + pr((1,0) \vee (1,1)) \times u(G,\neg D) = \\
&= pr[((A \gg \neg G) \wedge (D \gg \neg G)) \vee ((A \gg \neg G) \wedge (D \gg G))] \times u(\neg G,\neg D) + \\
&\quad pr[((A \gg G) \wedge (D \gg \neg G))] \vee ((A \gg G) \wedge (D \gg G))] \times u(G,\neg D) = \\
&= pr[(A \gg \neg G) \wedge ((D \gg \neg G) \vee (D \gg G))] \times u(\neg G,\neg D) + \\
&\quad prob\,[(A \gg G) \wedge ((D \gg \neg G) \vee (D \gg G))] \times u(G,\neg D).
\end{aligned}$$

Aplicando o Axioma 1 (ver §6.1), o qual consiste numa definição formal das condições de verdade para as contrafactuais, $(D \gg \neg G) \leftrightarrow \neg(D \gg G)$, obtemos o seguinte resultado,

$$\text{UES}(A) = pr(A \gg \neg G) \times u(\neg G,\neg D) + pr(A \gg G) \times u(G,\neg D).$$

Para sabermos que tipo de utilidade esperada é a de Savage, agora que cumprimos o requisito para a aplicação da sua teoria, temos de colocar uma questão: que relações de independência se verificam neste caso entre os estados e as acções? Essa independência não pode ser causal, pois, como sabemos, armar contribui para não haver guerra (embora os estados sejam probabilisticamente independentes das acções, pode-se afirmar que essa dependência causal se encontra reflectida na contrafactual).

Da perspectiva de um terceiro, mais informado do que o próprio agente, é perfeitamente possível um estado ser causalmente dependente de uma acção e, no entanto, ser evidencialmente independente, bastando para isso que o agente desconheça uma ou outra das consequências da sua acção. Isto não acontece efectivamente neste caso.[48]

Podemos também observar que a evidência oferecida pela suposição de que agiríamos no sentido de armar, não contribui para modificar a probabilidade que atribuímos à contrafactual que expressa a relação de causalidade entre a acção de armar e o estado de ausência de guerra. Podemos, assim, concluir que o tipo de independência a que chamámos 'evidencial' é equivalente àquilo que é expresso pela Condição *X*. A primeira foi definida como aquela que se verifica quando a escolha de uma acção não constitui indício ou evidência para se acreditar que um ou outro estado do mundo virá a verificar-se. Já a Condição *X*, designada por 'independência epistémica', diz respeito àqueles casos em que saber que estamos prestes a efectuar uma acção não altera a probabilidade que atribuímos à proposição segundo a qual se estivéssemos prestes a efectuar essa acção, uma determinada consequência seguir-se-ia.[49]

Da discussão acima sabemos que quando a Condição *X* se verifica, então a seguinte identidade é verdadeira,

$$pr(A \gg \neg G) = pr(\neg G|\mathrm{A}).$$

[48] Independência evidencial e indepedência causal encontram-se relacionadas (ou não) de várias maneiras: podem verificar-se ambas, como quando tenho de escolher entre dois caminhos alternativos para evitar o trânsito; no PN existe independência causal, mas não independência evidencial; no caso do armamento nuclear, verifica-se tanto dependência causal, como dependência evidencial. E, como referi, não conhecer alguma consequência possível da minha acção pode implicar, do ponto de vista de um terceiro, dependência causal, embora, do ponto de vista do agente, independência evidencial. Contudo, se a relação entre estes dois conceitos for apenas considerada do ponto-de-vista do agente, parece-me que a existência de independência evidencial se torna numa condição suficiente da independência causal.

[49] Esta equivalência entre a independência evidencial e a Condição *X* (a da independência epistémica) é tornada clara pela consideração da identidade já referida, $pr(\mathrm{A} \gg \mathrm{G}|\mathrm{A}) = pr(\mathrm{A} \gg \mathrm{G})$, a qual constitui uma definição das duas. Ou seja, $pr(\mathrm{Q}|\mathrm{P}) = pr(\mathrm{Q})$.

Isto é facilmente comprovável: como, neste caso, a relação de causalidade acompanha a relação evidencial, a $pr(A \gg \neg G)$ nunca pode ser menor do que a $pr(\neg G|\mathrm{A})$. Era precisamente isso que acontecia no exemplo de Salomão e Batsheba. Da mesma maneira, como causação parece implicar sempre correlação, a primeira destas probabilidades também não pode ser maior do que a segunda. O mesmo argumento mostrar-nos-á que $pr(D \gg G) = prob\ (\mathrm{G|D})$.

De tudo isto se se conclui que

$$\begin{aligned}\mathrm{UES}\ (A) &= pr(A \gg \neg G) \times u(\neg G, \neg D) + pr(A \gg G) \times u(G, \neg D) = \\ &= pr(\neg G|\mathrm{A}) \times u(\neg G, \neg D) + pr(G|\mathrm{A}) \times u(G, \neg D) = \\ &= \mathrm{UCE}\ (A).\end{aligned}$$

Quando a probabilidade de uma contrafactual é idêntica à probabilidade condicional correspondente, então a utilidade esperada de Savage - a utilidade esperada incondicional (após a reformulação dos estados) - é idêntica à utilidade condicional do modelo de Jeffrey. Aplicando o mesmo raciocínio, chegar-se também à conclusão de que a UCE (*D*) é idêntica à UES (*D*). Isto acontecerá sempre que as contrafactuais envolvidas forem epistemicamente/evidencialmente independentes das acções.[50] Logo, independentemente das probabilidades atribuídas aos novos estados do mundo, a acção recomendada pela teoria de Savage será sempre a mesma que é recomendada pela teoria de Jeffrey.

A introdução da teoria da Savage no contexto da discussão pode também agora ser melhor compreendida: se a reformulação dos estados do mundo for feita através do recurso a contrafactuais, então o princípio causal pode ser interpretado como uma extensão natural do princípio de Savage, na medida em que ambos acomodam não só os casos em que os estados são causalmente dependentes das acções, como no exemplo acima, mas também os casos em que, tal como no PN, os estados são causalmente independentes das acções, mas não evidencialmente independentes das mesmas. Isto torna-se claro quando se considera o modo como Gilboa define os estados do mundo.

Gilboa sugere que, de modo a tornar os estados independentes das acções, estes não devem ser tomados como simples dados do problema, mas sim como funções de acções para consequências. Esta proposta pode ser entendida como uma crítica ao modo como os estados são definidos na teoria de Savage apenas enquanto 'estados', como algo que

[50] Bruno Jacinto (2011: 12-14) apresenta uma demonstração do caso geral.

se encontra imune à influência que sobre eles podem ter as escolhas do decisor. Segundo Gilboa:

'Na medida em que os estados passam a ser funções de acções para consequências, os mesmos podem agora ser tomados como independentes destas últimas: a dependência da consequência em relação à acção passa a estar reflectida no argumento da função e não na própria função' (2009: 114).

O problema que resulta desta sugestão é imediatamente claro para um bayesiano e consiste em saber como é possível atribuírem-se probabilidades subjectivas a funções. Ou seja, quando os estados são entendidos como eventos (Savage) ou proposições (Jeffrey) não é difícil conceber que as probabilidades que lhes são atribuídas consistem em graus de crença na ocorrência de eventos ou na verdade de proposições. Mas se não soubermos exactamente como devem ser entendidos os estados, ou quando estes são entendidos como objectos matemáticos, não encontraremos qualquer fundamento para lhes atribuirmos probabilidades. Gilboa dá-nos algumas pistas ao considerar a sua proposta como uma extensão da ideia segundo a qual os estados são funções de verdade para proposições, identificando Gibbard e Harper como os primeiros a adoptar esta sugestão. Sigo aqui a sugestão de Bruno Jacinto (2011: 9-10), e tenho vindo a interpretar essas funções como proposições condicionais do tipo contrafactual: 'se eu fizesse X, então Y seria o caso'; ou seja, para qualquer $a \in A$, $(a \gg f_E(a))$. Em suma, a escolha de uma ou outra acção tornaria verdadeiras ou falsas as proposições contrafactuais que definem os estados (ao fornecer-lhes os antecedentes/argumentos da função). Neste ponto podemos considerar uma conclusão importante que já estava implícita na discussão em §6.1, e que agora se torna clara: se fizermos equivaler a teoria causal de Gibbard e Harper ao método de Gilboa/Jacinto (aplicado à teoria de Savage), então isso significa que a teoria causal apresenta obviamente os resultados correctos quando a teoria evidencial de Jeffrey também o faz.

Gilboa irá argumentar que a modelação do PN na teoria de Savage, utilizando para o efeito o método acima exemplificado, resulta numa recomendação clara da posição monocaixista. Aplicando este modelo e seguindo o exemplo anterior, em vez dos dois estados iniciais - 'o previsor previu a acção que consiste em escolher a caixa opaca'(*PC1*),

e ‘o previsor previu a acção que consiste em escolher as duas caixas’(*PC2*) - passamos a ter os quatro seguintes estados:

(0,0) – Faça eu o que fizer, o previsor coloca 1 milhão na caixa opaca.
(0,1) – O previsor coloca 1 milhão na caixa opaca, apenas se eu escolher a caixa opaca.
(1,0) – O previsor coloca 1 milhão na caixa opaca, apenas se eu escolher as duas caixas.
(1,1) – Faça eu o que fizer, a caixa opaca estará vazia.

Por exemplo, (0,0) – ($C1 \gg 1{,}000{,}000$) ∧ ($C2 \gg 1{,}001{,}000$), deve ser lida como ‘se eu escolhesse a caixa opaca, então receberia 1,000,000 e se eu escolhesse as duas caixas, então receberia 1,001,000’.
A matriz do PN após a reformulação dos estados passa a ser a seguinte:

	(0,0)	(0,1)	(1,0)	(1,1)
C1	1,000,000	1,000,000	0	0
C2	1,001,000	1,000	1,001,000	1,000

Gilboa apresenta os seguintes dados do problema: o previsor não é infalível, mas após mil sujeitos terem sido submetidos à tomada de decisão, os quinhentos que escolheram a caixa opaca ganharam 1,000,000 e os quinhentos que escolheram as duas caixas ganharam 1,000. Estes dados produzem claramente uma conclusão: quando o PN é representado pela nova matriz, o único estado que é compatível com as observações é (0,1), i.e., o milhão encontra-se na caixa opaca, apenas quando o agente escolhe a caixa opaca. Disto resulta que devemos atribuir uma probabilidade muitíssimo elevada a (0,1). Sabendo-se que os estados são evidencialmente independentes das acções, pois saber que se fará *C1* ou *C2* não alterará a probabilidade que se atribui aos quatro estados do mundo, e dado que estes são mutuamente exclusivos, $pr(0,1) = 1$ e $pr[(0,0) \vee (1,0) \vee (1,1)] = 0$. Portanto,

$$pr\,(0,1) = pr[(C1 \gg 1{,}000{,}000) \wedge (C2 \gg 1{,}000)] =$$
$$= pr[(C1 \gg 1{,}000{,}000)|(C2 \gg 1{,}000)] \times prob\,(C2 \gg 1{,}000).$$

Pressupondo a independência evidencial das contrafactuais, temos que

$$pr(0,1) = [(1{,}000{,}000|C1)|(1{,}000|C2)] \times pr(1{,}000|C2) =$$
$$= pr(1{,}000{,}000|C1) \times pr(1{,}000|C2) =$$
$$= \frac{pr[(1{,}000{,}000\ ) \wedge (C1)]}{pr(C1)} \times \frac{pr[(1{,}000) \wedge (C2)]}{pr(C2)} =$$
$$= \frac{500/1000}{500/1000} \times \frac{500/1000}{500/1000} = 1.$$

Sabendo-se que $pr(0,1) = 1$, basta-nos olhar para as duas células da nova matriz do PN que correspondem a (0,1) e constatar que UES (*C1*) > UES (*C2*).

Analisaram-se duas maneiras através das quais o cálculo da probabilidade condicional dos estados pode ser substituído pelo cálculo da probabilidade de contrafactuais, de modo a alcançar-se uma solução para o PN. Contudo, nos dois exemplos de aplicação de contrafactuais ao PN, de Gibbard e Harper (segundo a teoria causal) e Gilboa (segundo a reformulação de Savage), os resultados consistiram em recomendações diferentes. Apesar dos argumentos serem ambos persuasivos, apenas uma das conclusões pode estar correcta. Uma das hipóteses a considerar, parece-me, é que Gilboa, ao contrário de Gibbard e Harper, tenha feito uma interpretação diferente das contrafactuais envolvidas. No caso destes últimos, essas contrafactuais detectariam a ausência de uma relação de dependência causal entre as acções e as previsões, o que não só corresponde a uma descrição adequada da situação, mas também, segundo os bicaixistas, à prescrição da acção correcta. Já Gilboa, por seu lado, ao considerar como unicamente relevante para a tomada de decisão o estado do mundo descrito em (0,1) - ($C1 \gg 1{,}000{,}000$) $\wedge$ ($C2 \gg 1{,}000$) - parece estar a tomar as relações de dependência contrafactual aí presentes como reveladoras de uma verdadeira relação de causalidade (à luz de uma determinada semântica de contrafactuais (ver §7.1)). Esta hipótese, à primeira vista surpreendente e contra-intuitiva, é precisamente aquela que iremos explorar daqui em diante.

7. Contrafactuais (2) – Duas soluções para a vagueza

7.1. Semântica dos mundos possíveis

A esperança colocada na proposta de Stalnaker era a de que, por si só, esta pudesse trazer consigo uma solução clara para o PN. Em contraste com a teoria evidencial de Jeffrey, a proposta de Stalnaker deu origem a uma teoria da decisão que, de acordo com essa esperança, respondia à questão situada no âmago do problema: como tornar a teoria insensível à correlação sem causação, de modo a não recomendar a acção errada? Pelo facto de a nova teoria ter a capacidade de seguir os traços das relações de causalidade verificadas entre acções e estados, a mesma foi denominada como causal.

Contudo, da análise de §6 verificou-se que, afinal, a reconstrução dos estados como proposições condicionais de tipo contrafactual não conduz a uma única solução, mas que algo mais será necessário para justificar uma ou outra das conclusões: bicaixista, no caso da proposta de Gibbard e Harper, e monocaixista no caso da proposta de Gilboa. Enquanto os primeiros dão primazia à noção de independência causal dos estados relativamente às acções, o segundo parece focar-se na selecção dos únicos estados do mundo cuja obtenção é uma possibilidade efectiva, tendo em conta os dados do problema. Mas, como ambos analisam os estados como proposições contrafactuais, surgiu a hipótese de que essas contrafactuais pudessem estar a ser interpretadas de maneira diferente, algo que não pode deixar de parecer normal tendo em conta a vagueza inerente à análise de contrafactuais. Esta é a hipótese que se pretende explorar; daí ser necessário passar algum tempo com este tipo peculiar de proposições.

A ideia de que a estrutura lógica da noção de causalidade pode ser esclarecida através da relação de dependência condicional entre eventos não é de todo recente. Ela surge explicitamente em Stuart Mill (1919), onde uma causa consiste, precisamente, numa condição suficiente do seu efeito. A análise de Mill adequa-se à tentativa de encontrar um modelo de explicação que sirva a teoria científica e que dê conta dos processos inferenciais utilizados para determinar as causas que permitem compreender um determinado fenómeno (ver Zilhão 2010a, 2012). Nessa medida, e com naturalidade, é na filosofia da ciência que a análise das condicionais se tornará num dos temas e problemas mais importantes ao longo do século XX.

O entusiasmo provocado pelo surgimento da lógica de Frege, dada a sua capacidade de regimentar a linguagem científica, foi sentido pelos empiristas lógicos, os quais encontraram um tratamento simples e aparentemente claro para as condicionais. Segundo Dorothy Edgington (1995: 236), a esperança consistia, para os empiristas lógicos, em utilizar a lógica de Frege para 'fazer pela ciência aquilo que ele tinha feito pela matemática'. O cepticismo desses filósofos, na linha de Hume, relativamente à realidade física da causalidade, tornava apelativo o tratamento de certos poderes causais como propriedades disposicionais, observáveis, das coisas. O problema era que na análise deste tipo de propriedades, tais como a 'solubilidade' ou o 'magnetismo', se fazia sentir a presença de um outro tipo de condicional que não se deixava regimentar tão facilmente:

'A substância X é solúvel' se, e somente se, 'se fosse misturada em água, então dissolver-se-ia'.
'A barra de metal Y é magnética se, e somente se, 'se fosse colocada limalha de ferro junto dela, então esta pegar-se-ia às suas extremidades'.

O que a condicional material de Frege não podia explicar era a razão pela qual não seria possível determinar se uma substância seria solúvel, ou se uma barra seria magnética, caso a substância não fosse misturada em água ou não fosse colocada limalha de ferro junto à barra, num momento específico do tempo. O problema reside obviamente no facto de a condicional de Frege continuar a ser verdadeira, mesmo quando o antecedente é falso. Ainda que a noção de causalidade seja analisada, à maneira de Hume, como instanciação de regularidades, neste caso de regularidades associadas ao comportamento de certos objectos, um tratamento científico das propriedades disposicionais desses objectos tem de ser passível de generalização universal. Ou seja, há que distinguir entre meras regularidades acidentais e regularidades que se encontram relacionadas com leis da natureza, o que pode ser extremamente difícil através da mera análise de propriedades disposicionais. Foi talvez por esta razão que as primeiras teorias de contrafactuais surgiram associadas à noção de lei, esperando-se que uma compreensão das condições de verdade das contrafactuais pudesse simultaneamente elucidar este conceito, na medida em que, ao contrário das meras regularidades arbitrárias, as leis parecem ter implicações contrafactuais: 'se a constante gravitacional fosse diferente, …'; 'se o objecto não estivesse sujeito ao atrito, …; 'se a concentração de gás na cozinha fosse maior, …'.

A tentativa de definir as contrafactuais como proposições condicionais governadas por leis da natureza foi efectuada por Nelson Goodman (1947), sugerindo condições de verdade para frases como a seguinte: 'Se o fósforo tivesse sido riscado, então ter-se-ia acendido':

Uma condicional contrafactual A ≫ C é verdadeira se, e somente se, existe uma conjunção de verdades T, incluindo uma lei da natureza [e satisfazendo uma condição X], tal que A∧T implica C'.[51]

Se esta definição funcionasse como era esperado, talvez tivesse sido possível a Goodman sugerir uma análise do conceito de causalidade através da análise de contrafactuais. Mas as coisas não são tão simples como a definição dá a entender. Desde logo poderá ser difícil discriminar entre aquilo que é realmente necessário para satisfazer a condição X – como o fósforo estar seco, a superfície em que é raspado ser suficientemente rugosa, existir um comburente adequado – e aquilo que não é necessário ou irrelevante – ter sido uma pessoa a riscar o fósforo ou ter sido uma máquina a fazê-lo.[52]

Além disso, existem inúmeras contrafactuais que estamos dispostos a aceitar como verdadeiras e que, tanto quanto sabemos, a sua verdade não depende da identificação de qualquer lei da natureza, como por exemplo: 'Se eles tivessem saído, o carro não estaria na garagem' ou 'Se eles estivessem em casa, as luzes estariam acesas'.

Por outro lado, uma teoria acerca de contrafactuais deverá ser capaz de fazer uma distinção entre duas proposições, às quais estamos dispostos a atribuir condições de verdade bastante diferentes. Mas, de acordo com a definição de Goodman, não parece existir uma lei da natureza que torne a primeira das seguintes proposições verdadeira por comparação com a segunda:

---

[51] Segue-se desta análise que as condições de verdade das contrafactuais especificam um argumento nomológico-dedutivo escondido, do qual a contrafactual em questão seria uma espécie de corolário (ver §2.1).

[52] Goodman sugere uma outra definição para determinar que factos necessários são esses, descrevendo-os como compatíveis (*cotenable*) com a suposição da verdade do antecedente: *C* é *cotenable* com *A* se, e somente se, não é caso que se *A* fosse verdadeira, *B* não o seria'. Esta definição aplica-se seguramente à 'secura do fósforo' e à 'existência de um comburente'. Contudo, entramos numa circularidade indesejada: é necessária uma definição de *cotenable* para definir as contrafactuais, mas essa definição é ela própria expressa através de uma proposição contrafactual.

‘Se atirasse esta moeda (não-viciada) ao ar 10 vezes, então sairia caras pelo menos uma vez’.
‘Se atirasse esta moeda (não-viciada) ao ar mil vezes, então ela transformar-se-ia em ouro’.

De acordo com a teoria de Goodman, não existe propriamente uma lei que torne a primeira verdadeira e a segunda falsa, a menos que a relação de causalidade seja interpretada como uma relação de dependência probabilística, algo que é perfeitamente razoável.
Com o surgimento de uma semântica para a lógica modal (Kripke 1963) o conceito de mundo possível veio dar uma nova oportunidade à tentativa de formulação de condições de verdade para contrafactuais. A relação entre este conceito e as proposições em causa torna-se clara no momento em que nos deparamos com o exemplo utilizado por David Lewis nas primeiras linhas de *Counterfactuals* (1973a: 1):

‘Se os cangurus não tivessem cauda, então tombariam’. Ou seja, num mundo possível que fosse em tudo semelhante ao actual e no qual a única diferença seria os cangurus não terem cauda (alterando-se tudo o que fosse estritamente necessário para tornar isso possível), então os cangurus tombariam’.

As condições de verdade seguem-se de imediato: 1) Se A é falsa em todos os mundos possíveis, então A ≫ C é vacuamente verdadeira; 2) A ≫ C é (não-vacuamente) verdadeira se, e somente se, algum mundo A∧C é mais próximo do mundo actual do que um mundo A ∧ ¬C. Uma contrafactual é verdadeira se, e somente se, um mundo em que A é verdadeira e C é verdadeira for mais semelhante ao mundo actual do que um mundo em que A é verdadeira e C é falsa. Se for necessário efectuar mais alterações ao mundo actual, nomeadamente alterações às leis da natureza, para se obter um mundo em que os cangurus não têm cauda e mesmo assim não tombam, do que para obter um mundo em que os cangurus não têm cauda e tombam, então este segundo mundo é mais semelhante ao actual do que o primeiro, o que torna a contrafactual de Lewis verdadeira.
A noção de semelhança entre mundos possíveis é, contudo, naturalmente vaga, sendo muitas vezes difícil ou talvez impossível determinar que mundos possíveis são mais semelhantes ao mundo actual do que outros. O projecto de Lewis consiste em utilizar uma

noção extremamente vaga, a de semelhança entre mundos possíveis, para analisar uma outra também ela vaga, a de dependência contrafactual. Além disso, parece que temos uma maior compreensão intuitiva da segunda do que da primeira. Do mesmo modo, parece que também temos uma maior compreensão intuitiva do conceito de causalidade, do que daquele que pretende explicá-lo, o de dependência contrafactual.[53] Se a noção de causalidade for analisada através da noção de dependência contrafactual, conviria então que esta fosse tornada suficientemente clara para o efeito. Se for possível mostrar que as condições de verdade de Lewis levam vantagem relativamente à análise de Goodman, no que respeita a avaliar a verdade ou a falsidade de certas proposições problemáticas, então teremos uma base de trabalho segura para a análise da causalidade. Neste ponto surgem, todavia, boas e más notícias.

Proposições contrafactuais acerca do comportamento das pessoas parecem resultar facilmente verdadeiras (ou falsas) na definição de Lewis, dado que esta não faz um apelo explícito a leis da natureza para fazer com que o consequente de uma contrafactual se siga do antecedente; ou seja, basta termos um conhecimento razoável acerca dos hábitos de alguém para aceitarmos que 'se essa pessoa estivesse em casa (a uma determinada hora), então as luzes estariam acesas'. A pessoa em causa ainda não chegou a casa, mas é razoável aceitarmos que um mundo possível em que a pessoa está em casa com as luzes apagadas – ou porque está na cama doente ou porque se deu uma avaria no sistema eléctrico – é menos semelhante ao mundo actual do que um mundo possível em que a pessoa estaria em casa (à mesma hora) com as luzes acesas. Estas são as boas notícias.

As más resultam da possibilidade deixada em aberto pela definição de Lewis. Não se segue desta que exista sempre um único mundo possível mais semelhante ao mundo actual, podendo dar-se o caso de existirem vários que partilham o primeiro lugar da semelhança. Nesses casos, Lewis exige que a contrafactual seja verdadeira em todos esses mundos possíveis. Existem, por exemplo, várias maneiras de tornar verdadeira a antecedente de: 'se atirasse esta moeda (não-viciada) ao ar 10 vezes, então sairia cara pelo menos uma vez'. Qual dessas maneiras é mais semelhante ao mundo actual? Talvez o consequente seja verdadeiro de acordo com a esmagadora maioria das maneiras mais

[53] Stalnaker (1968) apresentou condições de verdade com base em mundos possíveis não apenas para contrafactuais, mas também para condicionais indicativas. Stalnaker, contudo, não apresenta qualquer análise do conceito de semelhança entre mundos possíveis e, nessa medida, a sua proposta não pressupõe que se possa determinar a verdade de contrafactuais sem recorrer a argumentos que envolvem eles próprios o recurso a contrafactuais.

próximas do mundo actual de tornar verdadeira a antecedente, mas parece não existir uma garantia de que todas elas garantam a verdade da contrafactual.
Um dos exemplos mais conhecidos que coloca em relevo este problema da definição de Lewis consiste no seguinte par de proposições:

'Se Bizet e Verdi fossem compatriotas, então seriam franceses'.
'Se Bizet e Verdi fossem compatriotas, então seriam italianos'.

Neste caso, temos duas proposições empatadas no primeiro lugar para mundo possível mais próximo, mas em que nenhuma nos soa como definitivamente verdadeira. Se existirem vários mundos igualmente próximos do actual em que Bizet e Verdi são franceses nuns e italianos noutros, então, de acordo com a definição de Lewis, ambas as proposições são falsas.[54]
Seria fugir demasiado ao tema encontrar razões para pronunciar um veredicto final acerca da análise de Lewis. Além disso, temos a sensação de que na esmagadora maioria dos casos, principalmente naqueles em que temos maior informação, os nossos juízos acerca da semelhança entre eventos, objectos ou mundos possíveis são relativamente claros. Mas talvez seja demasiado apressado colocar a definição de Lewis ao mesmo nível que a de Goodman, concluindo que a primeira não apresenta afinal melhores resultados que a segunda na determinação do valor de verdade de proposições que ajuizamos como sendo claramente verdadeiras.
A vagueza da noção de semelhança entre mundos possíveis talvez não seja prejudicial à definição de Lewis, dependendo do tipo de vagueza que estejamos a considerar. Por um lado, a vagueza em causa poderá estar nas próprias coisas que estão a ser comparadas e, nessa medida, poderá não haver facto acerca do qual se aplique a expressão 'mais próximo de' ou 'mais semelhante a'. Por outro lado, se essa vagueza for interpretada como sendo do tipo epistémico, então a correcção da definição de Lewis não sai afectada: existe realmente algo a que se aplicam os termos acima, mas nós simplesmente não sabemos identificá-lo. Ou seja, existe realmente um mundo mais próximo do actual em que Bizet e Verdi são compatriotas, embora não tenhamos ainda informação suficiente para saber qual deles é ou é-nos impossível, da nossa perspectiva, vir a descobri-lo. Tendo em conta

[54] De acordo com Stalnaker (1968) estas proposições não têm valor de verdade, o que talvez esteja mais de acordo com a nossa intuição.

o passado (possivelmente o das famílias de Bizet e Verdi) e as leis da natureza, poderá ser necessário, segundo o próprio termo de Lewis, um 'milagre' mais pequeno para tornar a antecedente e uma das consequentes simultaneamente verdadeiras do que a mesma antecedente e a outra das consequentes. Embora a exploração e justificação desta hipótese não caiba nos limites da presente investigação, o que importa reter é o seguinte: se a noção de dependência contrafactual vai ser utilizada para analisar a de causalidade, então qualquer vagueza que invada a primeira irá igualmente invadir a segunda, o que, na análise das contrafactuais do PN, poderá mostrar-se decisivo. Afinal, um dos lemas do bicaixismo é o de que o monocaixismo apenas faz sentido na suposição absurda de que a causalidade funciona para trás no tempo.

Na secção sete do *Ensaio Sobre o Entendimento Humano*, Hume definiu causalidade de duas maneiras:

'(…) podemos definir uma causa como um objecto seguindo-se de outro, sendo que todos os objectos semelhantes ao primeiro são seguidos por objectos semelhantes ao segundo. Ou, por outras palavras, se o primeiro objecto não tivesse existido, o segundo também não'. (Citado por Lewis 1973b: 556)

A primeira parte da definição corresponde à análise de regularidades, em que os objectos são definidos como causas ou efeitos através de uma descrição dos mesmos, sem que para o efeito se utilize a noção de semelhança entre mundos possíveis. A segunda parte é aquela em que Lewis se inspirou para conceber a sua própria análise, passando a causalidade a ser entendida como uma relação de dependência contrafactual. Se um evento é causa de outro, então a verdade da proposição segundo a qual o primeiro ocorre é necessária para a verdade da proposição segundo a qual o segundo também ocorre. Ou seja, *e* depende causalmente de *c* se, e somente se, se *c* não tivesse ocorrido, então *e* também não.

O aspecto mais importante desta definição, que será determinante para os nossos propósitos, consiste na assimetria da relação de dependência contrafactual, a qual permite explicar a assimetria temporal da causalidade, em que as causas objectivamente precedem sempre os seus efeitos. Aceita-se facilmente a ideia de que o futuro depende contrafactualmente do presente, que as acções por nós escolhidas determinam, pelo menos em parte, o modo como o futuro vai ser: se eu fizesse algo diferente do que

realmente farei, o futuro seria também diferente do que realmente será. Mas se o futuro se encontra em aberto, o passado é fixo e inalterável, o que fazemos agora não tem qualquer influência no passado, i.e., o passado é contrafactualmente independente do presente.

O enorme poder explicativo da análise da noção de causação em termos de dependência contrafactual pode não ser inteiramente óbvio, principalmente tendo em conta o recurso à semântica de mundos possíveis. Por exemplo, o aumento da temperatura do corpo é a causa das alterações na leitura do termómetro usado para medir essa temperatura, daí aceitarmos facilmente 'se ele tivesse febre, então o mercúrio do termómetro teria subido'. Contudo, temos também a tentação de aceitar a inversa: 'se o mercúrio do termómetro tivesse subido, então ele teria febre'. Este é um daqueles exemplos em que parece verificar-se que a causa não teria ocorrido, se o efeito não tivesse também ocorrido, o que parece ir contra o sentido único da relação de causalidade. A questão que deve ser colocada é seguinte: que vantagens apresenta a análise de contrafactuais, relativamente à análise de regularidades, na explicação deste aparente círculo causal? Esta última, consistindo na observação da conjunção regular de dois eventos, terá por vezes dificuldades em distinguir conceptualmente o efeito da causa, acrescendo o facto de existirem muitos casos em que o efeito é observado antes da causa. A análise de Lewis, por seu lado, garante-nos que a segunda contrafactual é falsa. No mundo possível mais próximo em que o mercúrio do termómetro sobe e tudo o resto permanece igual (eu não estou doente), eu não tenho febre.[55]

Para compreendermos melhor o exemplo, consideremos o mundo actual como aquele em que eu não estou doente, não tenho sintomas de febre e que, de acordo com a leitura do termómetro, a temperatura do meu corpo é normal. No mundo possível mais próximo do actual em que tudo se mantém igual e o mercúrio do termómetro sobe, o que estamos dispostos a sacrificar? Os efeitos normais, tal como a febre, que a doença provocaria em mim ou o funcionamento normal do termómetro? Se não estivermos seguros da resposta, consideremos um mundo actual em que se faz sentir uma temperatura sufocante e em que temos de determinar a verdade da contrafactual 'Se o termómetro indicasse uma

[55] Como veremos mais adiante, as contrafactuais podem, dependendo das circunstâncias, ter uma interpretação que autoriza a verdade destas contrafactuais invertidas. Consideremos um brilhante professor de matemática a resolver no quadro uma equação e a seguinte contrafactual: 'Se ele se tivesse enganado, então estaria distraído'. No mundo possível mais próximo do actual em que o professor continua a ser brilhante e se engana é perfeitamente possível que a melhor explicação para ele se ter enganado seja estar distraído.

temperatura menor, então estaria mais frio'. O que estaremos dispostos a sacrificar? As leis da natureza, e as circunstâncias que fazem com que alguém sinta calor, ou o bom funcionamento do termómetro? É razoável pensar que o primeiro sacrifício implica um desvio maior em relação ao mundo actual do que o desvio implicado pela avaria de um instrumento sensível, feito por mãos humanas e susceptível ao erro, o que tornaria esta contrafactual falsa.

A teoria de Lewis oferece também uma solução para o problema que consiste em distinguir conceptualmente uma causa genuína de um mero epifenómeno. Considere-se que *c* é simultaneamente causa de *e* e de um epifenómeno *f*, o qual não é um efeito de *e*. Acontece, porém, que *c* causa primeiro *f* e só depois *e*. A análise de regularidades terá certamente dificuldade em distinguir conceptualmente a causa genuína do epifenómeno, pois ambos ocorrem igualmente, e de forma repetida, em conjunção com o efeito, de onde resulta que aparentemente tenhamos uma relação de dependência causal entre *f* e *e*, contrariando o facto de que *f* não é causa de *e*, mas apenas um epifenómeno de *c*. A solução consiste em negar a contrafactual que expressa essa suposta dependência: 'Se *f* não tivesse ocorrido, então *e* também não'. Ou seja, eliminar *e* juntamente com *f* implica um desvio maior em relação à actualidade do que eliminar apenas *f* e preservar *e*. É mais semelhante ao mundo actual um mundo em que *c* causa *e*, mas não *f*, do que um mundo em que *c* se verifica, mas que não causa nem *f*, nem *e*. Assim, é preferível que algumas circunstâncias desfavoráveis, ou a mínima alteração possível das leis do mundo actual, tenham impedido *f* de ocorrer, e manter *c* e o seu efeito *e*, do que eliminar *f* e *e*, e manter apenas *c*. Do mesmo modo, faz mais sentido manter a temperatura do corpo estável e ter o termómetro estragado, do que ter febre e manter a fiabilidade do termómetro.

A assimetria da dependência contrafactual ajuda a determinar os critérios de avaliação da semelhança geral entre mundos possíveis que Lewis propõe. Segue-se que estes critérios constituem uma solução para a vagueza na interpretação de contrafactuais, ou, sabemos agora, da vagueza na identificação dos mundos possíveis mais próximos do actual. Estes critérios constituem o que Lewis designa como 'a resolução standard para a vagueza' e são os seguintes:

1) É de primeira importância evitar grandes violações das leis da natureza que impliquem um leque abrangente e variado de consequências.

2) É de segunda importância maximizar a região do espaço-tempo em que prevalece uma coincidência perfeita de factos particulares.
3) É de terceira importância evitar pequenas violações das leis da natureza, ainda que simples e localizadas.
4) É de pouca ou nenhuma importância assegurar uma semelhança aproximada entre factos particulares, mesmo em matérias de grande relevância para nós. (Lewis 1979: 47-8)

No início falámos de duas interpretações para as proposições contrafactuais. Os critérios acima apresentados oferecem-nos a chave para uma dessas interpretações, aquela que se encontra de acordo com a assimetria da dependência contrafactual, ou seja, aquela que respeita a assimetria temporal da causalidade. A proposta de Stalnaker apresenta-se agora com uma naturalidade que a princípio poderá ter escapado. Se o objectivo consistia em modelar um problema de decisão evitando que a mera correlação probabilística fosse confundida com causação, então a relação de dependência contrafactual garantiria que qualquer relação de dependência causal dos estados relativamente às acções seria detectada pelas contrafactuais que definem esses estados. Mais importante, se alguma das proposições em causa violasse a assimetria da dependência causal, a teoria seria também sensível a essa violação.
Convém ainda notar que, se a antecedente A de uma contrafactual for verdadeira no mundo actual, então o mundo actual é o mundo-A mais próximo, daí que uma contrafactual verdadeira implique uma condicional indicativa verdadeira com o mesmo antecedente e o mesmo consequente.

7.2. Interpretação retroactiva

Neste ponto, o que nos cabe determinar é se o argumento monocaixista dá como verdadeira alguma contrafactual que seja falsa de acordo com a resolução standard da vagueza. Pressupondo que a resolução standard para a vagueza da semelhança entre mundos possíveis é a correcta, consideremos novamente algumas das premissas do argumento monocaixista (ver §4.1):

1. Se eu escolhesse as duas caixas, então o previsor prevê-lo-ia.
4. Se eu escolhesse a caixa opaca, então o previsor prevê-lo-ia.

Estas duas proposições não podem ser ambas verdadeiras sob a resolução standard da vagueza. O previsor já efectuou a sua previsão, assim faça eu o que fizer, tal não irá alterar o conteúdo da caixa opaca. Mais precisamente, a verdade do antecedente, aquilo que decidimos fazer, não terá qualquer influência no valor de verdade do consequente. Se o previsor previu que iremos escolher as duas caixas, então, ao escolhermos as duas caixas, tornamos 1 verdadeira. Mas, nesse caso, 4 é falsa, pois se o previsor previu que vamos escolher as duas caixas e eu escolho apenas a caixa opaca, o antecedente de 4 é verdadeiro (nesse mundo possível), mas o consequente tem de ser falso.
Ou seja, no mundo possível mais próximo em que escolhemos as duas caixas o *estado-actual* da caixa opaca mantém-se inalterado e no mundo possível mais próximo em que escolhemos a caixa opaca o *estado-actual* da caixa opaca mantém-se igualmente inalterado. Para que as contrafactuais 1 e 4 sejam ambas verdadeiras sob a resolução standard, para que o previsor acerte seja qual for a minha escolha, num dos mundos possíveis que torne um ou outro dos antecedentes verdadeiros tem de ocorrer um *pequeno milagre* (ou violação das leis da natureza), para que o conteúdo da caixa opaca corresponda sempre à minha escolha, o que consiste claramente numa violação dos critérios de Lewis - 1) impedir grandes violações das leis da natureza, e 2) maximização da semelhança de factos no espaço-tempo. Essa grande violação consistiria na alteração, no momento da escolha, do conteúdo da caixa opaca. De acordo com esta interpretação standard das contrafactuais, se uma acção *x* não tiver eficácia causal na produção de *y*, então

$pr(x \gg y) = pr(\neg x \gg y)$.

Terence Horgan (1981) chamou a atenção para uma outra resolução da vagueza que faz com que as premissas 1 e 4 do argumento monocaixista sejam ambas verdadeiras. A resolução sugerida por Horgan favorece aquilo a que Lewis (1979a) chamou de 'backtracking argument' – algo como 'argumento retroactivo'. Segundo Lewis, um argumento deste tipo é válido se, e somente se, for admissível uma violação da assimetria da dependência contrafactual, dependendo essa admissibilidade das circunstâncias da

situação em causa. Essas circunstâncias dizem respeito àquilo que conhecemos acerca do caso particular que está a ser analisado e daquilo que nele deve contar como uma lei da natureza. Estes argumentos retroactivos são particularmente apropriados quando as leis em causa, neste caso menos inflexíveis, dizem respeito ao carácter e comportamento dos indivíduos. Estes são argumentos em que o passado depende contrafactualmente do presente, ou seja, casos em que 'se o presente fosse diferente, então o passado também teria sido'. Considere-se o exemplo do próprio Lewis. No mundo actual Maria e António tiveram no dia anterior uma enorme discussão. Maria é extremamente orgulhosa e tende a amuar durante dias a fio. António é não só orgulhoso, mas também vingativo. Assim, a contrafactual seguinte parece ser verdadeira: 'Se Maria tivesse hoje pedido ajuda a António, então não teriam tido uma discussão no dia anterior'. O exemplo fala por si: tendo em conta o carácter de Maria e António - e neste caso são as 'leis' que regulam esse carácter que devem contar como leis da natureza - no mundo possível mais próximo do actual, em que Maria pediu ajuda a António, eles não teriam discutido no dia anterior.

No PN, a ideia de Horgan é a de que a correcção da previsão 'é um parâmetro de semelhança mais importante do que a maximização da região do espaço-tempo em que prevalece uma perfeita semelhança entre factos particulares' (Horgan 1981: 164). Neste caso, é *como se* o presente, a acção que eu decido efectuar, tivesse influência no modo como o passado é. Ao contrário do que normalmente sucede, se o presente fosse diferente, então o passado também teria sido. Assim, é *como se* a minha acção tivesse o poder de transformar a previsão, de modo a que esta corresponda sempre à minha escolha. Ou seja, o passado torna-se contrafactualmente dependente do presente.

Aceitando uma interpretação retroactiva das contrafactuais, 1 e 4 podem ser ambas verdadeiras. No mundo possível mais próximo em que escolho ambas as caixas, o previsor já o terá previsto; no mundo possível mais próximo em que escolho apenas a caixa opaca, o previsor já o terá previsto. Isto significa que

pr(caixa opaca $\gg$ previsor prevê opaca) $>$ pr(2 caixas $\gg$ previsor prevê opaca).

Consideremos agora as premissas decisivas do argumento bicaixista e averiguemos, como acima, o que resulta das duas interpretações possíveis, standard e retroactiva:

2’. Se a caixa opaca contiver 1 Milhão (material), então eu ganharia 1 Milhão e Mil, se escolhesse (contrafactual) ambas as caixas.
5’. Se a caixa opaca estiver vazia (material), então eu ganharia 0, se escolhesse (contrafactual) apenas a caixa opaca.
6’. Das duas uma (base disjuntiva): *ou* ganharia 1 Milhão e Mil, se escolhesse ambas as caixas, e 1 Milhão, se escolhesse apenas a caixa opaca, *ou* ganharia Mil, se escolhesse ambas as caixas, e 0, se escolhesse apenas a caixa opaca.

Podemos verificar que 2’, 5’ e 6’ são todas verdadeiras sob a resolução standard da vagueza. 2’, por exemplo, é verdadeira se, e somente se, ou a caixa opaca não contém 1 milhão (a antecendente da condicional material é falsa) ou eu decido escolher ambas as caixas. A minha acção (que determina a verdade do antecedente da contrafactual) não tem qualquer influência no conteúdo da caixa opaca, i.e., no valor de verdade do consequente. A garantia de que esse consequente é verdadeiro é-nos oferecida pelo antecedente da condicional material. Por outras palavras, de acordo com a análise de Horgan, existe uma coincidência perfeita entre os factos particulares do passado e os do mundo que torno actual através da minha escolha, consistindo esses factos no conteúdo inalterado da caixa opaca. Por razões análogas, 5’ também é verdadeira, e a conjunção de 2’ e 5’ torna 6’ verdadeira.
Por outo lado, 2’ e 5’ não podem ser verdadeiras sob a resolução retroactiva da vagueza. Consideremos, por exemplo, 2’. Supondo que a caixa opaca contém o milhão (ou que o antecedente é verdadeiro), e que a correcção do previsor é um critério de semelhança entre mundos possíveis mais importante do que a previsão actual, então, se eu escolhesse ambas as caixas, receberia apenas 1000 e não 1 milhão e mil (ocorreu um *pequeno milagre* e o conteúdo da caixa opaca alterou-se); logo, o consequente da condicional material é falso. Pela mesma ordem de razões, 5’ será falsa quando a caixa opaca estiver vazia. Segue-se daqui que 6’, a conclusão do argumento bicaixista, é também falsa.
Esta dupla possibilidade de análise das contrafactuais pode, a meu ver, servir para interpretar as razões subjacentes aos argumentos de Gibbard e Harper e de Gilboa (mesmo que estes não as tenham assim entendido). Apesar de ambas as propostas promoverem o cálculo da probabilidade de contrafactuais, cada uma delas depende directamente de uma análise distinta das contrafactuais envolvidas. Os primeiros, focando-se na noção de independência causal, é como se estivessem a favorecer uma interpretação das

contrafactuais de acordo com a resolução standard da vagueza. Já o argumento de Gilboa, por seu lado, depende de uma resolução retroactiva da vagueza, em que apenas dois mundos possíveis se poderão tornar actuais.

Para tornar as coisas mais claras, relembremos algumas conclusões. De acordo com o argumento de Gibbard e Harper, a acção de escolher a caixa opaca, dada a enorme fiabilidade do previsor, não é epistemicamente/evidencialmente independente do estado do mundo que consiste em o previsor prever correctamente a minha escolha, ou seja,

$pr(C1 \gg PC1|C1) > (C1 \gg PC1)$.

Além disso, dado que a previsão é causalmente independente da acção,

$pr(C1 \gg PC1) = prob\ (PC1)$.

Do mesmo modo, dado que *PC1* é causalmente independente das acções,

$pr(C2 \gg PC1) = pr(C1 \gg PC1)$.

Esta última identidade reproduz, em termos probabilísticos (substituindo-se o símbolo de identidade pelo de conjunção), o conteúdo expresso pela premissa 6' do argumento bicaixista, nomeadamente, o primeiro disjunto: 'receberia 1.001.000 se escolhesse as duas caixas e 1 milhão se escolhesse apenas a caixa opaca'.[56]

Gilboa, por outro lado, ao focar-se na enorme fiabilidade do previsor, é como se estivesse a dar total primazia, na resolução da vagueza, à 'lei da natureza' que consiste nessa mesma fiabilidade, interpretando de maneira retroactiva as contrafactuais 1 e 4 do argumento monocaixista. Considerem-se os seguintes quatro mundos possíveis:

w1 – escolho as duas caixas e encontro um milhão na caixa opaca
w2 – escolho a caixa opaca e encontro um milhão na caixa opaca
w3 – escolho as duas caixas e a caixa opaca está vazia
w4 – escolho a caixa opaca e a caixa opaca esta vazia.

[56] Da mesma maneira, a ideia de que $pr(C2 \gg PC2) = pr(C1 \gg PC2)$ corresponde ao segundo disjunto da premissa 6':'receberia 1000 se escolhesse as duas caixas e nada se escolhesse apenas a caixa opaca'.

A perspectiva atribuída a Gilboa é a de que w2 e w3 são os únicos mundos possíveis consistentes com os dados do problema. w2 e w3 correspondem, respectivamente, às premissas 4 e 1 do argumento monocaixista. A atribuição de probabilidades de Gilboa compromete-o, precisamente, com essa leitura retroactiva das contrafactuais:

w1 – $pr(C2 \gg PC1) = 0$
w2 – $pr(C1 \gg PC1) = 1$
w3 – $pr(C2 \gg PC2) = 1$
w4 – $pr(C1 \gg PC2) = 0$

Ou seja, $pr(C1 \gg PC1) > pr(C1 \gg PC2)$ e $pr(C2 \gg PC2) > pr(C2 \gg PC1)$.
Por seu lado, os defensores do argumento bicaixista têm de admitir que a sua solução para o problema depende de uma consideração séria da possibilidade efectiva de actualizar w1ou w4. Ou seja, que a nossa escolha deve colocar-nos numa posição em que *podemos* ganhar 1.001.000 ou 1.000, e não apenas 1.000.000 ou zero. O argumento monocaixista diz-nos simplesmente que w1 é extremamente improvável em comparação com w3, e w4 é extremamente improvável em comparação com w2.
Mas qual das resoluções da vagueza devemos adoptar como sendo pragmaticamente a mais apropriada? As circunstâncias do problema favorecem qual das resoluções? Esta é questão que se impõe neste ponto da controvérsia, dela dependendo, possivelmente, a solução para o PN. Uma das maneiras de tentar encontrar a reposta poderá consistir em passar a discussão para um meta-nível, o qual tem como objecto as premissas originais de ambos os argumentos. Um bom argumento neste meta-nível seria um argumento que estabelecesse qual das resoluções da vagueza é mais adequada, sem que, para isso, recorresse a premissas que envolvessem proposições contrafactuais. Esta é a possibilidade explorada por Horgan, da qual trataremos em seguida.

8. Contrafactuais (3) – Uma solução pragmaticamente apropriada para a vagueza

8.1. Meta-argumentos: Eells vs Horgan

Os dois meta-argumentos favoráveis a cada uma das soluções para a vagueza das contrafactuais são-nos já familiares. Vale a pena, contudo, expô-los da forma mais clara possível.

O meta-argumento favorável à solução retroactiva e, a *fortiori*, ao argumento monocaixista baseia-se no seguinte raciocínio: w2 é o mundo-(caixa opaca) mais próximo e w3 é o mundo-(2 caixas) mais próximo. São, portanto, os únicos mundos possíveis que têm possibilidade de se tornarem actuais. Logo, o meta-argumento monocaixista depende do seguinte princípio:

(M_o) Tenho virtualmente a certeza, independentemente de quaisquer crenças que eu tenha acerca da probabilidade de vir a escolher a caixa opaca ou as duas caixas, que w2 ou w3 se tornarão actuais. (Horgan, 1981: 167).

Pode-se constatar que este meta-argumento não faz uso de quaisquer contrafactuais, nem de quaisquer premissas que façam parte do argumento monocaixista original, constituindo uma justificação independente para se adoptar a solução retroactiva para a vagueza. A questão que se coloca é a de saber se os defensores da solução standard e, a *fortiori*, do argumento bicaixista, têm também à sua disposição um meta-argumento deste tipo que constitua uma justificação independente para a sua solução.

O meta-argumento bicaixista baseia-se no seguinte raciocínio: no PN, as minhas acções não têm influência causal sobre o conteúdo da caixa opaca. Mas, se estiver um milhão na caixa opaca, eu tenho à minha disposição actualizar w1 ou w2; por outro lado, se a caixa opaca estiver vazia, eu tenho à minha disposição actualizar w3 ou w4; ou seja, eu devo encarar como o mundo possível mais próximo aquele que eu actualizarei através da minha escolha. Logo, o meta-argumento bicaixista depende do seguinte princípio:

(M_2) Das duas uma: eu actualizaria w1, se escolhesse as duas caixas, e actualizaria w2, se escolhesse a caixa opaca, *ou* actualizaria w3, se escolhesse as duas caixas, e actualizaria w4, se escolhesse a caixa opaca. (Horgan 1981: 166).

O problema deste último meta-argumento, apesar de inquestionavelmente válido sob a resolução standard, é o de que não existe uma justificação independente para se optar por esta solução, pois (M_2) não é mais do que uma reformulação da premissa fundamental 6' do argumento original bicaixista:

6'. Das duas uma (base disjuntiva): ou ganharia 1 Milhão e Mil, se escolhesse ambas as caixas, e 1 Milhão, se escolhesse apenas a caixa opaca, *ou* ganharia Mil, se escolhesse ambas as caixas, e 0, se escolhesse apenas a caixa opaca.

O meta-argumento bicaixista vê-se assim infectado por uma circularidade viciosa. A sua conclusão é a seguinte: se (M_2) for verdadeiro, então a solução standard deve ser adoptada. Mas (M_2) não é mais do que uma reformulação da conclusão do argumento ao nível-objecto: se 6' for verdadeira, então o agente deve escolher as duas caixas. Segundo Horgan, o bicaixista não tem, portanto, uma justificação independente para adoptar a resolução standard. A conclusão de Horgan é a de que, pelo facto de o único meta-argumento cogente ser o meta-argumento monocaixista, a discussão acerca de qual é a solução pragmaticamente adequada para a vagueza fica, assim, encerrada de maneira favorável à resolução retroactiva.

Ellery Eells (1985) formulou dois argumentos importantes com o objectivo de refutar as conclusões de Horgan. O primeiro desses argumentos pretende demonstrar que (M_o) não favorece a resolução retroactiva, na medida em que os defensores da resolução standard podem igualmente aceitar a verdade de (M_o). Isto levará a uma reformulação de (M_o), de modo a que este princípio passa a exibir o mesmo tipo de circularidade que (M_2) exibe. Este primeiro argumento de Eells parte da seguinte reformulação do princípio subjacente à resolução standard:

Ou a) w1 é o mundo-(2 caixas) mais próximo $\wedge$ w2 é o mundo-(caixa opaca) mais próximo, *ou* b) w3 é o mundo-(2 caixas) mais próximo $\wedge$ w4 é o mundo-(caixa opaca) mais próximo,

em que a) é verdadeira se, e somente se, o milhão está na caixa opaca, e b) é verdadeira se, e somente se, a caixa opaca estiver vazia.

Contudo, o defensor da resolução standard não é ingénuo, e, tal como os monocaixistas, conhece os dados do problema, acreditando também que a *pr*(1 milhão|escolhe caixa opaca) = quase 1, e que a *pr*(caixa vazia|escolhe duas caixas) = quase 1. Ora, daqui segue-se que ele acredita também no seguinte:

pr[w1 é o mundo-(2 caixas) mais próximo ∧ w2 é o mundo-(caixa opaca) mais próximo|escolhe caixa opaca] = quase 1, e
pr[w3 é o mundo-(2caixas) mais próximo ∧ w4 é o mundo-(caixa opaca) mais próximo|escolhe duas caixas] = quase 1.

Mas estas duas proposições têm as seguintes consequências:

pr(w2 será actualizado|escolhe caixa opaca) = quase 1
pr(w3 será actualizado|escolhe 2 caixas) = quase 1.

Como *pr*(escolhe caixa opaca ∨ escolhe duas caixas) = 1 e *pr*(w2 será actual ∨ w3 será actual) = 1 - independentemente da probabilidade subjectiva que se atribui às acções condicionantes acima (escolher caixa opaca ou escolher as duas caixas) - o defensor da resolução standard continua a ter virtualmente a certeza de que w2 ou w3 serão actuais. Aliás, mesmo que se atribuam probabilidades abaixo de 1 às proposições condicionantes, tal não tem qualquer influência na crença de que w2 ou w3 serão actuais.
Ora, estas duas consequências consistem precisamente naquilo em que o defensor do princípio (M_o) acredita. Em suma, a crença em (M_o) não é incompatível com a crença em (M_2), ou, por outras palavras, (M_o) não é capaz de caracterizar o princípio fundamental do meta-argumento monocaixista.
Neste ponto, ao questionarmo-nos acerca do que tem o meta-argumento monocaixista de estabelecer, convém relembrarmos a sua formulação:

(M_o) Tenho virtualmente a certeza, **independentemente de quaisquer crenças que eu tenha acerca da probabilidade de vir a escolher a caixa opaca ou as duas caixas**, que que w2 ou w3 se tornarão actuais.

A contenção de Eells é a seguinte: o que (M_o) tem de estabelecer não é que temos virtualmente a certeza de que w2 e w3 serão actuais, independentemente das nossas crenças na probabilidade de virmos a realizar uma ou outra acção. O que tem de se estabelecer é que temos virtualmente a certeza de que w2 e w3 serão actuais, independentemente das acções que serão realizadas. Ou seja, o princípio é acerca de acções e não de crenças. Portanto, (M_o) deverá transformar-se em (M_o'):

(M_o') w2 ou w3, um ou outro tornar-se-ão actuais; e este facto é independente de qual a acção que será efectuada, escolher as duas caixas ou escolher a caixa opaca (Eells 1985: 208).

Deste modo, a crença em (M_o') já se torna incompatível com a crença em (M_2). Se olharmos para 6', que se encontra subjacente a (M_2), veremos que a crença em (M_2) implica a verdade de pelo menos uma das seguintes proposições: 'Escolho as duas caixas e ganho 1.001.000' e 'Escolho a caixa opaca e não ganho nada'. A verdade de pelo menos uma delas é necessária à verdade da base disjuntiva, mas (M_o') não permite que nenhuma delas seja verdadeira.[57]
Este novo princípio (M_o') incorpora uma nova cláusula de independência que pode ser interpretada através da conjunção de duas contrafactuais:

(Se escolhesse as 2 caixas ≫ w2 será actual ou w3 será actual) ∧ (Se escolhesse a caixa opaca ≫ w2 será actual ou w3 será actual).

Esta conjunção, por sua vez, é equivalente a esta outra:

(Se eu escolhesse 2 caixas ≫ w3 será actual) ∧ (Se eu escolhesse caixa opaca ≫ w2 será actual).

Chegamos, assim, ao ponto essencial do argumento de Eells. Esta conjunção, que, segundo ele, caracteriza o novo princípio (M_o'), consiste em nada mais do que a conjunção das duas premissas fundamentais, 3 e 6, do argumento monocaixista original ao nível objecto, e não, como pretendia Horgan, num princípio independente ao meta-

[57] A base disjuntiva tem a seguinte forma lógica: ((C2≫PC1) ∧ (C1≫PC1)) ∨ ((C2≫PC2) ∧ (C1≫PC2)).

nível. Estas duas premissas estão na base da conclusão normativa do argumento monocaixista. Eells conclui, assim, que a circularidade que infecta (M_2), infecta também (M_o), deixando os dois princípios, para já, numa situação de igualdade, e estabelecendo um impasse na discussão entre monocaixismo e bicaixismo.

A resposta de Horgan a este argumento parece-me bastante convincente. O argumento de Eells depende da interpretação que é por ele dada ao princípio (M_o'). Ou seja, como uma conjunção de proposições contrafactuais que expressaria a independência contrafactual dos mundos w2 e w3 relativamente às acções. Contudo, quando nos referimos à independência que caracteriza (M_o') – 'É virtualmente certo que w2 ou w3 serão actuais' – podemos referir-nos apenas à independência desta conclusão relativamente a quaisquer premissas que digam respeito à probabilidade de o agente vir a escolher uma ou outra acção. Ou seja, é necessário que no corpo de premissas do agente (as crenças referidas no princípio original, M_o) que o levam a concluir que a resolução retroactiva da vagueza é a mais apropriada, não existam quaisquer premissas que incluam as contrafactuais do argumento-objecto e as suas probabilidades.

Com efeito, parecem, a meu ver, existir dois tipos de premissas que o defensor da resolução retroactiva tem ao seu dispor para formular o seu meta-argumento. As primeiras dizem respeito à caracterização dos quatro mundos possíveis - w1, w2, w3 e w4; essa caracterização pode ser feita simplesmente através de conjunções de acções e consequências. O segundo género de premissas, por seu lado, diz respeito às circunstâncias que definem o PN, nomeadamente, a fiabilidade do previsor; estas premissas podem consistir em atribuições de probabilidade aos quatro mundos possíveis. E mesmo que desejemos definir tal fiabilidade através de condicionais materiais - por exemplo, 'se escolher a caixa opaca, então ganho 1,000,000' – ainda que seja verdade que estas dependem necessariamente das contrafactuais do argumento no nível-objecto, feita a transição de um nível para o outro deixa de ser necessário invocar essa dependência. Ou seja, se no argumento ao nível-objecto a função das contrafactuais empregues, aí essenciais, era a de garantir a plausibilidade da conclusão normativa de ambos os argumentos, quando chegamos ao meta-nível o nosso objectivo consiste em estabelecer a conclusão de que a resolução retroactiva é a mais apropriada do ponto de vista pragmático, e as premissas agora utilizadas não têm necessariamente de ter a forma de contrafactuais.

Em suma, o princípio que, segundo Horgan, realmente caracteriza o argumento favorável à resolução retroactiva e, a *fortiori*, ao monocaixismo, é o seguinte:

(M_o'') w2 ou w3, um deles será actual; e esta proposição segue-se de um conjunto de proposições pertencentes ao meu corpo de crenças, o qual não contém qualquer proposição a respeito da probabilidade de eu vir a escolher a caixa opaca ou as duas caixas.

Pode-se, assim, constatar que o meta-argumento monocaixista, caracterizado por (M_o''), não necessita da reformulação (M_o'), evitando, desse modo, incorrer em circularidade.
Até aqui, as coisas parecem correr bem para o monocaixista. Contudo, Eells apresentou um segundo argumento, o qual tem como objectivo reformular o meta-argumento favorável à resolução standard das contrafactuais, de modo a que este deixe de exibir a referida circularidade, e possa, assim, ser colocado a par com o meta-argumento favorável à resolução retroactiva. Se esse novo argumento alcançar o seu objectivo, terá, então, de ser declarado um impasse na disputa entre os defensores da resolução standard e os defensores da resolução retroactiva. Se esse for o caso, talvez seja necessário procurar outras linhas de argumentação, recorrendo-se a outros argumentos ao nível-objecto que não façam uso de contrafactuais.
A estratégia que o novo meta-argumento de Eells adopta passa menção, no princípio reformulado, do facto que torna tão plausível o argumento bicaixista ao nível-objecto: que não existe *realmente* causação para trás no tempo e que a acção realizada não altera o conteúdo da caixa opaca. Além disso, o novo meta-argumento bicaixista deverá evitar o recurso a proposições contrafactuais, tais como aquelas que são necessárias para validar o argumento ao nível objecto. Segundo Eells, (M_2) deverá transformar-se em (M_2'):

(M_2') As únicas diferenças entre o mundo actual e os mundos duas-caixas e caixa-opaca mais próximos dizem respeito a aspectos pelos quais as acções são causalmente responsáveis; e efectuar qualquer uma destas acções não pode afectar causalmente o passado (Eells 1985: 210).

Ou seja, no mundo possível mais próximo em que escolho as duas caixas, o único aspecto diferente em relação ao mundo actual consiste no facto de eu ganhar uma determinada quantia de dinheiro, aspecto esse que não depende causalmente da minha acção. Por comparação, de acordo com a resolução retroactiva, um dos dois mundos possíveis mais próximos do actual, w2 ou w3, é um mundo cujo passado é diferente do actual num aspecto causado pela minha acção, o que contraria (M_2'): se escolher a caixa opaca,

actualizo w2, um mundo em que se encontra um milhão na caixa opaca; mas, se escolhesse as duas caixas, actualizaria w3, um mundo em que a caixa opaca se encontra vazia.

Sem dúvida que este meta-argumento monocaixista apresenta uma melhoria significativa relativamente a (M_2). Além disso, parece existir, a meu ver, um aspecto um tanto ou quanto contra-intuitivo relacionado com a ordenação de mundos possíveis, quanto à sua semelhança, associado à resolução retroactiva. Ou seja, um dos dois mundos – w1 e w4 - que não se conta entre os dois mais próximos – w2 e w3 - difere do mundo actual apenas quanto à acção efectuada, sendo o seu passado idêntico ao mundo actual.

Para tornar este ponto mais claro, consideremos, por exemplo, a hipótese em que escolho a caixa opaca. Qual é o ranking da proximidade ou semelhança entre os quatro mundos possíveis e o mundo actual? O seguinte parece-me ser o ranking correcto:

1º – w2 (escolho uma caixa e recebo 1 milhão);
2º – w3 (escolho duas caixas e recebo 1000);
3º – w1 (escolho duas caixas e recebo 1 milhão e mil);
4º - w4 (escolho uma caixa e recebo 0).

Se é certo que dificilmente temos intuições claras acerca da proximidade ou semelhança entre mundos possíveis, não pode deixar de parecer estranho que w3 surja aqui à frente de w1. Ou seja, é mais semelhante a w2 um mundo cujo passado é diferente de w2 (ou em que existe um *pequeno milagre* que altera o conteúdo da caixa opaca), do que um mundo que difere apenas quanto a um aspecto mínimo, acção realizada, e cujo futuro consistirá apenas em ganhar mais mil Euros. De outro modo, é necessário um 'milagre maior' para tornar w3 mais próximo de w2, do que aquele que é necessário (aliás, nenhum) para tornar w1 mais próximo de w2.

Finalmente, a questão crucial consiste, agora, em saber se (M_2') é, de algum modo, circular. Desde logo se pode constatar que este novo princípio não contém qualquer proposição contrafactual resolvida da maneira standard, nem parece ser possível analisá-lo de modo a reduzi-lo a quaisquer premissas que façam parte do argumento bicaixista original. Contudo, faz-se nele referência à noção de causalidade, a qual, se for analisada à maneira de Lewis, irá depender, por sua vez, da noção de dependência contrafactual. Além disso, essa dependência contrafactual tem de ser resolvida de maneira standard, de

modo a que as relações de causalidade envolvidas no meta-argumento (ou a ausência delas) possam funcionar devidamente. Parece, assim, que 'alguma' circularidade volta a infectar o meta-argumento bicaixista O objectivo deste era encontrar uma justificação para se aderir a uma resolução standard da vagueza, mas essa justificação pressupõe já a aceitação de uma visão das relações de causalidade, apenas compatível com uma leitura standard das contrafactuais.

Uma das defesas que o bicaixista tem ao seu dispor consiste em aderir a uma outra análise do conceito de causalidade, embora, a meu ver, seja difícil entender como poderá essa análise ser compatível com a aplicação da teoria de Gibbard e Harper, na medida em que essa aplicação dependerá da atribuição de probabilidades a proposições que são capazes de acomodar ou reflectir a dependência causal. O princípio (M_2') é claro em relação a este aspecto: a diferença entre mundos possíveis depende somente de uma ou de outra acção estar a funcionar como causa (neste caso, de não estar). Como veremos, existem outros tipos de teoria causal (ver §11.1 e conclusão), as quais, embora diferentes da de Gibbard e Harper, garantem os mesmos resultados que esta. Algumas destas teorias são naturalmente compatíveis com outras maneiras de entender o fenómeno da causalidade, nomeadamente, como uma relação de dependência probabilística entre causa e efeito. Contudo, mesmo no contexto dessas outras teorias, suponho que existirão divergências quanto à identificação dos factores que representam a interpretação mais adequada da estrutura causal do problema e que devem ser tidos em conta no cálculo da utilidade esperada. Mesmo aí, o monocaixista encontrará certamente uma estratégia para defender, por exemplo, uma ou outra maneira de interpretar as relações de dependência probabilística, de modo a favorecer a sua opção. Além disso, parece-me que a análise dessas relações poderá continuar a depender de uma análise semântica de contrafactuais, pois ainda que a relações em causa sejam apenas de natureza probabilística, parece-me difícil encontrar maneiras de formular os argumentos-objecto que não dependa do emprego deste tipo de proposições (embora admita essa possibilidade). A opção pela teoria de Gibbard e Harper apresenta-se, assim, como uma escolha natural face à nossa maneira de raciocinar contrafactualmente em problemas de decisão.

Horgan (1985) não é muito optimista quanto ao estado da disputa após a reformulação que Eells faz do meta-argumento bicaixista. Entendendo que a objecção de circularidade perde parte do seu alcance [*loses its bite*], e face à plausibilidade intuitiva de (M_2'), Horgan acaba por declarar um impasse. Segundo as suas palavras, o bicaixista pode

simplesmente, sem inconsistência, recusar-se a jogar o jogo que consiste em procurar uma meta-defesa dos argumentos originais, e recusar-se mesmo a ceder à ideia de que o sucesso nesse jogo determina qual é a resolução pragmaticamente apropriada para a vagueza das contrafactuais.

8.2. Uma solução monocaixista

Não é possível discordar de Horgan. Existem, de facto, outros jogos para jogar. Contudo, a meu ver, o jogo dos meta-argumentos e das contrafactuais é sério demais para nos darmos ao luxo de não jogar. Para se defender uma posição, há que apresentar razões favoráveis a um ou outro dos argumentos originais, e os meta-argumentos são a maneira natural de o fazer.

A meu ver, a importância e a necessidade de defender um ou outro dos argumentos-objecto prende-se com aquilo que está em causa no PN: um confronto entre dois princípios de racionalidade prática. Os dois argumentos-objecto originais captam o espírito dos princípios em confronto, nomeadamente, o princípio da maximização da utilidade condicional esperada (PMUCE) e o princípio da dominação.[58] Esses argumentos, como se viu, não são falaciosos (§5).

O argumento monocaixista original depende das probabilidades condicionais relevantes que permitem a aplicação do PMUCE. Ora, estas probabilidades condicionais encontram-se já implicitamente contidas nas premissas contrafactuais 3 e 6 do argumento-objecto monocaixista, das quais se seguem as seguintes condicionais materiais:

3 * Escolho duas caixas → ganho 1,000.
6 * Escolho caixa opaca → ganho 1 milhão.

Isto não significa que devemos defender a superioridade da teoria evidencial sobre a teoria causal. Esta última parece, para já, ser a única capaz de oferecer os resultados certos em exemplos como o de Salomão e Batsheba (mais adiante veremos o que caracteriza

[58] É certo que a discussão passou a fazer-se entre a teoria evidencial de Jeffrey e uma teoria causal da decisão como a de Gibbard & Harper. Contudo, tanto estes últimos (1978), como o próprio Horgan (1981), apresentam dois tipos de dominação, aos quais correspondem dois tipos de maximização da utilidade esperada, evidencial e causal.

exactamente este tipo de casos). Simplesmente não é abusivo afirmar que o argumento monocaixista original reconhece implicitamente uma interpretação retroactiva de contrafactuais, ao ignorar a ausência de verdadeira causalidade entre os estados e as acções, e funcionando como se essa causalidade estivesse, de facto, presente. Não é um problema para as condicionais materiais serem 'interpretadas retroactivamente', como é o caso de 3* e 6*, mas ambas herdam a sua elevada probabilidade (condicional) da elevada probabilidade de 3 e 6. É a probabilidade destas duas contrafactuais que a teoria causal irá utilizar para calcular a utilidade esperadas das acções, de acordo com a interpretação retroactiva das dessas mesmas contrafactuais.

O argumento bicaixista original, por seu lado, exprime o raciocínio subjacente à dominação: aconteça o que acontecer, ou independentemente da previsão, está sempre ao meu dispor actualizar um mundo possível melhor do que aquele que actualizaria escolhendo a acção contrária, i.e., um milhão e mil em vez de um milhão, e mil em vez de zero. Além disso, o argumento a partir da dominação depende de uma resolução standard desta premissa, garantindo que não se verifica qualquer relação de dependência contrafactual entre as acções e o passado.

Considerando, assim, que os dois argumentos originais ao nível-objecto são as duas maneiras naturais de exprimir, respectivamemte, os raciocínios a partir do PMUCE e a partir da dominação, devemos, então, levar a sério o jogo que consiste em evitar qualquer circularidade nos meta-argumentos que empregamos para justificar a cogência dos primeiros. Como vimos, dado o recurso à noção de causalidade no meta-argumento bicaixista, o meta-argumento monocaixista parece ser o único que consegue evitar completamente essa circularidade (ver §8.1).

Toda esta discussão que se tem vindo a desenvolver tem como base a formulação das premissas dos argumentos mono e bicaixista sob a forma lógica de condicionais contrafactuais. Mas não será possível formulá-los através de condicionais materiais? Se o fosse, não estaríamos a eliminar o fundamento de uma possível interpretação retroactiva, a qual parece ser a única maneira de defender a posição monocaixista?

Na verdade, empregando condicionais materiais, o argumento monocaixista perderia a sua força normativa. O único problema reside no facto de, pela mesma ordem de razões, o mesmo vir a acontecer com o argumento bicaixista. Por exemplo, suponha-se que, decidindo escolher a caixa opaca, eu acredito nas seguintes versões materiais das premissas do argumento monocaixista

a) Eu *escolho*/*escolherei* as duas caixas → Eu *ganho*/*ganharei* 1.000 Euros;
b) Eu *escolho*/*escolherei* a caixa opaca → Eu *ganho/ganharei* 1.000.000 Euros.

Contudo, dada a forma lógica destas condicionais, a crença em a) e b) não é incompatível com a crença na seguinte contrafactual: 'se eu *escolhesse* as duas caixas, então ganharia 1,001,000 Euros'. Deste modo, a conclusão normativa de que devo escolher a caixa opaca perde a sua plausibilidade original, pois dada esta crença contrafactual, eu deveria escolher antes as duas caixas.

É também possível oferecer *a posteriori* um motivo para a necessidade de formular estes argumentos através de contrafactuais. O bicaixista encontra nas suas conclusões acerca do PN motivo para defender uma teoria causal da decisão, a qual operará através das probabilidades das contrafactuais envolvidas. Por seu lado, um monocaixista que deseje abrir a possibilidade de a teoria causal acomodar as suas próprias conclusões, desejerá certamente que o seu argumento se preste a uma interpretação causal, o que só é possível tendo as contrafactuais relevantes interpretadas de forma retroactiva.

A questão final que desejo colocar é a seguinte: será que a permissibilidade de se efectuar esta leitura 'excêntrica' das contrafactuais é a única maneira de tornar razoável a aceitação da conclusão do argumento monocaixista original? A abordagem que se tem feito à semântica das contrafactuais tem sido a de preservar a sua bivalência. É dessa maneira que, de acordo com uma resolução standard, as premissas 1 e 4 se encontram em contradição (ver §4.1). Contudo, não é excêntrico aceitar uma abordagem que negue a possibilidade de se atribuirem os tradicionais valores de verdade a contrafactuais ou, até mesmo, às condicionais em geral. No caso presente ser-nos-ia útil uma abordagem que nos permitisse, por exemplo, lidar com estas expressões não em termos de verdade ou falsidade, mas sim em termos de grau de crença. Se isto for possível, talvez seja também possível eliminar a contradição que aparenta minar o argumento monocaixista e estabelecer novamente, contra os cépticos, uma paridade entre os dois argumentos, sem a necessidade de se apelar a quaisquer interpretações controversas das contrafactuais.

A problemática da atribuição de condições de verdade às condicionais tem uma longa história e são bem conhecidas, por exemplo, as limitações de uma perspectiva verofuncional das expressões do tipo 'Se A, B'. Não tão longa, embora igualmente controversa, é a discussão da semântica das contrafactuais. A análise tem vindo a ser feita

com base na aceitação das condições de verdade de Lewis, embora estas não estejam isentas de dificuldades. Considere-se:

(c) Se Oswald não tivesse matado Kennedy, mais ninguém o teria feito.

Suponhamos que (c) é verdadeira. O critério é o da semelhança entre mundos possíveis: no mundo possível mais próximo em que Oswald não matou Kennedy, Kennedy não teria sido assassinado.

Num mundo de leis determinísticas, se Oswald tivesse espirrado antes de disparar em *t*, e tivesse perdido a oportunidade de o fazer, então algo anterior a *t* teria de ter sido diferente do mundo actual, de modo a provocar o espirro, e assim sucessivamente, recuando-se até ao início dos tempos. Logo, o mundo possível aparentemente mais próximo do actual, em que Oswald espirra em *t*, teria sido em *t* muito diferente do mundo actual. Mas considere-se um outro mundo em que, tendo um passado idêntico ao do mundo actual, ocorre um 'pequeno milagre' em *t* e Oswald espirra. Qual destes dois mundos é mais semelhante ao mundo actual? De acordo com Lewis (ver §7.1), o mundo em que ocorre um 'pequeno milagre' é o mais semelhante ao mundo actual, pois o críterio 2 – 'maximizar a região do espaço-tempo em que prevalece uma coincidência perfeita de factos particulares – sobrepõe-se ao critério 3 – 'evitar pequenas violações das leis da natureza, ainda que simples e localizadas'. O referido 'pequeno milagre' consiste na ocorrência do espirro de Oswald.

Podemos, contudo, questionar o pressuposto de que, na análise da semelhança entre mundos possíveis, estamos sempre perante leis determinísticas. Existirá alguma razão pela qual uma teoria acerca de contrafactuais necessite que encaremos o mundo de uma forma determinista? É certo que as contrafactuais estão relacionadas com leis da natureza. Essa relação estava dada como primitiva nas primeiras teorias de contrafactuais, como a de Goodman (ver §7.1), e é explicada na teoria de Lewis através do critério da semelhança: dois mundos com leis da natureza idênticas são mais semelhantes entre si do que dois mundos com leis da natureza diferentes. Contudo, nada é exigido quanto à natureza determinística dessas mesmas leis.

Considere-se, então, uma terceira possibilidade: um mundo em tudo semelhante ao actual até *t*, altura em que um 'pequeno milagre' ocorre e Oswald espirra, mas em que um outro 'pequeno milagre' ocorre (talvez não tão pequeno como o anterior) e um espectador do

cortejo presidencial, que transporta sempre consigo uma arma e detesta Kennedy, decide num impulso súbito disparar e matar o presidente. Este mundo será, depois de *t*, muito 'mais semelhante' ao mundo actual do que aquele em que Oswlad espirra e perde a oportunidade de disparar. Os critérios de Lewis acomodam este facto: um mundo em que ocorrem dois 'pequenos milagres' é menos semelhante ao mundo actual, embora tenham passados e futuros semelhantes, do que um mundo em que ocorre apenas um milagre, embora com um futuro diferente do mundo actual. Em suma, de acordo com a semântica de Lewis, a contrafactual continua a ser verdadeira. Contudo, as nossas intuições acerca da semelhança entre mundos possíveis parecem aqui ir contra aquilo que os critérios de Lewis nos autorizam a acreditar.

Considere-se agora que vivemos num mundo não-determinista e que existia, de facto, uma real possibilidade tanto de Oswald espirrar, como de um outro espectador, particularmente perturbado, ter disparado e matado Kennedy. Nesse caso, não são necessários 'pequenos milagres' para que essas coisas aconteçam; portanto, um mundo sem milagres e virtualmente idêntico ao nosso quanto ao passado e futuro, parece ser mais próximo do mundo actual do que aquele em que Oswald espirra e Kennedy vive até aos oitenta.[59] Para que a contrafactual (c) seja verdadeira, para Lewis, o consequente tem de ser verdadeiro em *todos* os mundos possíveis mais próximos; portanto, levando a sério os critérios da resolução standard para a semelhança entre mundos possíveis, a contrafactual em causa parece afinal ser falsa (mesmo supondo que não houve conspiração).

Existem outros exemplos cuja avaliação intuitiva que deles fazemos parece ir contra a análise da semelhança proposta por Lewis. Por exemplo, ao fim de 94 minutos de jogo, e após desperdiçar inúmeras oportunidades de golo ao longo de toda a partida, o Sporting não conseguiu vencer o Barcelona, tendo o jogo acabado com um empate 0-0. De modo a cumprir o requisito 2 de Lewis, a seguinte contrafactual parece ter de ser verdadeira:

Se o Sporting tivesse ganho por 1-0, teria marcado um golo perto dos 94 minutos.

Como sabemos, o mundo possível mais próximo é sempre aquele que partilha com o mundo actual uma maior região do espaço-tempo em que a semelhança de leis e factos é total. Por exemplo, *t* é o momento do tempo em que um determinado mundo possível

[59] Este exemplo tem como base um outro com a mesma estrutura apresentado por Edgington (1995: 255-257).

começa a divergir do mundo actual; para tornarmos a antecedente de uma contrafactual verdadeira, tentamos atrasar o mais possível o momento *t* e mudar o menor número possível de coisas, antes de *t*, para pôr a funcionar o mundo possível que começa em *t*.
Ou, então, considerem-se as seguintes circunstâncias (Tichy 1976): quando António sai à rua, se o tempo estiver mau, ele usa sempre chapéu; se o tempo estiver bom, então existe uma probabilidade de 0,5 de ele usar chapéu. Suponha-se que no mundo actual o tempo está mau e António usa chapéu. De acordo com a análise da semelhança, a seguinte contrafactual é verdadeira:

Se o tempo estivesse bom, António teria usado o chapéu.

É mais semelhante ao mundo actual um mundo em que faz bom tempo e em que, como no mundo actual, António usa chapéu, do que um mundo em que faz bom tempo e António sai de cabeça descoberta. Contudo, temos alguma relutância em afirmar peremptoriamente que esta contrafatual é verdadeira.
Estes dois exemplos, particularmente o segundo, têm uma característica relevante para o que aqui nos interessa. Apesar de a eventual atribuição de verdade às contrafactuais em causa não parecer completamente chocante, a nossa intuição diz-nos que a sua falsidade também não pode ser declarada determinantemente. É possível que o Sporting tivesse marcado um golo aos 93,59 minutos. É possível, e até bastante provável, que António tivesse usado chapéu, caso o tempo tivesse estado bom. A avaliação das contrafactuais em causa presta-se a ser feita através de uma perspectiva probabilística ou em termos de grau de crença. Esta perspectiva acomoda perfeitamente a ideia bastante popular segundo a qual a determinação do valor de verdade das condicionais não pode ser independente da consideração do contexto em que são asseveradas. Uma maneira natural de definir esse contexto é atribuir ao falante uma certa distribuição de probabilidades sobre o conjunto de maneiras através das quais a antecedente da contrafactual pode ser verdadeira ou falsa. Isto significa, precisamente, que essa distribuição pode ser feita sobre um conjunto de mundos possíveis. Além disso, esta sugestão tem a vantagem de poder ser aplicada a todas as condicionais, indicativas e contrafactuais, no caso de sermos cépticos, como muitos o são, relativamente à verofuncionalidade da conectiva ‘se, então’. Como vimos, as condições de verdade de Stalnalker para contrafactuais – no essencial semelhantes às de Lewis – aplicam-se também às condicionais indicativas.

Esta sugestão terá consequências algo dramáticas e parece equivaler a uma desistência completa de tentar determinar 'verdadeiras' condições de verdade. Por exemplo, será que apenas condicionais como 'se António e Maria estão em Lisboa, então António está em Lisboa' é que podem ser designadas genuinamente como 'verdadeiras'? Por outro lado, qualquer condicional contingente parece poder ser considerada verdadeira, desde que o contexto seja propício, i.e., dependendo da distribuição de crenças do falante e da audiência, o que parece ser, à partida, uma desistência total de obter qualquer género de objectividade. Mas a proposta consiste, precisamente, em substituir a atribuição de verdade ou falsidade a este tipo de proposições pela atribuição de graus de crença às mesmas. Portanto, o grau de crença atribuído à condicional acima só pode ser 1, sob risco de incorrermos em grosseira irracionalidade. Do mesmo modo, existirão outras condicionais, tais que o grau de crença nas mesmas só pode ser 0, e todas as condicionais contingentes podem situar-se em qualquer posição do espectro entre 0 e 1. Isto respeita perfeitamente as condições de assertabilidade que estamos dispostos a aceitar para condicionais: existe um número enorme de condicionais contingentes que, tendo em conta o contexto em causa, estamos dispostos a asseverar, embora o grau de crença nas mesmas não chegue a 1.

A proposta considerada foi adoptada por Bas Van Fraassen (1976), enquanto extensão da semântica de Stalnaker, com base precisamente na ideia de que as condicionais não constituem um discurso acerca de factos:

'Se existem apenas factos respeitantes a este mundo, e não contra-factos ou factos respeitantes a outros mundos possíveis, as frases que envolvem condicionais não podem ser avaliadas com base na sua correspondência com os factos. Isto porque elas são acerca daquilo que não é um facto. Assim, exceptuando aquelas que não podem ser verdadeiras – como $[A \wedge (A \rightarrow \neg A)]$, que tem necessariamente de ser falsa, e outros casos limite - o valor de verdade de tais frases parece ser indeterminado. (…) as condicionais contrafactuais não têm como função apresentar factos, e, num sentido estrito, nenhuma delas merece ser chamada "verdadeira" ou "falsa". A sua função é diferente, tal como é diferente a função das frases no modo interrogativo ou imperativo (Van Fraassen 1976: 267-68).

As coisas funcionam da seguinte maneira: quando eu tenho um grau de crença de 0,9 numa determinada condicional 'A ≫ B', isso significa que, de acordo com a minha distribuição de crenças, 90% dos mundos-A são mundos-B. Podemos retomar aqui as noções de verdade e falsidade para compreendermos melhor o que está em causa: nos mundos-(A ∧ B), A ≫ B é verdadeira, nos mundos-(A ∧ ¬B), A ≫ B é falsa, mas se o mundo actual for um mundo-(¬A), então existe 90% de hipóteses de o mundo possível mais próximo do actual se tratar de um mundo-B e 10% de hipóteses de se tratar de um mundo-(¬B).

O projecto de Van Fraassen é bastante mais ambicioso do que esta simplificação, procurando na verdade estabelecer que o grau de crença (*c*) numa determinada condicional, *c*(A ≫ B), é idêntico ao grau de crença em B, na condição de A, *c*(B|A). De acordo com a semântica formal de Stalnaker, existe uma função de selecção *f* que seleciona para qualquer mundo *m* e uma qualquer proposição A, o mundo possível *m'* mais próximo de *m* em que A é verdadeira. Se A for verdadeira no mundo actual, então *f* seleciona o mundo actual *m*, mas se A for falsa no mundo actual, então *f* seleciona o mundo-A mais próximo. A essência da proposta de Van Fraassen consiste no seguinte: quando A é falsa, a selecção faz-se ao acaso (*at random*) de acordo com o grau de crença na respectiva condicional, tal como se estivéssemos a retirar bolas marcadas de dentro de um saco, em que 90% delas são mundos-B. É neste sentido probabilístico bastante preciso que repousa a ideia segundo a qual, salvo em casos limite pouco interessantes, o valor de verdade das contrafactuais é indeterminado.[60]

Ao contrário do que acontece com Lewis, que pretende realmente que se possa fazer uma análise das contrafactuais através da noção vaga de semelhança entre mundos possíveis, a proposta de Van Fraassen não necessita que se invoquem quaisquer factos acerca da maior ou menor similitude entre mundos possíveis, nem se acredita que é possível determinar exactamente o modo através do qual a função realmente funciona. Tal como Stalnaker acreditava, não se pretende que a teoria seja informativa acerca de questões de

[60] Jeffrey (1991) (mencionado em Edgington (1995: 309)) apresentou igualmente uma proposta para atribuir a uma condicional 'Se A, B' um valor semântico intermédio (o que parece ser uma outra maneira de dizer 'indeterminado'), equivalente ao grau de crença na mesma quando A é falsa, dada uma distribuição de crenças específica. Suponha-se que o meu grau de crença em 'Se A, B' é de 90%, e que a minha distribuição de crenças relativamente às proposições relevantes é a seguinte: *c*(A ∧ B) = 0,4, *c*(A ∧ ¬B) = 0,1 e *c*(¬A) = 0,5. A maneira de calcular o valor semântico da condicional consiste, então, numa média ponderada dos valores semânticos atribuídos às respectivas proposições relevantes (ocupando 1 e 0, respectivamente, o lugar de V e F): *c*(Se A, B) = (0,4 × 1) + (0,1 × 0) + (0,5 × 0,9) = 0,85 = *c*(B|A). Stalnaker e Jeffrey (1994) (mencionado em Edgington (1995: 309)) mostraram que esta construção é semelhante à de Van Fraassen).

similitude, independentemente daquilo que acreditamos acerca da verdade de B, caso A fosse verdade. Este aspecto reveste-se de particular importância para os nossos propósitos, nomeadamente, na avaliação das premissas 1 e 4 do argumento monocaixista:

1. Se eu escolhesse as duas caixas, então o previsor prevê-lo-ia.
4. Se eu escolhesse a caixa opaca, então o previsor prevê-lo-ia.

Como vimos, de acordo com uma resolução standard da vagueza, estas duas premissas não podem ser ambas verdadeiras. A verdade do antecendente não tem qualquer influência causal na produção do consequente, e se, por exemplo, 1 for verdadeira no mundo actual *m*, o mundo *m*' mais próximo de *m* é um mundo em que a previsão é idêntica, o que faz com que 4 seja falsa.
Contudo, se abordarmos estas contrafactuais através da perspectiva segundo a qual o seu valor de verdade é indeterminado, a contradição que mina o argumento-objecto monocaixista dissolve-se. Reitera-se que é necessário que estas duas proposições tenham a forma lógica da condicional de Stalnaker/Lewis (A ≫ B), independentemente da sua forma gramatical. Se quisermos podermos formulá-las com a subordinada no futuro do conjuntivo,

1'. Se eu escolher as duas caixas, então o previsor tê-lo-á previsto,

sem que, com isso, se perca o que está em causa: uma condicional que expressa uma expectativa não-realizada, em que o antecedente não é verdadeiro (nem falso) no mundo actual.[61] Como vimos acima, o uso da condicional material torna implausível as conclusões normativas de ambos os argumentos.
Adoptando a semântica proposta por Van Fraassen, e considerando a elevada taxa de sucesso do previsor, o que temos perante nós é o seguinte: o agente terá um grau de crença de 0,9 na premissa 1, e um grau de crença de 0,9 na premissa 4. Ou seja, o agente

[61] A distinção que é feita, em termos de condições de verdade, entre condicionais indicativas, por um lado, e contrafactuais ou conjuntivas, por outro, não é de todo pacífica. Edgington (1995: 311-323) argumenta no sentido de mostrar que a diferença entre uma contrafactual como 'se eu tivesse escolhido…' e uma indicativa como 'se eu escolher…', consiste apenas no seguinte: a primeira diz, num momente posterior à acção, aquilo que a segunda dizia, apropriadamente, no momento anterior à acção. Do mesmo modo, a proposição 1 acima parece dizer exactamente o mesmo, perante um caso hipotético, que 1' diria se essa hipótese fosse uma possibilidade real. Não pretendo, contudo, tomar posição sobre este tópico.

considera que 1 tem 90% de hipóteses de ser verdadeira, caso um mundo-(duas caixas) venha a ser o caso, e que 4 tem 90% de hipóteses de ser verdadeira, caso um mundo-(caixa opaca) venha a ser o caso. É preciso ter em conta que nesta perspectiva ambas as premissas têm um valor de verdade indeterminado. Ou seja, não é o caso que, considerando a premissa 1, por exemplo, os mundos em que 'eu escolho as duas caixas' é falsa se dividam entre mundos-(duas caixas ∧ previsão correcta) e mundos-(duas caixas ∧ previsão errada). Como se mencionou, de acordo com esta perspectiva, não existem factos acerca de mundos possíveis que tornem qualquer uma destas premissas verdadeira ou falsa. O uso destas expressões tem aqui apenas um papel expositivo. Ao contrário do que acontecia com a semântica de Lewis, não há qualquer análise do modo como funciona a função de selecção.[62] Por estas razões, um agente que mantenha os respectivos graus de crença acerca das duas premissas não parece demonstrar inconsistência ou irracionalidade. Cabe, agora, questionarmo-nos acerca das consequências que este resultado acarreta no que respeita à avaliação dos argumentos mono e bicaixista no PN.

O argumento-objecto monocaixista dependia, para a sua validade, de uma leitura retroactiva das contrafactuais envolvidas. Agora, a sua validade já não depende desta interpretação, e os dois argumentos originais, ao nível-objecto, podem ser colocados par a par. A premissa 7 do argumento monocaixista faz uso da noção de verdade – se 3 e 6 forem verdadeiras, então devo escolher caixa opaca. Se em vez disto eu disser 'se 3 e 6 tiverem 90% de hipóteses de ser verdadeiras, então devo escolher a caixa opaca', a força normativa da conclusão não parece sair debilitada. 7 é uma conclusão normativa e não se exige, portanto, que as premissas das quais se segue tenham necessariamente de ser verdadeiras ou falsas – afinal existem argumentos válidos acerca de questões morais e existem sérias dúvidas quanto à posse de valor de verdade das premissas envolvidas. De qualquer modo, o argumento depende de inferências válidas que partem de premissas com valor de verdade indeterminado, para outras que, assim o desejaríamos, deveriam pelo menos herdar o grau de crença que as primeiras merecem – assim acontece de 1 para 3 e de 4 para 6. Ou seja, parece ser necessária uma noção de validade que acomode a noção

[62] É devido a esta desistência de avançar com genuínos valores de verdade para a condicional - quando o antecedente é falso – que a proposta de Van Fraassen consegue escapar à demonstração de Lewis (ver §6.1 e Apêndice 2). Por outro lado, existe uma outra razão pela qual Van Fraassen não está preocupado com a prova de Lewis; Van Fraassen (1976: 252) acredita que esta prova depende de um pressuposto (nas suas palavras, 'realismo metafísico') que, segundo ele, deve ser rejeitado: que em todas as distribuições de crença esteja sempre em causa a mesma proposição. A discussão da rejeição deste pressuposto não se encontra, todavia, nos limites deste trabalho. Para o que aqui nos interessa basta saber que, para Van, Fraassen qualquer proposição contingente que tenha ¬A como antecedente tem um valor de verdade indeterminado.

de preservação de probabilidade. Intuitivamente isto pode parecer-nos uma evidência, pois deparamos constantemente com raciocínios que partem de premissas contingentes, que temos como incertas, incluindo algumas sob a forma de condicionais indeterminadas. O aparato técnico que demonstra esta possibilidade garante valores bastante aceitáveis, os quais não minam a confiança que temos na aceitabilidade das conclusões (Adams 1975: Cap. 2).

Por si só, o resultado obtido não constitui, contudo, um avanço significativo ao ponto de encerrar a discussão da avaliação geral dos argumentos mono e bicaixista. Isto porque teremos sempre de decidir acerca do modo como lemos as condicionais em causa: causalmente ou 'apenas' evidencialmente. Como sabemos, Lewis pretende oferecer-nos uma análise da noção de causalidade em termos de dependência contrafactual, e é essa noção que se encontra na base da teoria causal de Gibbard e Harper. Como veremos mais adiante, encontram-se expressas nessas contrafactuais, que Lewis designa como *dependency hypothesis* (ver §10.1), as nossas crenças acerca da estrutura causal do mundo, crenças essas que são relevantes para a resolução do problema de decisão em causa. Mas será possível adoptar a semântica de Stalnaker/Van Fraassen e preservar a análise da causalidade de Lewis? Segundo Edgington (1995), Stalnaker (1984: 157-160) não envereda por este caminho, acreditando que o conceito de causalidade não é sequer passível de análise.

De qualquer modo, temos de estar conscientes de que se verifica aqui uma tensão: a semântica de Stalnaker/Van Fraassen não pretende ser informativa quanto a questões de semelhança entre mundos possíveis; contudo, esta noção tem de ser inevitavelmente invocada para efeitos de cálculo da probabilidade dessas contrafactuais (ver §9.2). Mais, tal semântica implica a atribuição de um grau de crença nessas proposições equivalente ao grau de crença no consequente, na condição de se verificar o antecedente, $c(B|A)$. Ora, isto parece funcionar bem quando a relação de causalidade segue a relação evidencial e o valor da probabilidade da contrafactual é idêntico à probabilidade condicional respectiva. Mas o que fazer em casos como o de Salomão e Batsheba, em que a probabilidade da contrafactual é diferente da probabilidade condicional?

Esta tensão só pode, a meu ver, ser resolvida de uma forma. Temos de distinguir entre dois tipos de empreendimento: aquele que consiste em determinar o valor de verdade da condicional de Stalnaker, e aquele que consiste em determinar a probabilidade das contrafactuais envolvidas nos problemas de decisão e, *a fortiori*, a racionalidade das

nossas acções. No primeiro, declaramos como indeterminado o valor de verdade de A ≫ B (excepto quando A e B são verdadeiras no mundo actual, e quando A é verdadeira e B falsa), ou seja, quando temos um grau de crença condicional, *c*(B|A), entre 0 e 1. No segundo empreendimento, fazemos uma estipulação e pressupomos não só que existe um e um único mundo possível mais próximo do actual em que A é verdadeira, mas que também é possível determinarmos que mundo é esse, nomeadamente, se é um mundo-(A ∧ B) ou se é um mundo-(A ∧ ¬B). É razoável aceitar que o nosso grau de crença numa proposição como 'se eu escolhesse a caixa-opaca, ganharia 1,000,000' é de 0,9, grau de crença coincidente com a nossa visão das relações evidenciais em causa (e até mesmo com as observações verificadas no passado), e, no entanto, para efeitos de avaliação do impacto causal da nossa acção, atribuirmos uma probabilidade a esta proposição que não se encontra de acordo com esse grau de crença 'inicial'. Se se fizer uma interpretação causal dessa contrafactual, então essa probabilidade coincidirá com o nosso grau de crença 'inicial'; se, por outro lado, essa interpretação for do tipo não-causal, como no caso do bicaixismo (e no exemplo de Salomão e Batsheba), então essa probabilidade será menor. A correcção das interpretações em causa encontrar-se-á dependente, portanto, dos meta-argumentos que acharmos mais convincentes.

Esta divisão do trabalho está de acordo com uma maneira natural e intuitiva de abordarmos estes assuntos, pois o empreendimento que consiste em decidirmos o que fazer apresenta características muito próprias. Quando avaliamos uma decisão em geral, ou uma decisão tomada por outros, estamos atentos *prima facie* às relações evidenciais entre a decisão tomada e aquilo que essa decisão nos diz acerca de quem tomou a decisão. Por exemplo, quando socilitados a apresentar o nosso grau de crença relativamente à proposição 'se Salomão roubasse Batsheba, então haveria uma revolta', avaliamos a decisão tomada por Salomão e aceitamos naturalmente a crença em como este é provavelmente um rei não-carismático; logo, temos uma crença forte em como haveria uma revolta, se Salomão roubasse Batsheba. Algo diferente parece acontecer quando avaliamos a situação como se estivéssemos no lugar de Salomão: neste caso, estamos atentos *prima facie* aos poderes causais da 'nossa' decisão e às vantagens que podemos colher da execução da acção. Ou seja, dependendo do ponto-de-vista, as consequências da acção são bastante distintas, pois o agente é o único que tem algo a ganhar com a situação: um milhão, no caso do PN, e Batsheba no caso de Salomão. Se quisermos,

podemos dizer que estamos perante duas perspectivas 'epistémicas' distintas, as quais se encontram relacionadas com os interesses dos sujeitos em causa.

À primeira vista, teremos de admitir que este é um argumento 'meramente' pragmático, embora, mesmo que isso seja o caso, não creio que nos devemos preocupar demasiado com esse facto. Não nos devemos admirar que a solução para um problema de decisão recorra a argumentos deste tipo, nem que considerações pragmáticas sejam relevantes para efeitos de análise semântica. Contudo, não é verdade que esta divisão de trabalho repouse apenas em considerações pragmáticas. Uma observação de Frank Ramsey (1931), num contexto aparentemente danoso para o monocaixista, vem inesperadamente em nossa ajuda:

'(…) qualquer possível volição nossa, no presente, é (para nós) irrelevante para qualquer evento passado. Para outro (ou para nós próprios no futuro) pode servir de sinal do passado, mas para nós, agora, o que fazemos afecta apenas a probabilidade do futuro'. (Citado de Ahmed (2014: 216)).

Ou seja, Ramsey distingue duas maneiras de encarar o mundo – duas perspectivas epistémicas, se quisermos: enquanto observadores ou enquanto agentes deliberadores. Para um observador, qualquer acção efectuada no presente por um terceiro contribui legitimamente para tirar conclusões ou modificar as suas crenças acerca de eventos passados; de outro modo não seria possível acreditar no relato de testemunhas com base no seu depoimento. Por outro lado, uma acção que efectuámos no momento *t*0 pode, para nós próprios em *t*1, constituir evidência para fazermos inferências fidedignas acerca de eventos ocorridos em *t*-1 em que estivemos envolvidos; por exemplo, recordarmos que no jantar festivo da noite anterior ofendemos uma amiga com sugestões inconvenientes, leva-nos a acreditar que muito provavelmente estaríamos embriagados. Ora, quando avaliamos o valor de verdade de proposições como 1 e 4, é precisamente isso que estamos a fazer: a tirar conclusões acerca do passado (ou a atribuir probabilidades a proposições acerca do passado), com base em acções que um agente hipotético (um terceiro ou nós próprios) poderá ter feito ou vir a fazer. Não me parece possível negar que o empreendimento que consiste em atribuir valor de verdade a proposições, a partir de uma perspectiva epistémica adequada, deve ser feito a partir do ponto de vista imparcial do observador. Por outro lado, calcular a probabilidade de contrafactuais é um

empreendimento efectuado no contexto da teoria da decisão e, portanto, fazemo-lo sempre do ponto de vista do agente ou, pelo menos, colocando-nos na sua posição. No nosso caso, dada a escolha da teoria de Gibbard e Harper, é inevitável proceder a esse empreendimento técnico.

A perspectiva de Ramsey foi classificada por Ahmed (2014) como 'dualismo evidencial': devemos ter políticas diferentes de lidar com as relações evidencias entre acções presentes e eventos passados, dependendo da nossa perspectiva epistémica.[63] Note-se que a defesa da tese de Ramsey não é incompatível com a defesa da teoria evidencial. Enquanto esta estabelece recomendações acerca do que *fazer*, a primeira diz respeito àquilo que devemos *pensar*, sendo as duas logicamente independentes uma da outra. Esta distinção permite-nos, ainda assim, constatar e confirmar a plausibilidade da proposta de divisão de trabalho apresentada, pois quando se trata de estabelecer o valor de verdade de proposições, estamos certamente a falar daquilo em que devemos *pensar*/*acreditar*. Contudo, a tese de Ramsey tem implicações no que respeita à maneira como entendemos o confronto entre teorias da decisão rivais. Ao indicar-nos o tipo de evidência que devemos ter em consideração quando colocados na perspectiva do agente, a sua aceitação permite-nos identificar a teoria causal como a teoria evidencial 'verdadeira'. Ou seja, a tese de Ramsey permite uma reconciliação entre a teoria causal e a teoria evidencial, ao permitir que as duas coincidam uma com a outra, o que constitui uma derrota *de facto* para os evidencialistas.

A meu ver, a recomendação avançada (é isso que interpreto a tese de Ramsey) constitui uma exigência de racionalidade no que respeita à divisão de trabalho proposta, principalmente quanto à necessidade de se considerar certo tipo de evidência para efeitos

[63] Parte da motivação para se aceitar esta perspectiva está relacionada com a necessidade de se salvaguardar o livre-arbítrio dos agentes, de modo a dar-se sentido a qualquer procedimento de tomada de decisão. Ou seja, para que um agente leve a sério o seu papel deliberativo na escolha da acção, ele tem (pelo menos) de supor o seu livre-arbítrio, algo que só é possível quando o agente se considera livre de influências exteriores, prévias à sua acção, que influenciam a probabilidade da sua ocorrência. Existe, a meu ver, um argumento forte favorável ao dualismo evidencial: sou informado da estreita correlação entre eventos passados do tipo *E* e acções subsequentes do tipo *A*; se estiver em meu poder efectuar uma acção do tipo *A*, então eu posso decidir efectuar *A* se, e somente se, o evento correspondente do tipo *E* não tiver ocorrido; a constatação de que eu posso sempre agir de modo a contrariar a correlação funciona como uma espécie de *tickle* (ver §10.2) que me permite quebrar a relação evidencial entre o passado e as minhas acções futuras. Ainda relacionada com a tese de Ramsey, oferecendo-lhe maior plausibilidade, encontra-se a ideia de que a deliberação *sufoca* [*crowds out*] a previsão (Price 2012): se alguém ainda não tomou uma decisão, então dificilmente se deverá levar a sério a sua confiança em como fará isto ou aquilo, pois permanece em aberto a possibilidade de mudar de ideias. Ou seja, a questão 'Farei x?' é transparente relativamente à questão 'Devo fazer x?'. A resposta à primeira é dada através da resposta à segunda. Para uma crítica à tese de Ramsey, ver Ahmed (2014, Cap. 8).

de formação de crenças. Além disso é certamente uma política racionalmente adequada de lidar com a evidência para efeitos de tomada de decisão, como se viu no caso de David e Bathsheba (e noutros casos igualmente persuasivos, ver §10.1). Contudo, vimos que a necessidade de se proceder a uma resolução da vagueza das contrafactuais, para efeitos do cálculo de probabilidades, é inevitável e inerente à aplicação da teoria causal (ver a análise da teoria de Savage). Daí que, embora a recomendação se aplique à grande maioria de problemas de decisão em que as relações evidenciais não espelham causação, existem alguns casos - PN e Dilema do Prisioneiro (ver §9) – que, tal como se tem argumentado, constituem excepções.

Finalmente, existe um outro problema aparente relacionado com a adopção da semântica de Van Fraassen: ao reduzirmos o valor semântico de contrafactuais a crenças condicionais incertas, não estaremos a comprometer-nos com uma perspectiva indeterminista do mundo? Quando avaliamos o valor semântico de uma contrafactual que expressa causalidade, como, por exemplo,

Se o fósforo tivesse sido riscado, então ter-se-ia acendido,

estaremos dispostos a atribuir-lhe um grau de crença de valor 1 e, desse modo, a considerá-la como ‘genuinamente verdadeira’? Creio que isso não será necessário, pois afinal estamos perante uma condicional contingente e o grau de crença que este nos merece, ainda que extremamente elevado, deve ter em conta as circunstâncias naturais (os factos ‘não-compatíveis’ (ver §7.1, n. 49) que podem prevenir o acendimento do fósforo: estar molhado, a superfície não ser suficientemente rugosa, etc. Isto é perfeitamente compatível com uma perspectiva determinista. De facto, nenhuma das maneiras possíveis de analisar conceptualmente o fenómeno da causação se encontra comprometida com qualquer uma das duas perspectivas, determinista ou indeterminista.

Resumindo, existirá algo de novo relativamente ao argumento monocaixista, ao nível-objecto, que favoreça uma interpretação causal das condicionais nele envolvidas? Como avaliaremos a premissa 6’ do argumento bicaixista, quanto ao nosso grau de crença na mesma? Esta premissa consiste naquilo que se designou como a *base disjuntiva* da inferência para a conclusão normativa e, como é fácil de constatar, trata-se de uma partição; ou seja, a sua verdade é necessária. E o que dizer do grau de crença que mantemos nas condicionais 2’-5’, a partir das quais inferimos 6’? Mais uma vez, a sua

verdade não está em causa. Contudo, estas premissas não parecem conseguir responder a uma questão que interessa e que parece ser essencial para um decisor racional: qual é a probabilidade de ganhar o milhão, caso escolha a caixa opaca? E qual é a probabilidade de ganhar o milhão, caso escolha as duas caixas? As circunstâncias são bem conhecidas e sabemos perfeitamente quais são estas probabilidades. O problema reside no facto de o argumento-objecto bicaixista não fazer qualquer menção das mesmas, tal como se desejasse evitar uma questão que lhe é penosa. Pelo contrário, sabemos que o argumento monocaixista inclui agora uma menção explícita, nas premissas que o compõem, a essas probabilidades; ou seja, essas premissas podem ser lidas da seguinte maneira: 'Se eu escolhesse a caixa opaca, então existiria 90% de hipótese de o previsor o prever', e 'Se eu escolhesse as duas caixas, então existiria 90% de hipótese de o previsor o prever'.[64] Mais do que isso, sabemos agora que essas premissas não necessitam de qualquer interpretação pouco ortodoxa de modo a providenciarem a informação exigida pelo agente racional.

A vantagem que a semântica de Van Fraassen nos traz torna-se, assim, dupla: retira peso ao papel da interpretação retroactiva das contrafactuais, remetendo-a ao segundo empreendimento - calcular a probabilidade das mesmas para efeitos do cálculo - e expõe nas premissas dos argumentos-objecto a informação ou os valores que o agente considera relevantes para efectuar esse cálculo.

Estes valores, que dizem respeito às probabilidades dos dois únicos mundos possíveis viáveis - w2 e w3 –, complementam ou tornam mais informativo, se quisermos, o meta-argumento favorável à resolução retroactiva da vagueza das contrafactuais:

(M_o''') w2 ou w3 têm ambos uma probabilidade de 0,9 de virem a ser actuais, e w1 e w2 uma probabilidade de 0,1; e esta proposição segue-se de um conjunto de proposições pertencentes ao meu corpo de crenças (acerca da fiabilidade do previsor), o qual não contém qualquer proposição a respeito da probabilidade de eu vir a escolher a caixa opaca ou as duas caixas.

Este meta-argumento parece ser tanto mais convincente quanto maior foi a fiabilidade do previsor e quanto maior for o grau de crença do agente nas premissas 3 e 6 do argumento

[64] Se tivermos um grau de crença condicional *c*(B|A), torna-se plausível ler uma condicional da seguinte maneira: se A for o caso, então existe uma probabilidade *x* de B ser o caso.

ao nível-objecto. Isto parece implicar que quando essa fiabilidade é pouco superior a 0,5 a plausibilidade do meta-argumento sai um pouco enfraquecida. Esta é uma conclusão algo incómoda para os monocaixistas, pois, como sabemos, estamos na presença de um PN, desde que o grau de fiabilidade do previsor seja superior a 0,5005. Mas este enfraquecimento não tem necessariamente de se verificar, dependendo do ponto de vista a partir do qual a discussão em redor dos meta-argumentos é abordada. Se a discussão estiver a ser tida num nível intuitivo, pré-teórico, então é verdade que, dada a menção das probabilidades relevantes no meta-argumento moncaixista, quanto menores forem os valores das probabilidades de w2 e w3, menos convincente parece ser o argumento. Contudo, dada a quantia astronómica presente na caixa opaca, os agentes que são confrontados com o problema, e que são leigos quanto ao PMUCE (evidencial ou causal), dificilmente se deixarão tentar pelo argumento bicaixista, à medida que os valores de w2 e w3 forem progressivamente diminuindo até 0,5005, caso a sua tendência inicial para o monocaixismo seja particularmente forte. Além disso, se neste nível pré-teórico as nossas intuições tendem a vacilar quando os valores das duas caixas se vão aproximando, então não me parece descabido acreditar que essas intuições também vacilarão à medida que a fiabilidade do previsor vai progressivamente diminuindo.

Por outro lado, se a discussão dos meta-argumentos estiver a ser tida ao nível do confronto entre teorias da decisão alternativas – ou entre interpretações diferentes das contrafactuais envolvidas nos argumentos-objecto -, então nenhum monocaixista se deixará dissuadir pelo meta-argumento bicaixista, mesmo quando a fiabilidade do previsor é de 0,5005. Neste nível, o que é importante para a discussão é que o meta-argumento monocaixista contém informação indispensável quanto à probabilidade dos estados do mundo relevantes, ao contrário do que acontece com o meta-argumento bicaixista, que nada nos diz quanto àquilo que provavelmente irá acontecer (ficarmos ricos, ou não), dada a nossa escolha. O meta-argumento monocaixista, convém repetir, ao contrário da sua contra-parte bicaixista, não recorre à noção de causalidade e por esse mesmo motivo, não pressupõe já qualquer tipo de resolução para a referida vagueza.

Convém relembrar, por último, que esta solução monocaixista é obtida no contexto da teoria causal. Ou seja, não se está aqui a defender, ao contrário de Horgan (1981), que se deve adoptar sempre a interpretação retroactiva de contrafactuais para efeitos do cálculo da utilidade esperada. Assim, para sustentar tal solução, há que ter confiança na solidez

da teoria e testá-la contra possíveis contra-exemplos. Tal será o empreendimento levado a cabo na terceira parte.

9. Uma solução evidencialista para o dilema do prisioneiro

9.1. Tipos de jogos

À primeira vista, O PN parece ter apenas um interesse teórico. Numa dada situação hipotética, bastante remota e irrealista, o PMUCE recomenda uma acção que, de um certo ponto de vista, não parece ser a acção racional. Daí a necessidade de revisão da teoria e o debate entre evidencialismo e causalismo. Contudo, essa situação hipotética não é tão remota quanto possamos pensar. Na verdade, existe uma versão do problema no contexto da teoria económica (Clark 2002). Os economistas acreditam que os cidadãos são capazes de prever, com uma elevada taxa de sucesso, as alterações da situação económica. Por exemplo, se os governos decidem introduzir dinheiro na economia, com o intuito de diminuir o desemprego, os cidadãos prevêem-no e, de uma maneira involuntária – pois não têm intenção de aumentar o desemprego - agem de um modo que contraria as vantagens da medida, dando assim origem à inflação. Neste cenário, os cidadãos fazem o papel do previsor e o governo faz o papel do decisor. Tal como o previsor é extremamente fiável a prever o conteúdo caixa opaca, os cidadãos são extremamente fiáveis a prever o rumo da situação económica. O problema de decisão pode ser ilustrado pela seguinte matriz:

	Prevêem introduzir	Prevêem constante
Introduzir	inflação	cai desemprego
Constante	recessão	sem alteração

O governo tem à sua disposição duas acções: introduzir dinheiro na economia ou manter o fluxo de dinheiro constante. Constata-se que introduzir dinheiro na economia é a acção dominante, pois seja qual for a previsão, introduzir dinheiro tem sempre melhores

consequências do que a sua alternativa: se os cidadãos prevêem a introdução, resulta daí a inflação, a qual é preferível à recessão; se os cidadãos prevêem que o fluxo de dinheiro é constante, então o desemprego cairá, o que é preferível à situação inicial. Como fariam notar os bicaixistas, a acção do governo não tem qualquer eficácia causal na formação das expectativas dos cidadãos. Por sua vez, os monocaixistas argumentarão que, tal como no PN, existem apenas duas consequências possíveis, a inflação ou nenhuma alteração, pois a probabilidade condicional das duas previsões, dadas as acções respectivas, é suficientemente elevada.

Parece existir um consenso entre os economistas de que a acção racional consiste em introduzir dinheiro na economia, com base exactamente nos mesmos argumentos que os bicaixistas utilizam: a previsão dos cidadãos já está feita e nada do que os governos possam fazer pode alterá-la. Se essa acção tem obtido, ou não, os resultados desejáveis, trata-se de uma questão empírica para a qual não parece existir uma resposta suficientemente clara.[65] Além disso, podem existir dúvidas quanto a saber se este exemplo constitui realmente um PN. Não é absolutamente certo que a introdução de moeda na economia faça realmente parte do tipo de coisas acerca das quais os cidadãos em geral têm expectativas.

Mas o PN pode ser ainda mais comum do que este exemplo nos leva a crer. Mais precisamente, o PN pode ser tão comum quanto o dilema do prisioneiro. A relação entre estes dois puzzles tem sido realçada por vários autores e muitos defendem que os dois são um e o mesmo problema, nomeadamente, que cada um dos prisioneiros do dilema enfrenta isoladamente um PN. Se esta identidade for verdadeira, poder-se-á construir um novo argumento para defender uma ou outra solução para o PN, adoptando para este a solução encontrada para o dilema. Ou seja, a acção dominante no PN é a solução racional se, e somente se, a acção dominante no dilema for a solução racional, o mesmo se verificando para a acção contrária. Esta é a possibilidade que pretendo explorar, estando em causa saber se o dilema dos prisioneiros é, de facto, um PN. Para esse efeito é necessário familiarizar-nos com o primeiro, o que, no contexto da teoria dos jogos, consiste em saber que tipo de jogo é o dilema dos prisioneiros.

Considere-se o jogo conhecido em língua inglesa por *chicken* (algo como o nosso *medricas*): os dois jogadores encontram-se dentro de automóveis alinhados na mesma direcção, mas em sentido contrário e a uma certa distância um do outro. Cada um acelera

[65] Ver, acerca deste PN económico, John Broome (1989: 220-222).

ferozmente contra o adversário e o primeiro a desviar o seu automóvel, evitando a colisão, passa a ser o *chicken*.

A seguinte matriz representa a estrutura do jogo. As linhas correspondem às opções do jogador A, as colunas às opções do jogador B, e os números às consequências da decisão tomada, tendo em conta a decisão do adversário:

		B	
		Desviar	Centro
A	Desviar	(3,3)	(2,4)
	Centro	(4,2)	(1,1)

A análise da matriz mostra-nos várias coisas importantes. Primeiro, ambos os jogadores têm uma escala de preferências similar: ambos preferem, acima de tudo, manter a direcção e deixar o seu adversário desviar-se, pois é essa a preferência que corresponde à vitória no jogo (adquirir uma reputação de bravura). A preferência imediatamente inferior consiste em ambos desviarem simultaneamente os carros e evitarem a colisão. Por último, resta a consequência que pode ser desastrosa para ambos, a colisão frontal. Segundo, ambos os jogadores, ao tomarem uma decisão, têm de ter em conta aquilo que pensam que o seu adversário irá fazer. Chama-se a isto escolher uma estratégia. Terceiro, duas das consequências possíveis, que correspondem à escolha de estratégias diferentes, possuem uma característica que se destaca: nenhum deles consegue sair-se melhor do que o adversário mudando unilateralmente de estratégia. Ou seja, se eu escolher a estratégia que consiste em desviar-me, e o meu adversário escolher a estratégia que consiste em seguir em frente, eu não consigo aumentar os meus ganhos mudando de estratégia, a menos que ele também o faça e vice-versa. Aos pares de resultados que correspondem a estas duas consequências, nos cantos superior direito e inferior esquerdo da tabela, chamam-se pontos de equilíbrio. A questão que parece colocar-se para ambos os jogadores é a seguinte: haverá algo como uma solução para este jogo? Em caso de resposta afirmativa, em que poderá consistir tal solução? A resposta seria algo como a existência de uma estratégia, disponível para cada um dos jogadores, tal que, seja o que for que o adversário faça, a escolha dessa estratégia permite alcançar sempre um melhor

resultado do que a escolha da estratégia contrária. Neste exemplo, não existe uma tal estratégia. Um tipo de jogo em que, pelo contrário, se verifica a existência um tal par de estratégias é o dilema do prisioneiro.

Dois membros de um gangue são presos por assalto à mão armada e colocados em celas separadas, sem possibilidade de comunicação entre si. A acusação oferece a cada um dos prisioneiros o seguinte acordo: 'se confessares e o teu parceiro não confessar, ficas livre e ele leva a pena máxima. Se ambos confessarem, são os dois presos, mas nenhum de vós leva a pena máxima. Se nenhum confessar, trataremos de vos condenar a ambos por evasão fiscal'. A tabela de decisão que representa o dilema enfrentado por cada um dos prisioneiros é apresentada abaixo. Os valores numéricos das consequências correspondem aos anos de prisão que cada um receberá, consoante a combinação das estratégias seguidas por si e pelo adversário:

		B	
		confessar	não-confessar
A	confessar	(5,5)	(0,10)
	não-confessar	(10,0)	(1,1)

O dilema do prisioneiro partilha algumas características com o jogo anterior. Primeiro, é notório que, mais uma vez, as escalas de preferência dos jogadores são idênticas, e quando ambos escolhem estratégias distintas, os pares de ganhos e perdas que lhes correspondem são simétricos. Ambos preferem a liberdade à pena menor e esta à pena maior. Segundo, a decisão de cada um é também ela tomada de acordo com a expectativa acerca do comportamento do adversário.

No contexto da teoria matemática dos jogos, um dos dados de partida é que os jogadores são agentes racionais. Significa isto que a sua única motivação consiste em maximizar os seus ganhos e minimizar as suas perdas, consoante as possibilidades que cada situação oferece. Além disso, supõe-se que, no caso de existir uma solução para o jogo, ou seja, uma estratégia que garante a cada um dos jogadores a obtenção do objectivo referido, ambos irão escolhê-la.

Apesar destas semelhanças, há um aspecto em que os jogos divergem entre si. Se, no primeiro, tínhamos dois pontos de equilíbrio, neste temos apenas um. Se o jogador A confessar, tentando alcançar o prémio maior (a sua liberdade), duas coisas podem acontecer: 1) se o jogador B também confessar, ambos recebem cinco anos de prisão; 2) se o jogador B não confessar, o jogador A obtém a sua liberdade. Por outro lado, se o jogador A não confessar, duas coisas podem acontecer: 3) se o jogador B também não confessar, ambos recebem 1 ano de prisão; 4) Se o jogador B confessar, o jogador A será condenado à pena máxima.

Os factos são estes: se ambos jogam para o prémio máximo, ambos recebem apenas o terceiro resultado mais desejado, cinco anos de prisão; se um ou outro jogar para o segundo prémio, duas coisas podem acontecer: ou ganha o prémio máximo ou expõe-se a ser explorado pela estratégia do adversário, sujeitando-se a uma condenação à pena máxima. Verifica-se, assim, que o único ponto de equilíbrio é o que corresponde à escolha da estratégia que consiste em confessar, pois se algum deles alterar unilateralmente a sua decisão, alcançará um pior resultado. O mesmo não acontece quando ambos permanecem em silêncio, pois se um dos dois alterar a sua decisão, o seu adversário ficará exposto ao pior dos resultados possíveis.

Assim, ao contrário do *chicken*, este jogo parece ter aparentemente uma solução: existe uma estratégia que, independentemente da estratégia adoptada pelo adversário, alcança sempre melhores resultados do que a estratégia contrária. Chama-se a uma tal estratégia ‘dominante’ e ‘dominada’ à sua contrária. Um dos mais elementares critérios para determinar a racionalidade de uma decisão é o seguinte: nunca adoptar uma estratégia dominada. Logo, na suposição da racionalidade de ambos os jogadores, ambos escolherão a estratégia dominante, vendo-se ambos condenados a permanecer na prisão por cinco anos.

O dilema é gerado pelo seguinte aspecto do problema: se é racional escolher uma certa estratégia, esperar-se-ia que essa estratégia maximizasse os ganhos do agente, pois ao definirmos os agentes como racionais foi esse o critério utilizado. Contudo, no caso do dilema do prisioneiro, a estratégia dominante não alcança esse objectivo, pois os jogadores poderiam ter alcançado um resultado melhor para ambos, caso ambos tivessem escolhido a estratégia dominada. A estratégia dominante não é, portanto, a mais eficiente, i.e., não é *Pareto-Óptima*.

Para compreendermos este conceito, considere-se uma situação em que será distribuída uma certa quantidade de bens a dois indivíduos. Atribuamos o valor 100 à totalidade dos bens. Qualquer distribuição de bens que conduza à obtenção de parcelas que, após somadas, resultam em valor 100, será uma distribuição eficiente. Assim, tanto uma distribuição de 91-9, como uma de 50-50 são *Pareto-Óptimas*. Uma distribuição, por exemplo, de 49-49 já não será *Pareto-Óptima*, pois ambos os indivíduos podem ver a sua situação melhorada sem que o outro tenha qualquer prejuízo, independentemente dos restantes bens serem distribuídos equitativamente ou de um deles obter bens de valor 51. Assim, no dilema do prisioneiro, se entendermos os anos de prisão como prejuízos a serem distribuídos, o resultado que minimiza para ambos os jogadores esses prejuízos é aquele em que ambos recebem apenas um ano de prisão.

Do dilema enfrentado surge um outro aspecto distinto. Se o *chicken* assumia um carácter estritamente competitivo, pois qualquer outro resultado para além da vitória poderia ser considerado subjectivamente como uma derrota, o dilema do prisioneiro apresenta um aspecto cooperativo: se ambos os jogadores cooperarem, poderão obter um resultado melhor para ambos do que aquele que é obtido caso ambos escolham a estratégia dominante. Tal resultado pode ser subjectivamente considerado por ambos como uma semi-vitória, pois apesar de nenhum deles ter obtido o primeiro prémio, ambos evitaram o pior e escaparam à fatalidade, obtendo o segundo melhor resultado da sua lista de preferências, minimizando, assim, as suas perdas. É por este motivo que muitas vezes as estratégias dos agentes são classificadas não como ‘confessar/não-confessar’, mas sim como ‘não-cooperar/cooperar’, respectivamente.

Ao procurar-se uma solução para o dilema do prisioneiro através da opção pela estratégia dominante, concluiu-se que, afinal, essa solução apresenta um carácter insatisfatório; logo, o argumento que lhe é favorável não é tão persuasivo quanto à primeira vista poderia parecer. Do ponto de vista da teoria dos jogos, qual é a razão técnica pela qual não podemos atribuir ao dilema do prisioneiro uma solução cooperativa? Tal acontece porque o dilema não é um jogo de soma-zero. Ou seja, não sendo um jogo estritamente competitivo, os ganhos de um jogador não correspondem necessariamente a perdas equivalentes por parte do seu adversário. Pelo contrário, num jogo de soma-zero não se gera, nem se perde, qualquer riqueza. Um exemplo claro é o poker. Começa-se com uma determinada quantia e chega-se ao final com a mesma quantia. A única diferença é que no final esta mudou de mãos ou se encontra distribuída de maneira diferente da inicial.

Em suma, se estivermos perante um jogo de soma-zero, é realmente possível encontrar uma solução para o mesmo. Considere-se, para o efeito, a seguinte tabela de decisão (Davis 1997: 12):

		Adversário			
		I	II	III	IV
	A	-3	17	-5	21
Nós	B	7	9	5	7
	C	3	-7	1	13
	D	1	-19	3	11

Os valores correspondem às nossas perdas e ganhos, sendo que, para o nosso adversário, os valores são inversos aos nossos. O primeiro passo para encontrar a solução de um jogo de soma-zero é aplicar o princípio da eliminação das estratégias dominadas.

Os valores devem ser interpretados da seguinte maneira: se nós escolhermos a estratégia A e o nosso adversário a estratégia II, teremos um ganho de 17. Se escolhermos a estratégia A, mas o nosso adversário escolher a estratégia III, teremos uma perda de 5. A simples análise da tabela mostra-nos que nenhuma das nossas estratégias é dominada por qualquer outra. Contudo, ao considerarmos as estratégias do adversário, vemos claramente que as suas estratégias I e III dominam as estratégias II e IV. Assim, podemos eliminar as colunas respectivas, juntamente com os ganhos e perdas correspondentes. Voltando a analisar a tabela, vemos agora que a nossas estratégias A, C e D são dominadas pela estratégia B. Logo, resta ao adversário escolher a estratégia III, pois só assim conseguirá minimizar as suas perdas. Fica, pois, encontrado o ponto de equilíbrio deste jogo.

Como podemos, então, encontrar pontos de equilíbrio? Primeiro, suponhamos que temos dois jogadores, A e B, e que B conhece as estratégias à disposição de A. Como B irá escolher o valor *mínimo* de qualquer linha (ou coluna) que A escolherá, então A irá escolher uma estratégia que obterá o *máximo* de entre estes valores *mínimos*. Assim, este valor é chamado de *maximin*. Segundo, suponhamos que A conhece as estratégias à disposição de B. Como A irá escolher o valor máximo de qualquer coluna (ou linha) que B escolherá, então B irá escolher uma estratégia que obterá o mínimo de entre estes valores *máximos*. Assim, este valor é chamado de *minimax*. Quando, num jogo, os valores

maximin e *minimax* são idênticos, encontramo-nos perante um ponto de equilíbrio e um respectivo par de estratégias de equilíbrio. O ganho (e respectiva perda) associado a um ponto de equilíbrio corresponderá ao valor máximo na respectiva coluna (ou linha) e ao valor mínimo na respectiva linha (ou coluna). Quando é encontrado um ponto de equilíbrio num jogo de duas pessoas com soma-zero, considera-se encontrada a solução (ou o valor) desse jogo. No caso acima, o valor do jogo é 5.[66]

O argumento a favor da escolha da estratégia de equilíbrio num jogo de duas pessoas de soma-zero resulta, assim, claro. Primeiro, o jogador irá obter, pelo menos, o valor do jogo; segundo, o jogador irá impedir que o seu adversário ganhe mais do que o valor do jogo; terceiro, como o jogo é de soma-zero, cada jogador, na suposição da sua racionalidade, estará motivado para minimizar as suas perdas e maximizar os seus ganhos e, como tal, a jogar a sua estratégia dominante. Significa isto que o carácter satisfatório das soluções que a teoria dos jogos apresenta reside na natureza estritamente competitiva dos jogos de soma-zero. Como sabemos, o dilema dos prisioneiros não só não é um jogo de soma-zero, como incorpora também um elemento cooperativo.

[66] Uma exposição da demonstração do teorema minimax – segundo o qual pode ser atribuído um valor *V* a todos os jogos finitos de soma-zero - pode ser encontrada em Luce e Raiffa (1957).

9.2 Cooperar ou esperar para ver?

À luz dos pressupostos fundamentais da teoria dos jogos - que existem expectativas mútuas de racionalidade e informação completa - este resultado não pode deixar de ser incómodo. Tal parece resultar da seguinte constatação: dois agentes 'irracionais', que decidam ambos cooperar, obterão um resultado melhor do que o resultado obtido por dois agentes racionais. Embora esta seja a raiz do dilema, ou o seu aspecto trágico, será que a racionalidade tem necessariamente de ser garantia da obtenção do melhor resultado?

Lawrence H. Davis (1977) sugere um exemplo que ilustra esta infelicidade, ou aspecto trágico, aliado ao exercício da racionalidade. Suponha-se que um agente é forçado a escolher entre passar uma hora com uma cobra venenosa ou passar uma hora com um leão. A escolha racional parece ser a cobra, pois apesar de existir uma enorme probabilidade de sermos mordidos, podemos sempre contar com o antídoto. Por outro lado, ser mordido pelo leão é garantia de morte certa. No entanto, o leão pode estar a dormir e escusávamos de passar pelas dores excruciantes associadas à mordida da cobra. Do mesmo modo, no dilema do prisioneiro, a opção pela escolha dominante parece equivaler à escolha da cobra: se o outro prisioneiro decidir não confessar, e nós fizermos o mesmo, escusamos de cumprir uma pena 'excruciante' de 5 anos. Mas o que transforma o dilema do prisioneiro num verdadeiro paradoxo, para a solução do qual existem dois argumentos contrários, é o pressuposto de que não estamos a lidar com um agente irracional ou imprevisível, como uma cobra ou um leão, mas com outro agente racional igual a nós. Ou seja, temos de agir de acordo com o que esperamos que a sua racionalidade lhe recomenda. Mas isto é algo acerca do qual não temos, aparentemente, quaisquer certezas. Ou, mesmo que as tenhamos, o problema é que, se ele agir 'racionalmente', o resultado mútuo será sub-óptimo.

Abandonando o argumento e o raciocínio favorável à escolha dominante, cairemos num tipo de raciocínio contaminado por uma inevitável circularidade. Para sabermos qual é a alternativa racional, colocando-nos na posição de um dos jogadores, precisamos de saber com segurança quais são as consequências dessas alternativas. Mas, para sabermos quais são essas consequências, precisamos de saber o que o outro irá fazer. O que o outro irá fazer depende daquela que para ele é a alternativa racional. Mas, o que para ele é racional fazer, depende de quais são as consequências das suas alternativas. Para ele saber quais

são essas consequências, precisa de saber o que nós iremos fazer. Para sabermos o que iremos fazer… E assim por diante.

Haverá maneira de sair desde círculo? Afinal, sabe-se que ambos os prisioneiros são racionais e que, na suposição da sua comum racionalidade, ambos optarão pela mesma alternativa. Essa alternativa só pode ser aquela que maximiza a sua utilidade. Davis (1985: 53) apresenta uma versão do argumento favorável à cooperação, incorporando nas premissas (1) e (2) os princípios necessários para se fugir à circularidade:

(1) Uma alternativa *X* encontra-se racionalmente prescrita para um agente *y*, se *y* sabe que existem apenas duas consequências possíveis *m* e *n*, tal que se *y* fizer *X*, a consequência é *m*, se *y* não fizer *X* a consequência é *n*, e *m* é melhor do que *n* (segundo o juízo de *y*).
(2) Cada prisioneiro sabe que ambos optarão pela alternativa racionalmente prescrita.
(3) Cada um sabe que uma alternativa se encontra racionalmente prescrita para um deles, apenas se também se encontrar racionalmente prescrita para o outro.
(4) Cada um sabe que manterá o silêncio, apenas se o outro o fizer, e que confessará, apenas se o outro o fizer.
(5) Cada um sabe que, se o silêncio for racionalmente prescrito, a consequência será (1,1); e que, se confessar for racionalmente prescrito, a consequência será (5,5).
(6) Cada um sabe que (1,1) e (5,5) são as únicas consequências possíveis.
(7) Cada um sabe que ambos sabem que (1,1) é melhor do que (5,5).
(8) Logo, manter o silêncio é a alternativa racionalmente prescrita para cada um.

A estratégia do argumento de Davis consiste em raciocinar a partir da suposição de que existe uma alternativa racionalmente prescrita, mesmo sem se saber ainda que alternativa é essa. Para que o argumento possa prosseguir é, então, necessário pressupor-se não só que existe essa alternativa, mas que ambos sabem que ambos irão escolhê-la, ou seja, que (2) é verdadeira.

A questão consiste, assim, em averiguar a plausibilidade das premissas (1) e (2). Quanto à primeira, suponha-se que não existe uma solução racional para o dilema. Nesse caso, deixam de vigorar os pressupostos acerca da racionalidade mútua dos agentes, deixando de haver qualquer segurança acerca do que o outro irá fazer. É como se estivéssemos novamente a jogar o dilema com uma cobra ou um leão. Como foi visto, nessa situação,

a alternativa racional consiste em optar pela acção dominante, o que, por sua vez, contraria o pressuposto inicial de que não existia uma acção racional.

Por seu lado, é difícil determinar se, fora do contexto idealizado da teoria, a crença na verdade de (2) é tão implausível quanto, à primeira vista, poderá parecer, implausibilidade essa que poderá roubar ao dilema alguma relevância prática. Com efeito, a definição das circunstâncias tem de ser suficientemente minuciosa para se poder determinar se estamos, ou não, perante um verdadeiro dilema do prisioneiro. Costuma-se, por vezes, incluir nessa definição a ideia de que não existe "honra entre bandidos". Esta circunstância é bastante relevante, pois se existisse um código de conduta pré-definido, nomeadamente, uma certeza de punição caso se agisse contra as regras desse código, ter-se-ia de atribuir diferentes *payoffs* às consequências. Nesse caso, a natureza do problema iria provavelmente alterar-se, deixando de se poder classificar como um dilema do prisioneiro. Mas, independentemente de acharmos a premissa (2) mais ou menos realista, o importante é saber se esta torna o dilema semelhante ao PN. Com efeito, parece existir uma relação próxima entre o pressuposto da racionalidade mútua dos agentes e um aspecto essencial da caracterização do PN, a elevada taxa de sucesso do previsor. Se isto for suficiente para tornar o dilema e o PN semelhantes, então a suposta implausibilidade de (2) não retira qualquer importância teórica ao dilema, muito pelo contrário.

Um primeiro ponto em comum entre estes dois problemas é que em ambos se verifica um apelo controverso à dominação. Pressupondo, como em (2), a semelhança entre os prisioneiros, podemos redefinir o dilema num confronto entre esse apelo à dominação e o PMUCE. Colocando-nos na perspectiva de um dos jogadores do dilema,

UCE (confessar) = $u(5,5) \times pr((5,5)|\text{eu confesso}) + u(0,10) \times pr((0,10)|\text{eu confesso})$.

Como $pr((0,10)|\text{eu confesso}) = 0$, pois ambos faremos a mesma coisa, e como, pelo mesmo motivo, $pr((5,5)|\text{eu confesso}) = 1$, então a utilidade de confessar é 5. Por outro lado,

UCE (cooperar) = $u(10,0) \times pr((10,0)|\text{eu coopero}) + u(1,1) \times pr((1,1) \mid \text{eu coopero})$.

Como $pr((10,0)|\text{eu coopero}) = 0$, pois ambos faremos a mesma coisa, e como, pelo mesmo motivo, $pr((1,1)|\text{eu coopero}) = 1$, então a utilidade de cooperar é 1. Como a utilidade é

aqui inversamente proporcional ao tempo de prisão, segue-se que a utilidade de cooperar é maior do que a utilidade de confessar.

Comparando as tabelas de decisão podemos constatar, desde logo, que ambos os problemas têm a mesma estrutura. Sejam quais forem os valores exactos da utilidade, desde que $x > y > z > 0$, estaremos sempre na presença de um dilema do prisioneiro,

	A_2	B_2
A_1	(y, y)	(0, x)
B_1	(x, 0)	(z, z)

Na matriz do PN, sabendo-se que a utilidade dessas consequências é inversa à do dilema, tal que $0 > z > y > x$, temos o seguinte:

	Caixa opaca vazia	1 Milhão na caixa opaca
Duas caixas	1000 = y	1.001.000 = 0
Caixa opaca	0 = x	1.000.000 = z

O segundo aspecto em comum consiste na independência causal que se verifica entre os estados do mundo e as acções disponíveis. No dilema, a minha decisão não tem qualquer influência causal sobre a decisão do meu oponente; no PN, a minha decisão não tem qualquer influência causal sobre o conteúdo da caixa opaca.

Finalmente, os dois problemas têm um terceiro aspecto em comum: os estados do mundo não são evidencialmente independentes das acções. A minha decisão de confessar é um sinal muito forte de que o meu oponente irá confessar, o mesmo acontecendo com a minha decisão de cooperar. No PN, a minha decisão de escolher as duas caixas é um sinal muito forte de que o previsor deixou a caixa opaca vazia, enquanto a minha decisão de escolher a caixa opaca é um sinal muito forte de que o previsor colocou um milhão na caixa opaca. Este terceiro aspecto é específico daqueles problemas que Sobel (1985) caracteriza como Problemas de Newcomb de Previsão Certa [*near-certainty*] e dilemas do prisioneiro de expectativa comum [*near-certainty*], nos quais, respectivamente, a taxa de sucesso do previsor é quase um, e a probabilidade de o outro prisioneiro agir como eu é quase um.

Temos, assim, justificada a relação de simetria, acima aludida, entre a elevada taxa de sucesso do previsor e o pressuposto da racionalidade comum dos prisioneiros. Estes três factores são responsáveis por tornar os dois problemas estruturalmente idênticos.
Consideremos uma nova matriz, comum aos dois problemas:

	E_1	E_2
A_1	$(A_1 \wedge E_1)$	$(A_1 \wedge E_2)$
A_2	$(A_2 \wedge E_1)$	$(A_2 \wedge E_2)$

Dada a dependência epistémica dos estados relativamente às acções, temos, para ambos os problemas, as seguintes igualdades:

$pr(E_1|A_1) \cong 1 \cong pr(E_2|A_2)$,
$pr(E_2|A_1) \cong 0 \cong pr(E_1|A_2)$.

De onde se seguem:

$pr\,(A_1 \wedge E_1) \cong 1 \cong pr(A_2 \wedge E_2)$,
$pr(A_1 \wedge E_2) \cong 0 \cong pr(A_2 \wedge E_1)$.

Sendo os dois problemas estruturalmente idênticos, a acção que é racionalmente prescrita num deles tem de corresponder à acção equivalente no outro. A acção de cooperar, que é a acção racionalmente prescrita na suposição da racionalidade comum dos agentes, equivale à acção de escolher a caixa opaca. Logo, a acção de escolher a caixa opaca é a acção racionalmente prescrita, na suposição de que o previsor é quase infalível. Como afirma Sobel, cada Dilema do Prisioneiro de Expectativa Comum é um Problema de Newcomb de Previsão Certa. Por outras palavras, ambos os problemas têm uma solução evidencialista.
David Lewis (1979b), argumentando a favor da identidade entre os dois problemas, coloca as coisas em termos que fazem realçar essa simetria:

1. São-me oferecidos 1.000 Euros e outros 1.000 ao outro prisioneiro; é pegar ou largar.
2. Talvez venha também a receber 1.000.000, mas venha isso a acontecer, ou não, tal é causalmente independente do que eu farei.
3. Receberei o milhão extra se, e somente se, o outro recusar os seus mil.

O dilema e o PN são semelhantes quanto aos pontos 1 e 2. O PN difere apenas quanto ao ponto 3, que deve ser reformulado, segundo Lewis, da seguinte maneira:

3'. Receberei o milhão extra se, e somente se, for previsto que eu recusei os mil.

3''. Receberei o milhão extra se, e somente se, um qualquer processo predictivo (que pode decorrer antes, durante ou depois da minha escolha) tem como resultado uma previsão de que não escolho ficar com os mil.

A meu ver, esta maneira de colocar o problema é, desde logo, enviesada, adequando-se perfeitamente a quem defende a acção dominante, que é o caso de Lewis.[67] Isto porque em 3' e 3'' está implícito que 'eu confesso e espero para ver'; transpondo para o PN, ficamos com: 'eu escolho as duas caixas e espero que o previsor erre'.
A meu ver, a simetria entre os dois problemas pode também ser apresentada nos seguintes termos:

a. São-me oferecidos 1.000 Euros (5 anos de prisão) e outros 1.000 ao outro prisioneiro; é pegar ou largar (confesso ou não confesso).
b. Posso aceitar os mil, ou recusá-los e ganhar um milhão (1 ano de prisão); mas, consiga ou não ganhar o milhão, tal é causalmente independente do que eu farei (aceitar ou não aceitar os mil).
c. Ganharei o milhão (1 ano de prisão) se, e somente se, o outro recusar os seus mil (o outro não confessa).
c'. Ganharei o milhão (1 ano de prisão) se, e somente se, for previsto que recusei os mil.

[67] Lewis (1979b: 236) usa os seguintes termos: '(2) Perhaps also I will be given a million (…)'.

Um aspecto importante da caracterização de Lewis é que parece excluir do cenário duas consequências da acção de cooperar: ganhar 1.000.000 Euros (extremamente provável) ou 0 (extremamente improvável). Nenhuma referência a estas consequências é feita na caracterização do problema. Para um defensor da estratégia anti-cooperação, e do bicaixismo, como Lewis, a segunda parte do ponto b é claramente decisiva para a sua tomada de posição, o que não é de todo novidade. Para um defensor da cooperação e do monocaixismo, apesar da crença inabalável em b, dois novos pontos da maior importância devem ser incluídos na caracterização:

d. Se eu recusasse os 1000, então (probabilisticamente) o outro recusaria os 1000; e se eu aceitasse os mil, então (probabilisticamente) o outro aceitaria os mil.

d'. A probabilidade de o outro prever que eu não aceitaria os mil, caso eu não os aceitasse, é quase 1.

Se os pontos 2 e b realçam a independência causal entre acções e previsões, d e d' realçam a dependência epistémica entre as mesmas. Ou seja, a *pr*(eu recuso ≫ o outro recusa) ≠ *pr*(outro recusa|eu recuso) = 1. Dada esta desigualdade, como já sabemos, a utilidade causal esperada de recusar os mil, e a utilidade condicional esperada de recusar os mil, irão diferir. Ou, transpondo o conflito para o seio da teoria causal, como fizemos no PN, o valor da utilidade causal esperada, com as contrafactuais interpretadas de maneira standard, e o cálculo da mesma, com as contrafactuais interpretadas de maneira retroactiva, irão diferir. Esta sim, é uma maneira imparcial de caracterizar o problema.
Resumindo, para tornar o dilema do prisioneiro num PN, temos de nos assegurar que o primeiro se qualifica como um dilema de expectativa comum [*near-certain*]. A forma mais apropriada de garantir que assim seja, consiste em identificar o outro prisioneiro como uma réplica quase perfeita de nós próprios, como um espelho onde a nossa acção será reflectida. Este aspecto do problema toma precedência sobre a verdade de 2 ou b, e a escolha racional passa por seleccionar a acção que nos transforma em milionários ou, respectivamente, a estratégia que nos livra de quatro anos de prisão. A aceitação desta precedência também não é de todo novidade; como vimos atrás, a opção por uma interpretação retroactiva das contrafactuais, no caso do PN, corresponde a uma leitura das circunstâncias que atribui o primeiro lugar de importância à infalibilidade do previsor ou,

respectivamente, à semelhança entre os prisioneiros, como no caso do argumento de Davis.

Com o intuito de contestar esta identidade entre os dois problemas, poder-se-á apontar o seguinte: temos um PN se, e somente se, a probabilidade de o previsor acertar for superior a 0,5005; contudo, um dilema em que a probabilidade com que cada um dos prisioneiros prevê a acção do outro seja baixa, deixa de apresentar um aspecto paradoxal; nestes casos, como vimos, é racional confessar (o leão pode estar acordado).

Temos, no entanto, de nos lembrar que um dilema deste último tipo não é um PN; como vimos acima, nem todos os dilemas são PN's, apenas o são os dilemas de expectativa comum; se removermos do dilema a cláusula da dependência epistémica, deixamos de ter identidade entre os dois problemas e, como tal, o argumento favorável ao monocaixismo que se ensaiou não sai beliscado. Ainda assim, não se pode negar que estamos sempre na presença de um PN, mesmo quando a taxa de sucesso do previsor é tão baixa quanto 0,5005; para continuarmos a defender a identidade entre os dois problemas, temos de qualificar um PN deste tipo como um PN de previsão certa [*near-certain*], o que poderá parecer algo forçado. Contudo, não vale a pena termos este escrúpulo: desde que, num PN, tenhamos um confronto entre princípios de racionalidade (ou entre interpretações da utilidade esperada), e desde que admitamos que temos dependência epistémica, mesmo com uma taxa de sucesso tão baixa, então podemos legitimamente qualificar um tal PN como sendo de 'previsão certa' *near-certain*. Em suma, todos os dilemas de expectativa-comum são PN's de previsão certa.

Para quem acredita que a cláusula da expectativa mútua de racionalidade é extremamente implausível, pertencendo apenas ao domínio da teoria idealizada dos jogos, a importância prática tanto do dilema, quanto do PN, vê-se algo diminuída. Talvez venha daí a relutância de Jeffrey quanto à necessidade de ter em conta o PN para efeitos de revisão do seu modelo:

"I see Newcomb's Problem as a prisioner's dilema for space cadets: a secular, sci-fi sucessor to the problems of predestination that exercised such thinkers as Jonathan Edwards (1703-58)". (Jeffrey 1983: 25).

Mas, tal como se viu no início, talvez existam mais PN's do que pensamos. Existem dados empíricos muito interessantes que mostram que o emprego de estratégias cooperativas

para lidar com versões iteradas do dilema conduzem a resultados claramente mais eficientes do que aqueles que resultam do emprego de estratégias dominadoras, sem que seja necessário atribuir aos agentes outros motivos para além da satisfação do seu interesse próprio (Axelrod 1984). Contudo, para versões *one-round* do dilema, os dados são mais difícieis de interpretar. Por exemplo, o matemático, físico e cientista cognitivo Douglas Hofstadter (1983) decidiu convidar 20 amigos para jogarem entre si uma ronda do dilema, cada um deles contra cada um dos outros, amigos esses cuja inteligência não merecia contestação - 'You are very bright. So are the others. All about equally bright, I would say' (Hofstadter 1983: 14). Cada um deles deveria responder-lhe apenas com uma letra: *C* para cooperar (*cooperate*) e *D* para desertar (*defect*); além disso, cada um deveria ainda oferecer uma breve justificação para a sua resposta. Os *payoffs* eram os seguintes: 3$ para a cooperação mútua, 1$ para a deserção mútua, e, em caso de respostas diferentes, 0$ para o cooperador e 5$ para o desertor. A esperança de alguns jogadores era, obviamente, ser o único desertor, o que garantiria o prémio máximo de 95$, contra os 54$ (18 × 3$) de cada um dos cooperadores.

Os resultados, para Hofstadter, foram não só desapontantes, como também contrários à sua expectativa: 14 desertores e 6 cooperadores. As razões apresentadas pelos primeiros podiam, quase na sua totalidade, ser reconduzidas à escolha da acção dominante. Por exemplo, Robert Axelrod, que defendeu uma solução cooperativa para versões iteradas do dilema, desertou sem a mínima hesitação. Outro dos desertores respondeu simplesmente da seguinte maneira: 'Hofstadter, passa-me os meus 19$'! Este valor corresponde ao prémio mínimo, que cada um dos participantes receberia caso todos os vinte desertassem. Mas, pior ainda, de acordo com a perspectiva de Hofstadter, a maioria dos cooperadores justificou a sua escolha pelas *razões erradas*, nomeadamente, através do recurso a considerações de ordem moral.[68] Apesar destes resultados, Hofstadter defendeu a escolha cooperativa (evidencialista, se quisermos), utilizando o mesmo argumento apresentado acima por Lawrence H. Davis. Ou seja, partindo de uma expectativa mútua de racionalidade, e pressupondo que existe de facto uma solução racional para o dilema, restava aos participantes maximizarem os seus ganhos de acordo com os dados do problema:

[68] Dan Dennett, um dos cooperadores, justificou a sua escolha da seguinte maneira: 'I would rather be the person who bought the Brooklin Bridge then the person who sold it. Similarly, I'd feel better spending $3 gained by cooperating than $10 gained by defecting" (Hofstadter 1983: 27). Tal como afirmou outro dos cooperadores, o psiquiatra Charles Brenner, Dennett provavelmente não quis ver o seu nome incluído numa lista de desertores publicada numa revista internacional de grande circulação como a *Scientific American*.

‘Um qualquer número de pensadores idealmente racionais, colocados perante a mesma situação, e experimentando de igual modo a agonia que acompanha este tipo de raciocínio, chegará inevitavelmente à mesma conclusão, desde que a justificação última da sua escolha resida apenas no exercício da sua racionalidade. Caso contrário, a racionalidade seria algo de subjectivo, e não uma coisa objectiva, como é o caso da aritmética. […] Se aceitarmos isto, então teremos percorrido 90% do caminho. Todo o que necessitamos agora de saber é o seguinte: “Dado que todos vamos apresentar a mesma [solução], qual destas será a mais lógica? Ou seja, que mundo é melhor para o pensador racional *individual*: um mundo constituído apenas por cooperadores ou outro constituído apenas por desertores”? A resposta é imediata: “Obtenho 57$ se todos cooperarmos e obtenho 19$ se todos desertarem. Eu prefiro claramente 57$, logo este pensador racional particular prefere cooperar. Como eu sou um caso típico, *todos* os pensadores racionais devem preferir cooperar’ (Hofstadter 1983: 22).

Segundo Hofstadter, os resultados do jogo deveram-se ao facto de os participantes não terem levado a sério a expectativa mútua de racionalidade (a fiabilidade do previsor, se quisermos), a qual foi sublinhada quando no convite se mencionou o brilhantismo intelectual de todos os participantes. Ora, levar a sério esse aspecto não é exactamente o mesmo que efectuar uma simulação mental do raciocínio de cada um dos outros participantes. Por exemplo, eu posso ponderar as razões favoráveis a cada uma das escolhas e considerar que durante um terço do tempo que levo a reflectir eu desejo cooperar, e durante dois terços de tempo desejo desertar. Se me encarar como um participante típico, concluo então que dois terços dos meus pares vão desertar e, efectuando um cálculo simples, concluo que para maximizar os meus ganhos devo desertar. Contudo, este tipo de raciocínio levará sempre à mesma conclusão, independentemente do número de cooperadores que estimamos que venha a haver, de acordo com os resultados da nossa simulação. Segundo o argumento utilizado, isto não é suficiente, pois uma consideração adequada do pensamento estratégico deve ter em conta que, para além de idealmente racionais, todos os outros jogadores são também idealmente *super-racionais,* ou seja, cada pensador racional deve ter em conta que cada um dos outros pensadores racionais deve ter em conta que cada um dos outros pensadores racionais deve ter em conta que cada um dos outros pensadores racionais deve ter em

conta… e assim *ad infinitum*. Portanto, se levarmos a sério a ideia de que o nosso 'adversário' no dilema é uma réplica nossa, então devemos cooperar, o que nos permitirá escapar a este mecanismo recursivo. Pela mesma ordem de razões, no PN devemos escolher a caixa opaca.

Dá-se, assim, por encerrada a discussão do PN. Apresentaram-se dois argumentos favoráveis a uma solução monocaixista. O segundo foi discutido nesta secção. Já o primeiro resultou de uma análise dos prós e contras de teorias que divergem entre si quanto ao valor de verdade das contrafactuais. Concluiu-se que, para efeitos de atribuição de crenças, a adopção de uma semântica probabilística como a de Van Fraassen vai ao encontro daquilo que estamos dispostos a aceitar como verdadeiro ou falso. Esta semântica mostrou também ser favorável a uma certa interpretação de contrafactuais, designada como 'retroactiva', da qual depende a solução monocaixista. Contudo, dado que no decorrer desta análise se mostrou que a teoria causal da decisão é a mais adequada para lidar com problemas de decisão estruturalmente semelhantes ao PN, a solução monocaixista foi enquadrada no contexto desta teoria. Isto significa que, para efeitos do cálculo da probabilidade de contrafactuais, têm de se aceitar alguns pressupostos rejeitados pela semântica de Van Fraassen, nomeadamente, que, para qualquer proposição $A \gg B$, os mundos-A se dividem, de facto, entre aqueles em que $A \gg B$ é verdadeira e aqueles que $A \gg B$ é falsa; e, o que é altamente contestável, que para qualquer proposição contingente A existe um único mundo-A possível, no qual A é verdadeira, que é 'mais semelhante' ao mundo actual do que qualquer outro mundo-A possível em que A também é verdadeira.[69] A esta distinção na aplicação das duas semânticas, Van Fraassen e Stalnaker/Lewis, designou-se como uma divisão do trabalho no contexto da aplicação da teoria causal da decisão.

[69] Contudo, Lewis (1973) mostrou que o pressuposto em causa é insustentável. Numa grande quantidade de casos, existem vários mundos possíveis que ocupam simultaneamente o primeiro lugar da semelhança, sendo a consequente de uma contrafactual verdadeira nuns e falsa noutros. Contudo, penso que é seguro afirmar-se que quando Stalnaker (1972) apresentou a sua proposta, este pressuposto estava a ser considerado como uma condição necessária para se efectuar uma distribuição de probabilidades sobre mundos possíveis. Ou seja, para que a função *pr*($A \gg B$) possa ser bem definida, temos de ter a certeza de que a mesma obedece aos axiomas do cálculo de probabilidades. Tal será o caso desde que $A \gg B$ satisfaça, entre outras, a *lei do terceiro excluído condicional*, ou seja, desde que $(A \gg B) \vee (A \gg \neg B)$ seja uma verdade lógica. Assim, a fórmula de Stalnaker/Gibbard e Harper apenas faz sentido debaixo do pressuposto de que existe apenas um único mundo-A possível mais próximo do actual do que qualquer outro mundo-A possível. Lewis (1976) concebeu um método para calcular a probabilidade de contrafactuais, designado por *visualização* [*imaging*], que pode ser generalizado (Gardenfors 1988), de maneira a lidar com os casos em que o pressuposto acima não se verifica

Estando, assim, encontrada a interpretação correcta do princípio da maximização da utilidade esperada, torna-se agora necessário confrontá-la com as críticas e analisar algumas propostas de reformulação do princípio evidencial.

Parte 3 – A Teoria

10. Em defesa de Jeffrey

10.1. A receita do paradoxo

Em §6.1 considerámos um exemplo de um problema de decisão com a mesma estrutura do PN: o caso de Salomão e Batsheba. O argumento no contexto do qual esse problema foi apresentado consistia em tentar mostrar o seguinte: se o princípio da maximização da utilidade causal esperada fosse adequado para o resolver, então, dado que a sua estrutura é idêntica à do PN, o mesmo princípio deveria ser aplicado para resolver o PN. Como o princípio causal nos oferece aí a solução correcta, uma solução do PN que resulte da aplicação desse princípio deveria também ela estar correcta.

Após a consideração da análise semântica das contrafactuais, encontra-se em disputa, não a adequação da teoria causal – nomeadamente em casos como o de Salomão e Batsheba - mas a ideia de que o bicaixismo constitui a solução correcta para o PN. Ou seja, a disputa passa a dar-se no seio da teoria causal, entre interpretações diferentes das contrafactuais envolvidas, e já não entre a teoria causal de Gibbard e Harper e a teoria evidencial de Jeffrey.

Contudo, Ellery Eells apresentou uma defesa do princípio de Jeffrey, de acordo com a qual uma certa revisão da teoria evidencial permite acomodar casos com a mesma estrutura do PN. Se esta defesa for adequada, a vantagem da teoria evidencial consiste em poder dispensar o aparato conceptual da teoria causal. Como Eells afirma (1985: 195), para pôr a funcionar a teoria evidencial são apenas necessários os conectivos lógicos, as operações de adição e multiplicação, as proposições relevantes para o problema e os conceitos de probabilidade subjectiva e utilidade (ou *desejabilidade*, o termo utilizado por Jeffrey). A teoria causal, por seu lado, necessita de tudo isto e ainda do operador contrafactual. Temos, portanto, de ser capazes de atribuir probabilidades subjectivas não a simples proposições categóricas, que descrevem os estados do mundo da maneira como intuitivamente os compreendemos, mas sim a condicionais contrafactuais que exprimem relações de causalidade.

Como foi visto, o nosso entendimento do conceito de causalidade é imperfeito e sujeito a análises distintas: como eventos seguidos um ao outro no tempo, como uma simples

relação de dependência condicional, como uma relação de dependência contrafactual, etc. Mesmo adoptando esta última, a prova de que a nossa capacidade de lidar com o conceito de causalidade é imperfeita pode encontrar-se na vagueza inerente à determinação da verdade de proposições contrafactuais. Esta é a razão pela qual, segundo Eells, uma teoria evidencial que consiga acomodar os casos com uma estrutura idêntica à do PN leva vantagem sobre a teoria causal.[70]

Estes casos têm todos a mesma estrutura e existe uma fórmula para os produzir. Todos eles podem ser designados como 'Problemas de Newcomb de Causa Comum' (PNCC). Ao revelarmos esta estrutura, poderemos mais facilmente compreender como funcionam as relações de dependência probabilística e os motivos pelas quais a teoria evidencial oferece uma solução incorrecta. Um dos exemplos mais conhecidos deste tipo de problemas é designado como 'O Sonho do Fumador'. Os cientistas descobrem que, ao contrário do que se pensava, o cancro do pulmão não é provocado pelo consumo de tabaco, mas sim por um gene. Este gene não só é a causa do cancro, mas também é responsável por tornar os seus portadores extremamente vulneráveis à aquisição de vícios. Isto faz com que a probabilidade de alguém desenvolver cancro sendo fumador seja muitíssimo elevada, por comparação com a baixa probabilidade de desenvolver cancro, sendo não-fumador. Tudo isto é confirmado pelas estatísticas disponíveis.

Supondo que um determinado agente sente uma vontade irresistível de fumar e retira daí um grande prazer, o que deve ele fazer? Fumar ou não fumar? Como se pode facilmente constatar, fumar domina a opção contrária: fumar e ter cancro é melhor do que não fumar e ter cancro, e fumar e não ter cancro é melhor do que não ter cancro e não fumar. Não havendo qualquer relação causal entre fumar e o desenvolvimento da doença, fumar é claramente a opção racional. Contudo, dada a relação de dependência probabilística em causa, o princípio evidencial recomenda que não se fume.

Consideremos um outro exemplo de Brian Skyrms (1980: 129). O endurecimento das artérias, ao contrário do que se pensava, não é provocado pelo elevado consumo de

[70] Até 1985, a data do mais recente artigo de Eells em que o argumento favorável à revisão da teoria evidencial é apresentado, existiam duas outras vantagens desta última sobre a teoria causal. Em primeiro lugar, não existia um teorema da representação para a teoria causal. Dada a consequente ausência de uma relação 'testável' entre os axiomas da preferência e o cálculo da utilidade causal esperada, nenhuma garantia existia da aplicabilidade geral da teoria aos 'calculadores' desse tipo de utilidade. Em segundo lugar, a utilidade causal esperada de uma acção variava de acordo com a partição do espaço de possibilidades, o que não acontece com a teoria evidencial. Contudo, existem hoje teorias causais às quais isto já não se aplica, ou seja, teorias como a de Joyce (1999), para a qual existe um teorema da representação e em que a utilidade das acções não varia de acordo com a partição dos estados do mundo.

colesterol, mas sim por uma lesão na parede da artéria. Quando atinge um estado avançado, esta lesão começa a acumular o colesterol presente no sangue. Mais, sabe-se que a partir do momento em que a lesão se desenvolve, aumenta também o desejo de consumir alimentos com colesterol. Não se conhece a causa da lesão, nem o mecanismo através do qual ela se relaciona com a ingestão de colesterol. Sabe-se, no entanto, que o aumento da ingestão de colesterol atrasa o desenvolvimento da lesão e não tem qualquer efeito sobre a saúde vascular de quem não a tem.

Supondo que um determinado agente gosta de comer, todos os dias, ovos com bacon ao pequeno-almoço - em vez de cereais integrais - deve ele comer, ou não, os ovos? Tendo em conta que a acção de comer os ovos domina a acção de comer cereais – sofrendo ou não da lesão é sempre preferível comer os ovos a comer cereais - e sabendo-se que comer os ovos não provoca a lesão, nem o endurecimento das artérias, comer os ovos é claramente a acção racional. Contudo, a probabilidade condicional de se ter a lesão, dado um elevado consumo de colesterol, é muitíssimo elevada. Por seu lado, a probabilidade condicional de não se ter a lesão, dada uma alimentação livre de colesterol, é muitíssimo baixa. Portanto, o princípio evidencial recomenda que não se coma ovos.

Já sabemos que, em ambos os casos, o problema da aplicação do princípio evidencial reside no facto de as relações de correlação estatística não espelharem as relações causais realmente presentes. Mas, além disso, estes casos partilham algo mais em comum. Verifica-se em ambos a existência de um factor que se encontra fora do controlo dos agentes, façam estes o que fizerem, factor esse que é a causa comum das acções e das *supostas* consequências dessas mesmas acções. No caso do fumador, o gene é não só responsável pelo surgimento do cancro, mas também é a causa do desejo irresistível de fumar. No caso do colesterol, a lesão nas artérias é não só responsável pelo endurecimento das mesmas, mas também pelo desejo irresistível de comer alimentos ricos em colesterol. A estrutura causal de ambos os problemas pode ser, então, representada da seguinte maneira:

----- Cancro

Gene -

----- Fumar ---- Prazer de fumar e benefício para o doente

----- Endurecimento artérias

Lesão -

----- Consumo colesterol ---- Atrasa desenvolvimento lesão

Temos, assim, duas causas comuns (CC) e dois tipos de sintomas das mesmas: as acções sintomáticas (AS) e as consequências sintomáticas (CS). A estrutura geral de um PNCC pode ser representada pelo seguinte diagrama:

------ *CS*

CC -

------ *AS*

Numa tabela de decisão, as causas comuns correspondem aos estados do mundo; as acções sintomáticas, da presença da causa comum, correspondem às acções disponíveis; e as consequências sintomáticas, da presença da causa comum, correspondem aos valores que se encontram no interior da tabela.

Gibbard e Harper referiam-se à escolha da acção não-dominante – não fumar e não ingerir ovos - como se tratando da escolha de uma acção que, apesar de auspiciosa, não tinha qualquer eficácia na produção da consequência desejada. Skyrms refere-se à escolha dessas acções como se tratando de tentativas de eliminar as causas (o gene e a lesão), através de uma manipulação dos sintomas (ou das acções sintomáticas). O que temos, naturalmente, é que as acções sintomáticas constituem evidência da presença das causas comuns, tal como, naturalmente, um efeito constitui evidência da presença de uma causa. Contudo, neste ponto, não é inteiramente clara a maneira como o PN se encaixa nesta estrutura geral. Os estados do mundo – a correcção ou incorrecção da previsão – não parecem constituir causas de qualquer uma das acções disponíveis. Mas, se tornarmos o problema menos mágico e supusermos que existe uma causa comum à previsão e à acção, ficamos com um típico PN de causa comum. Esta possibilidade não é de todo implausível e já a ela se fez referência: existiriam dois tipos de personalidade, bicaixista e monocaixista, as quais determinariam as escolhas dos agentes e as quais o previsor consegue identificar com grande sucesso.

Ainda assim, a avaliação de uma personalidade, que um psicólogo efectua com base num raciocínio de tipo indutivo, não me parece poder ser caracterizada como uma relação de

causa-efeito. Portanto, de modo a classificar o próprio PN como um PNCC, a estrutura interna do problema, ou o modo através do qual uma certa causa comum afecta o previsor e a nossa escolha, terá talvez de permanecer misteriosa. Isto não altera, obviamente, a relevância teórica do problema, nem sequer a sua relevância prática, pois, como vimos, existe uma grande variedade de PN's.

Vejamos, então, qual é a receita para gerar PNCC (Eells 1985: 189). A seguinte tabela de decisão servir-nos-á de referência:

	¬E	E
A	(A∧¬E)	(A∧E)
¬A	(¬A∧¬E)	(¬A∧E)

Temos de encontrar um estado E e uma acção A, tal que,

u(A∧¬E) − u(A∧E) e u(¬A∧¬E) − u(¬A∧E),

têm ambas o mesmo sinal e são grandes quando comparadas, respectivamente, com

u(A∧¬E) – u(¬A∧¬E) e u(A∧E) – u(¬A∧E).

Seja A fumar e ¬A não-fumar, e E ter o gene e ¬E não ter o gene, e introduzam-se, por exemplo, as respectivas utilidades:

	¬E	E
A	10	1
¬A	9	0

Temos, assim, que (10 – 1) e (9 – 0) são bastante maiores do que, respectivamente, (10-9) e (1 – 0). Estes dados significam várias coisas. Em primeiro lugar, o estado do mundo E é, obviamente, muito pior para o agente do que o estado do mundo ¬E. Segundo, é também óbvio que A domina ¬A nas preferências do agente: seja qual for o estado do mundo, A tem sempre melhores consequências do que ¬A. Mas existe ainda outra relação entre estados e acções que a simples dominação não garante: seja qual for a acção

efectuada, existe um estado, neste caso ¬E, que é sempre melhor do que o outro. Esta relação é-nos garantida pelas desigualdades apresentadas acima, as quais constituem o primeiro ingrediente da receita. Quando estas desigualdades se verificam, podemos dizer, por exemplo, que o estado ¬E *dita* a acção A.

O segundo ingrediente irá esclarecer-nos da necessidade do primeiro. Nos PN's de causa comum, como o caso do fumador acima, o agente acredita que o estado E é causa da acção A e que o estado ¬E é causa da acção ¬A; logo, é plausível que o agente tenha as seguintes probabilidades condicionais subjectivas: $pr(A|E) > pr(\neg A|E)$ e $pr(\neg A|\neg E) > pr(A|\neg E)$. Como a relação de dependência probabilística é simétrica, temos o segundo ingrediente da receita, caracterizado pelas seguintes probabilidades do agente:

$pr(E|A) > pr(E|\neg A)$ e $pr(\neg E|\neg A) > pr(\neg E|A)$.

Como é fácil de constatar, quando estes dois ingredientes estão presentes, o princípio da dominação e o princípio evidencial aconselham acções distintas, respectivamente A e ¬A. Para ilustrar este resultado, gostaria apenas de acrescentar um exemplo à exposição de Eells. Consideremos, para o efeito, um caso qualquer em que, apesar de se verificarem estas relações probabilísticas, e A dominar ¬A, ¬E *não dita* A:

	-E	E
A	10	<u>6</u>
¬A	<u>2</u>	4

Neste exemplo continua a existir dominação de A sobre ¬A, mas já não é o caso que exista um estado que é sempre melhor do que o outro, seja qual for a acção efectuada. Se A for efectuada, ¬E é melhor do que E, mas se ¬A for efectuada, E é melhor do que ¬E. Repare-se também que, neste exemplo, $(10 - 6) < (10 - 2)$ e $(2 - 4)$ tem um sinal diferente de $(10 - 6)$.

Como se pode constatar, o princípio evidencial, no caso de as probabilidades condicionais serem suficientemente elevadas, recomenda também a acção dominante, não se verificando o conflito entre princípios, característico dos PNCC. Em suma, para criarmos um problema deste tipo os seguintes ingredientes têm de estar presentes:

1. ¬E *dita* A nas preferências do agente (as diferenças acima têm de ser positivas, garantindo que A domina ¬A e que E é muito pior do que ¬E).
2. *pr*(E|A) > *pr*((E|¬A) e *pr*(¬E|¬A) > *pr*(¬E|A).
3. Nenhuma das acções é causa de qualquer um dos estados do mundo.

O ingrediente 3 garante que o sentido da relação de causalidade presente, dos estados do mundo para as acções, torna correcta a recomendação do princípio da dominação. Torna-se, assim, fácil constatar que num PN com dois estados e duas acções, E (ter o gene) é a causa e A é a acção sintomática (fumar). Está assim encontrada a receita do paradoxo.

10.2. Ratificacionismo e outros 'tickles'

Como foi mencionado, a teoria causal resolve de forma adequada os dois PN's de causa comum que analisámos; ou seja, a teoria recomenda aquelas acções que todos aceitamos como as mais intuitivamente plausíveis – fumar e comer os ovos. Aqui não está em causa a correcção destas recomendações específicas, mas sim descobrir se uma certa revisão da teoria evidencial consegue lidar com este tipo de problemas com o mesmo sucesso da teoria causal, recomendando aquelas acções que, segundo o argumento bicaixista, sabemos serem as correctas.

Essa revisão é também conhecida como 'tickle defense' (ver Eells: 1984b). É difícil traduzir esta expressão de uma forma relativamente elegante que preserve o sentido que está em causa; por isso, daqui em diante, usaremos o termo original.[71]

Concentremo-nos no caso do 'sonho do fumador'. Existe uma relação evidencial entre o desenvolvimento do cancro, a consequência sintomática, e a acção sintomática, fumar. Esta relação evidencial encontra-se naturalmente espelhada na relação de dependência condicional que justifica a aplicação do princípio evidencial. Assim, há que quebrar, de algum modo, esse vínculo evidencial. Uma maneira de alcançar este objectivo seria conseguir detectar a presença da causa comum. Se a causa comum estiver presente, não será através da manipulação dos seus sintomas que se conseguirá erradicá-la.

[71] 'Tickle' significa literalmente 'cócega' ou 'irritação'. A ideia fundamental consiste em referir algo que o agente sente: uma sensação propriamente dita, como uma comichão, ou um desejo que se faz sentir em maior ou menor grau.

Existirá, então, algum elemento da deliberação do agente que o possa ajudar a identificar a presença da causa comum? O modo como os dois problemas foram apresentados oferece-nos uma resposta. No caso do fumador, a acção sintomática é acompanhada pelo desejo de fumar – o *tickle*; e, no caso do colesterol, a acção sintomática é acompanhada pelo desejo de comer os ovos – o *tickle*. Em suma, se o agente monitorizar de forma atenta toda a informação que tem ao seu dispor, ele poderá aí encontrar um indício da presença da causa comum. Não quer isto dizer que tenha de existir sempre um *tickle*, tal como não é de todo necessário que um gene faça sempre sentir a sua presença através de um desejo. Segue-se, portanto, que a *tickle defense* não constitui *prima facie* uma defesa da teoria evidencial para todos os casos possíveis.[72] Contudo, nos casos em que isso acontece, a aplicação do princípio evidencial é acrescida dessa recomendação: monitorizar aquelas variáveis que, num problema de decisão, podem oferecer confirmação da presença de uma causa comum.

Existem três elementos que determinam qual é a acção maximizadora: a fórmula utilizada para o cálculo da utilidade, as probabilidades subjectivas do agente e os seus desejos. Como se supõe, de acordo com a teoria, que o agente é racional, o 'gene' (ou outro mecanismo que funcione como causa comum) não irá afectar a escolha da fórmula, mas apenas as crenças e desejos do agente. Dado que o agente conhece essas crenças e desejos, ele irá reconhecer o *tickle*. A revisão do princípio evidencial consiste, portanto, na adopção de uma regra de evidência total: condicionalizar as probabilidades dos estados não apenas das acções sintomáticas, mas também da existência de eventuais *tickles*. Se antes tínhamos a seguinte desigualdade:

pr(cancro|fumar) > pr(cancro|ñ fumar).

Agora, acrescentando os *tickles* aos condicionantes, passamos a ter a seguinte desigualdade:

pr(cancro|tenho *tickle* ∧ fumo) = pr(cancro|tenho *tickle* ∧ ñ fumo).

[72] Embora seja difícil aceitar que não existe sempre um elemento fenomenológico que corresponde à vontade interior do agente, na verdade, segundo Skyrms (1980: 131), '(…) there need not be a tickle. The mechanism responsible for the increased intake of cholesterol might operate in any number of ways, conscious or inconscious, and the agent making the decision might not have a clue as to how it operates. For him the example retains its force'.

Por outro lado, considere-se o caso em que fumar, ao contrário de acarretar efeitos nefastos para a saúde, é, em geral, benéfico, prevenindo a activação do gene. Teríamos, então,

pr(cancro|ñ tenho *tickle* ∧ fumo) < *pr*(cancro|ñ tenho *tickle* ∧ ñ fumo).

Ou seja, o *tickle* bloqueia/tapa [*screens off*] a relação evidencial entre fumar e o desenvolvimento do cancro, legitimando, como inicialmente, o raciocínio por dominação. Esta noção de bloqueio/tapamento [*screening*] encontra-se na base da *tickle defense* e depende dos pressupostos fundamentais da teoria com que estamos a trabalhar: que estamos a lidar com acções voluntárias e que essas acções possuem causas mentais (ver §2.2). Para uma definição mais formal do efeito de bloqueio (Eells 1984c: 182), considere-se que C causa uma acção A através de R, em que R especifica um conjunto particular de crenças e desejos de um agente. Assim, se o agente souber que os estados expressos por R são uma condição suficiente de A, C deixa de ter, para o agente, qualquer eficácia na produção de A, de onde podemos supor que a seguinte igualdade é verdadeira:

pr(A|R∧C) = *pr*(A|R∧¬C);

de onde, por simetria da independência probabilística, se segue:

pr(C|R∧A) = *pr*(C|R∧A).[73]

Isto significa que a acção se torna evidencialmente irrelevante em relação ao estado do mundo e à consequência sintomática que dele resulta; mais especificamente, fumar (por si só) torna-se evidencialmente irrelevante para a presença do gene e para o desenvolvimento do cancro. No exemplo do fumador, incorporando já em R a presença ou ausência do *tickle*, temos a seguinte igualdade:

[73] É argumentável que um agente não possa decidir ter as crenças e os desejos que tem. Se isto for o caso, então o facto de o efeito causal de C ser mediado pelos estados descritos em R em nada afecta o papel causal de C. Pressupondo o pano-de-fundo da Psicologia Popular – de acordo com a qual a possibilidade de explicar a acção depende da atribuição de estados mentais ao agente – será difícil articular os conceitos de livre-arbítrio e de responsabilidade moral no contexto desta tese. Como se referiu, a *tickle defense* pressupõe que estamos a lidar com acções voluntárias.

pr(gene-cancro|R∧fumo) = *pr*(gene-cancro|R∧¬fumo).

Na verdade, existe um caso em que o bloqueio 'falha', embora isso não seja prejudicial à *tickle defense*. Trata-se do caso em que fumar pode prevenir a activação do gene. Aí poderíamos dizer que a *tickle defense* 'falha' por excesso, pois a acção sintomática, embora seja na mesma evidencialmente irrelevante, exerce também uma relação causal de sentido contrário àquela outra, ilusória, que o bloqueio existe para combater:

pr(gene-cancro|R∧fumo) < *pr*(gene-cancro|R∧¬fumo).

Mas será que este último caso nos faz retirar alguma confiança à *tickle defense*? Existirão outras situações em que as características da acção sintomática permitem iludir o bloqueio? Desde que a causa seja má (cancro, lesão na artéria) e a acção sintomática seja de modo a prevenir ou a contrariar os efeitos da causa comum (a probabilidade da causa, dada a acção, é maior do que sem ela), a recomendação da teoria evidencial coincide sempre, e de forma correcta, com a acção dominante. Por outro lado, se a acção sintomática é de modo a potenciar o efeito da 'causa comum', então já não teremos um PN de causa comum, mas apenas, possivelmente, um problema de controlo da vontade por parte do agente.[74]

Como se disse, dada a possibilidade de não existir um *tickle* associado a uma acção sintomática, a *tickle defense* não pode constituir uma defesa geral da teoria evidencial. Insistir não só na presença do *tickle*, mas também, caso ele exista, no perfeito acesso fenomenológico ao mesmo, conduz-nos, talvez, a uma idealização exagerada. Como tal, não é difícil conceber a hipótese de o agente não se aperceber de certos elementos de R que são necessários ao bloqueio, podendo, no decorrer da sua deliberação, acreditar no seguinte:

pr(A|R∧C) > *pr*(A|R∧¬C).

De onde se segue que

[74] Esta seria uma situação particularmente dramática: não só fumar contribui para o desenvolvimento de cancro, como o próprio gene que causa cancro, causa também uma acção que contribui para o seu efeito. Continua a ser um problema de causa comum, embora a aplicação da teoria evidencial seja incontroversa.

$pr(R \wedge C|A) > pr(R \wedge \neg C|A)$,

concluindo irracionalmente que ¬A (não fumar, por exemplo) é a acção racional.[75]

É desejável, contudo, que a teoria se aplique a agentes menos sofisticados que nem sempre têm acesso total aos seus conteúdos mentais - agentes que não são, portanto, idealmente racionais. Aliás, nenhum agente tem um acesso completo aos seus estados mentais. Talvez seja este o motivo que levou Jeffrey (1983) a fazer um importante acrescento à sua teoria, acrescento esse designado por *ratificacionismo*. Passamos, assim, a ter o seguinte princípio: é racional empreender uma determinada acção se, e somente se, essa acção for ratificável.

No seu espírito, o ratificacionismo é uma *tickle* defense. A diferença é que, agora, o *tickle* consiste na própria decisão de escolher uma ou outra acção. Isto implica aceitar que a decisão é conceptualmente distinta da acção final e que, para qualquer decisão *d* e quaisquer duas acções alternativas, *a* e *b*,

$pr(d(a)\ \&\ b) > 0$.

Ou seja, é sempre possível que o agente não consiga concretizar a sua decisão, seja por fraqueza da vontade, seja por algum impedimento ou imprevisto. Optar por uma acção confere uma altíssima probabilidade a essa acção, mas sempre menos do que 1. Convém também acrescentar que para a teoria poder funcionar, as decisões têm necessariamente de ser estados mentais com conteúdo fenomenológico explícito e reconhecível.

A ideia fundamental, em termos de aplicação, consiste em calcular a utilidade esperada de uma acção na suposição de que se tomou essa decisão, ou, de outra maneira, calcular a utilidade esperada de cada acção, pós-escolha. Esta estratégia consiste numa *tickle defense*, na medida em que a própria decisão constitui evidência para a presença da causa comum, bloqueando, desse modo, a relação de dependência evidencial entre essa mesma

[75] Frank Jackson e Robert Pargetter (1983) apresentaram um argumento importante contra a 'tickle defense': a teoria da decisão deve ter um alcance não apenas subjectivo, mas também objectivo. Devo poder decidir o que é melhor para mim próprio, de acordo com os meus desejos e a minha função de probabilidade, mas também devo poder escolher para um terceiro, de acordo com os *seus* desejos e a *minha* função de probabilidade. Contudo, a mente de um terceiro não é transparente para mim e os seus *tickles* são para mim desconhecidos. Logo, se eu tentar decidir o que é melhor para o outro, num exemplo como o sonho do fumador, aplicando o princípio de Jeffrey obterei a recomendação errada. Contudo, aplicando a teoria causal, a ausência de causalidade entre fumar e o desenvolvimento de cancro é para mim transparente, determinando, portanto, a escolha correcta.

causa e a acção sintomática. Convém notar que esta estratégia apresenta uma vantagem em relação à abordagem genérica da *tickle* defense: ao especificar-se que é a decisão do agente que passa a contar como *tickle*, garante-se que o *tickle* está sempre presente e à total disposição do agente. Na verdade, a decisão é aquilo que se designa por 'metatickle'. Dada a possibilidadde não existir uma manifestação fenomenológica da causa comum, a influência dessa causa manifestar-se-á através das funções de utilidade e probabilidade do agente, e, em última instância, na decisão tomada. O próprio Eells (1984) aceita esta designação como sendo a mais apropriada.

Se quisermos, podemos definir a regra ratificacionista da seguinte maneira: ao agirmos, devemos ter em consideração a evidência que a nossa própria decisão nos oferece acerca do tipo de pessoa que somos – se temos, ou não, o gene, ou se temos, ou não, a lesão na artéria. Nas palavras de Jeffrey (1983: 16):

'(…) escolhe para a pessoa que esperas vir a ser quando tiveres escolhido'.

Uma maneira de tornar tudo isto mais claro é considerar a necessidade de rever as tabelas de decisão de um problema, com vista a incorporar nelas o efeito de bloqueio que a suposição de uma escolha provoca. A seguinte é a tabela de decisão 'clássica' do 'sonho do fumador, estando identificadas as respectivas probabilidades condicionais (quase 1 e quase 0):

	Tenho gene	Ñ tenho gene
Fumar	cancro + prazer de fumar (1)	saudável + prazer de fumar (0)
Ñ fumar	cancro – prazer De fumar (0)	saudável – prazer de fumar (1)

Dadas estas probabilidades, *pr*(tenho gene|fumo) ≅ 1 e *pr*(ñ tenho gene|ñ fumo) ≅ 1, a teoria evidencial oferece a recomendação de não fumar. Consideremos agora as tabelas de decisão revistas:

d(fumar):

	Tenho gene	Ñ tenho gene
Fumar	(1)	(0)
Ñ fumar	(1)	(0)

d(ñ fumar):

	Tenho gene	Ñ tenho gene
Fumar	(0)	(1)
Ñ fumar	(0)	(1)

Como se pode constatar, as decisões em causa tornam as acções evidencialmente irrelevantes para a ocorrência dos estados. Dada a decisão de fumar, a probabilidade de ter o gene passa a ser quase 1, e como fumar, neste caso, é melhor do que não fumar, a recomendação é fumar. Dada a decisão de não fumar, a probabilidade de não ter o gene passa a ser quase 1, e como fumar, neste caso, é melhor do que não fumar, a recomendação é também fumar. Conclui-se, portanto, que fumar é a acção ratificável e que, como tal, fumar é a acção recomendada pela teoria evidencial. O que agora passa a contar para a determinação das probabilidades subjectivas dos estados é aquilo que a decisão nos diz acerca da presença ou ausência dos genes e não a simples probabilidade condicional de ter, ou não, o gene, dada esta ou aquela acção. Constata-se, também, que da análise das tabelas de decisão revistas, a acção agora recomendada pela teoria coincide com a acção dominante.

Apliquemos, agora, a análise ratificacionista ao PN. Consideremos, primeiro, a decisão de escolher a caixa opaca. O agente fica convencido de que o dinheiro estará na caixa, acabe ele por fazer seja o que for, logo

$pr_{d(\text{OPACA})}$(prevê opaca|escolhe opaca) = $pr_{d(\text{OPACA})}$(prevê opaca|escolhe duas).

Ou seja, a probabilidade de uma previsão de monocaixismo, dada a escolha da caixa opaca e a decisão de escolher a caixa opaca é igual à probabilidade de uma previsão de

monocaixismo, dada a escolha das duas caixas e uma decisão de escolher a caixa opaca. Do mesmo modo, ao decidir escolher as duas caixas, o agente fica convencido de que o dinheiro não estará na caixa opaca, acabe ele por fazer seja o que for, logo

$pr_{d(\text{2 CAIXAS})}$(prevê opaca|escolhe opaca) = $pr_{d(\text{2 CAIXAS})}$(prevê opaca|escolhe duas).

Estas duas igualdades mostram-nos que os estados, ou as previsões, passam a ser evidencialmente independentes das acções. E, como sabemos que 'escolher duas caixas' domina 'escolher caixa opaca', escolher as duas caixas é a acção recomendada pelo princípio ratificacionista.

Seja PC1 'o previsor prevê caixa opaca' e PC2 'o previsor prevê duas caixas'. Calculemos, agora para a pessoa que o agente é, após ter tomado a sua decisão, a utilidade condicional esperada de cada acção:

$$\begin{aligned}\text{UCE}_{d(\text{C1})}(\text{C1}) &= pr_{d(\text{C1})}(\text{PC1}|\text{C1}) \times u(\text{PC1}\wedge\text{C1}) + pr_{d(\text{C1})}(\text{PC2}|\text{C1}) \times u(\text{PC2}\wedge\text{C1}) = \\ &= 1 \times 1.000.000 + 0 \times 0 = \\ &= 1.000.000\end{aligned}$$

$$\begin{aligned}\text{UCE}_{d(\text{C1})}(\text{C2}) &= pr_{d(\text{C1})}(\text{PC1}|\text{C2}) \times u(\text{PC1}\wedge\text{C2}) + pr_{d(\text{C1})}(\text{PC1}|\text{C1}) \times u(\text{PC2}\wedge\text{C1}) = \\ &= 1 \times 1.001.000 + 0 \times 1000 = \\ &= 1.001.000\end{aligned}$$

$$\begin{aligned}\text{UCE}_{d(\text{C2})}(\text{C1}) &= pr_{d(\text{C2})}(\text{PC1}|\text{C1}) \times u(\text{PC1}\wedge\text{C1}) + pr_{d(\text{C2})}(\text{PC2}|\text{C1}) \times u(\text{PC2}\wedge\text{C1}) = \\ &= 0 \times 1.000.000 + 1 \times 0 = \\ &= 0\end{aligned}$$

$$\begin{aligned}\text{UCE}_{d(\text{C2})}(\text{C2}) &= pr_{d(\text{C2})}(\text{PC1}|\text{C2}) \times u(\text{PC1}\wedge\text{C2}) + pr_{d(\text{C2})}(\text{PC2}|\text{C2}) \times u(\text{PC2}\wedge\text{C2}) = \\ &= 0 \times 1.001.000 + 1 \times 1000 = \\ &= 1000\end{aligned}$$

Como se constata, a utilidade condicional esperada de escolher as duas caixas é sempre maior, seja qual for a decisão, ou o *tickle*, do agente. Portanto, a acção ratificável no PN é escolher as duas caixas

Para tornar esta estratégia ainda mais intuitivamente plausível, imagine-se que a decisão do agente é equiparável a carregar num botão que irá bloquear, de uma vez para sempre, essa mesma decisão. De seguida afastamo-nos e deixamos que um terceiro indivíduo, um calculador da utilidade idealmente racional, calcule a utilidade das duas acções para a pessoa que mostrámos ser ao carregar no botão. Podemos, então, ter a certeza de que, seja qual for a decisão que tenhamos bloqueado, a sua recomendação será sempre a de escolher as duas caixas.

É também certo que esta estratégia implica expandir a nossa ontologia de estados mentais, passando esta a incluir as decisões dos agentes, enquanto coisas distintas das acções propriamente ditas. Mas não só isto está de acordo com a maneira usual de ver as coisas, como também é um preço baixo a pagar, quando comparado com a adopção de uma teoria alternativa que nos obriga a atribuir probabilidades a condicionais contrafactuais. Em suma, a regra ratificacionista parece conseguir obter os resultados desejáveis nos PN's de causa comum. Partindo do princípio de que o PN tem a mesma estrutura destes, pode-se construir um bom argumento bicaixista, do ponto de vista da teoria evidencial.

Tendo em conta o seu papel fundamental na defesa de uma solução bicaixista para um problema tão fracturante como o PN - voltando aparentemente a colocar a teoria evidencial numa posição de rivalidade com a teoria causal - algumas considerações e clarificações encontram-se ainda em ordem.

Tal como apresentei a estratégia, esta constitui uma experiência de pensamento que o agente efectua no decorrer da sua deliberação, independentemente da sua tendência para favorecer uma ou outra acção, inclusivamente quando a sua tendência é neutra. É a própria aplicação do ratificacionismo, e o resultado do argumento que a ele recorre, que deve constituir uma razão para agir e persuadir em definitivo. Mais precisamente, quando a probabilidade da nossa decisão de escolher as duas caixas é elevada, a aplicação do princípio ratificacionista 'ratifica' a nossa decisão'; quando a probabilidade da nossa decisão em escolher a caixa opaca é elevada, a aplicação do princípio reverte a nossa decisão. É este carácter de experiência de pensamento que faz com que o princípio possa persuadir indivíduos completamente indecisos, que atribuem a uma e outra decisão uma probabilidade de 0,5. Não é, portanto, correcto afirmar que o princípio ratificacionista falha por não poder aplicar-se a indivíduos indecisos. O princípio pode servir de guia para a acção, mesmo quando os agentes estão indecisos. Aliás, como já vimos, o próprio Jeffrey reconhece o carácter hipotético da estratégia, afirmando que devemos agir

maximizando o valor da evidência (que a nossa decisão oferece), não como a estimamos *agora*, mas como a *estimaríamos* caso tomássemos uma ou outra decisão. A estratégia ratificacionista tem, portanto, como base intuitiva a noção muito plausível de que as nossas decisões podem muitas vezes alterar as nossas crenças e, como tal, os nossos desejos. Este aspecto de experiência de pensamento permite também evitar que tenhamos de alargar desnecessariamente a nossa ontologia para incluir não apenas acções e decisões, mas também meta-decisões. Ou seja, as nossas decisões finais constituiriam meta-decisões que seriam tomadas apenas depois de termos tomado decisões iniciais e com base nelas. Ora, isto não está de acordo com o modo como intuitivamente encaramos o nosso processo de raciocínio em problemas de decisão, pois se uma decisão não é final, então temos tendência para não a considerar uma verdadeira decisão. Mas, se este procedimento for apenas hipotético, então não necessitamos de meta-decisões, pois basta-nos apenas decidir (uma única vez) com base no que concluiríamos acerca de nós próprios, caso tomássemos uma determinada decisão.

Mas será que a *tickle-defense* e a estratégia ratificacionista colocam realmente a teoria evidencial em pé de igualdade com a teoria causal? Considere-se primeiro a *tickle defense*. Uma formulação influente da objecção tradicional ao argumento monocaixista – *Por que razão não és rico*? (ver §4.1) - foi apresentada por Joyce: o bicaixista obteve tudo aquilo que podia obter, tendo em conta a sua situação específica. Ou seja, alguém com uma personalidade bicaixista (alguém 'racional', concedamos), está condenado a quase nunca receber mais de 1000 Euros, pois aquilo que o previsor faz não é mais do que adivinhar, com enorme precisão, a personalidade dos decisores. Como tal, seria completamente irracional, da parte do bicaixista, recusar os 1000 com base numa hipótese ínfima de o previsor se ter enganado. Note-se que este argumento só funciona se o agente souber que tipo de personalidade é a sua, embora tal não constitua um problema irresolúvel para o bicaixista, pois presume-se que este obterá essa informação mediante uma análise da sua inclinação a favor de uma das escolhas.[76]

Contudo, Ahmed (2014) fez notar o seguinte: dado que o tipo de personalidade do agente pode ser entendido como uma causa comum do estado do mundo e da acção empreendida,

[76] Parece-me implausível que, em todas as situações, o agente, no decorrer da sua deliberação, possua sempre um grau de crença superior a 0.5 acerca de qual é a sua inclinação relativamente às opções disponíveis. Esta ideia pode ser sustentada por uma outra (ver Price 2012), segundo a qual a deliberação *sufoca* [*crowds out*] a previsão: se alguém ainda não tomou uma decisão, então dificilmente se deverá levar a sério a sua confiança em como fará isto ou aquilo, pois permanece em aberto a possibilidade de mudar de ideias. Ou seja, a questão 'Farei x?" é transparente relativamente à questão 'Devo fazer x?'Isto significa que a resposta à primeira é dada através da resposta dada à segunda.

o mesmo funciona para o agente como um *tickle*, quebrando a relação evidencial entre as acções e os respectivos estados. Portanto, dadas estas circunstâncias, tanto a teoria evidencial como a teoria causal recomendarão que se escolham as duas caixas. O bicaixista não-comprometido com a adopção de qualquer teoria da decisão verá este resultado como uma confirmação da racionalidade da sua posição, assim como o evidencialista bicaixista. Contudo, a meu ver, o causalista bicaixista não terá dificuldade em reagir: ou aceita este resultado e tentará mostrar que continuam a existir outras razões que tornam a teoria causal superior à teoria evidencial, ou nega que a interpretação do PN como um PNCC mantém inalterada a natureza do PN apresentado por Nozick.

Ambas estas respostas me parecem bastante plausíveis, especialmente a primeira: se, como se viu, a *tickle defense* se deixa reduzir a uma forma específica de ratificacionismo – quando o agente não tem nenhuma inclinação em particular, ele pode sempre recorrer à estratégia ratificacionista - então existem boas razões para duvidar do sucesso do resultado apontado por Ahmed, na medida em que existem boas razões para se duvidar da viabilidade da estratégia ratificacionista.

Existe uma versão do PN em que a aplicação do princípio ratificacionista resulta numa recomendação de monocaixismo. Tal recomendação mina a possibilidade de se aplicar universalmente a estratégia ratificacionista aos PN's de causa comum. Considere-se um problema de decisão em que o previsor não só prevê com elevado grau de probabilidade a nossa acção, como prevê também qual foi a nossa decisão, mesmo quando, por algum motivo, a nossa acção final acaba por ir contra essa decisão (Joyce 1999: 158-159). Ou seja, ele não só é fiável a prever (C1 ∧ *d*(C1)), como também o é a prever, por exemplo, (C2 ∧ *d*(C1)). Considere-se a seguinte matriz com os *payoffs* desta nova versão do problema:

	P(C2 ∧ *d*(C2))	P(C1 ∧ *d*(C2))	P(C2 ∧ *d*(C1))	P(C1 ∧ *d*(C1))
C2 ∧ *d*(C2)	1000	1,001,000	1000	1,001,000
C2 ∧ *d*(C1)	"	"	"	"
C1 ∧ *d*(C2)	0	1,000,000	0	1,000,000
C1 ∧ *d*(C1)	"	"	"	"

De modo a aplicar-se o princípio de Jeffrey, temos de saber quais são as probabilidades condicionais em causa, ou seja, a probabilidade dos estados do mundo nas colunas acima, dada a conjunção de uma acção com uma decisão. Mantendo a elevada fiabilidade do previsor, suponha-se que essas probabilidades são as seguintes (os asteriscos no lugar do condicionado devem ser substituídos por cada uma das 4 previsões possíveis):

	P(C2 ∧ *d*(C2))	P(C1 ∧ *d*(C2))	P(C2 ∧ *d*(C1))	P(C1 ∧ *d*(C1))
pr(*|(C2 ∧ *d*(C2))	0,9	0,01	0,08	0,01
pr(*|C2 ∧ *d*(C1))	0,08	0,01	0,9	0,01
pr(*|C1 ∧ *d*(C2))	0,01	0,9	0,01	0,08
pr(*|C1 ∧ *d*(C1)	0,01	0,08	0,01	0,9

Esta distribuição de probabilidades revela que o agente é extremamente fiável a prever qualquer que seja a conjunção entre a acção final e a decisão inicial; raramente prevê correctamente a acção e incorrectamente a decisão; e muito raramente prevê incorrectamente tanto a acçao final, como a decisão inicial.

Se considerarmos, por exemplo, que a UCE (C2 ∧ *d*(C2)) = UCE $_{d(C2)}$ C2, então podemos aplicar a estratégia ratificacionista e calcular a UCE das duas acções disponíveis, dada uma ou outra decisão, utilizando para o efeito as probabilidades da tabela acima. O resultado obtido é o seguinte:

UCE (C1 ∧ *d*(C2)) > UCE (C2 ∧ *d*(C2)),
UCE (C1 ∧ *d*(C1)) > UCE (C2 ∧ *d*(C1)).

Conclui-se, assim, que nesta versão do PN, a acção ratificável é a acção monocaixista. Mas por que motivo não funciona aqui a estratégia ratificacionista? A experiência de pensamento consistia, nas palavras de Jeffrey, em supor que poderíamos fazer uma escolha para um decisor que já tivesse tomado a sua decisão final. Ou seja, consideraríamos a evidência que a sua decisão nos oferecia e, tendo em conta essa evidência, ou *ratificávamos* essa decisão, ou reverteríamos essa decisão: 'Não, espera, se decidiste fazer *x*, então é melhor fazeres *y*'. O problema reside aqui no facto de o previsor prever também essa hipotética reversão da decisão inicial e, portanto, reverter a decisão

não nos permite *enganar* o previsor; este consegue topar o esquema do decisor e do seu, hipotético, *amigo ratificador*.

Defendi atrás a plausibilidade da correcção da escolha monocaixista e estou de acordo com a conclusão obtida pelo emprego da estratégia ratificacionista neste último caso, embora considere que esta foi obtida pelos motivos errados – como veremos a seguir, a fiabilidade desta estratégia é duvidosa. A plausibilidade dessa correcção encontrava-se associada a uma certa aplicação da teoria causal, e a uma dada interpretação de contrafactuais, e não a quaisquer considerações relacionadas com a ratificabilidade das decisões. Não existirão, portanto, outros problemas de decisão suficientemente relevantes em que a aplicação do ratificacionismo apresenta falhas irreparáveis? É isso que averiguaremos de seguida.

11. Contra-exemplos à teoria causal da decisão

11.1. Psicopatas e assassinos

Richard Holton (2016) criou o conceito de auto-sinalização [*self-signalling*], com o intuito de designar a tentativa de ganharmos conhecimento acerca de nós mesmos, e de aspectos do nosso eu mais profundo, através da análise do nosso próprio comportamento. Este empreendimento encontra-se, contudo, sujeito a um princípio de incerteza: dado que podemos manipular o nosso comportamento, a evidência que dele retiramos pode ser enganadora e chegar-nos distorcida.

Existem dois casos de tentativa de auto-conhecimento: quando o objecto a que tentamos aceder é o próprio comportamento, e quando o objecto é algo distinto do próprio comportamento. Um exemplo do primeiro caso é quando agimos moralmente de modo a convencer-nos a nós próprios de que somos pessoas morais; e, dado que agimos moralmente, é plausível concluir daí que realmente o somos. Neste caso, a evidência poderá ser boa. No segundo caso, a qualidade da evidência é bastante mais duvidosa. Exemplos deste são aqueles casos em que o comportamento não tem qualquer eficácia causal sobre o aspecto de nós mesmos do qual desejamos e gostaríamos de obter evidência. Um exemplo é o da doutrina calvinista: a nossa salvação encontra-se pré-determinada pela divindade e, embora agir moralmente possa constituir evidência de que somos um

dos escolhidos, nada nos garante que essa evidência não é enganadora. Ou seja, apesar de existir uma correlação estatística entre agir moralmente e ser-se escolhido, agir moralmente não é causa da nossa salvação.

O caso do PN é exactamente análogo ao do calvinismo: tenta-se produzir evidência de um estado auspicioso, sem que se contribua, com a nossa acção, para causar esse mesmo estado. Daí a expressão de David Lewis (1981), segundo a qual a teoria evidencial apresenta 'uma política irracional de lidar com as notícias', ou seja, com a evidência que as nossas acções nos oferecem. Deste modo, a distinção entre a teoria evidencial e a teoria causal pode ser apresentada nos seguintes termos: a primeira diz-nos para fazermos aquilo que nos satisfaria saber que teríamos feito, e a segunda diz-nos para fazermos aquilo que trará, com maior probabilidade, resultados auspiciosos.

Vimos como a teoria causal de Gibbard e Harper consegue obter no PN os resultados advogados pelos bicaixistas e também de que modo consegue obter os resultados correctos nos outros casos de causa comum, como o 'gene do fumador' e 'Salomão e Bathsheba'. Mas a teoria de Gibbard e Harper não é a única teoria causal que existe. Brian Skyrms (1980) e David Lewis (1981) apresentaram as suas próprias versões da teoria. As três têm em comum a exigência de incorporação de distinções e nexos causais no cálculo da utilidade esperada. Apesar das diferenças entre elas, Lewis argumentou também que as três constituem basicamente a mesma teoria. Por este motivo, será útil considerarmos de maneira formal uma generalização das várias teorias causais existentes.

A ideia central é a de que o valor causal, ou a eficácia, das acções consiste em utilidades esperadas do tipo incondicional (à maneira de Savage), calculadas relativamente a uma partição *K*, cujos elementos são proposições maximamente específicas acerca de como aquilo que interessa ao agente depende causalmente daquilo que o agente faz. Lewis (1981) chama-lhes *dependency hypotheses*. Para uma dada acção A,

$$\mathrm{UCausalE}\ (\mathrm{A}) = \sum\nolimits_{K} pr(K) \times u(\mathrm{A} \wedge K).$$

As proposições em *K* mantêm fixa a nossa perspectiva acerca da estrutura causal do mundo, independentemente da acção que venha a ser realizada, constituindo um pano-de-fundo (*background conditions*, segundo Skyrms) que o agente considera relevante para o resultado das suas acções. Assim, o quão auspiciosa uma acção é pode agora identificar-se com a sua eficácia causal, pois se o agente conhecer, sem margem para dúvidas, a

estrutura causal do mundo, ele atribuirá a um específico $k \in K$ uma probabilidade subjectiva de valor 1, de onde se segue que ele escolherá uma acção com base no 'valor das notícias' que essa acção traz; ou seja, quando $pr(K) = 1$, a $u(\mathrm{A}\ \mathrm{K}) = u(\mathrm{A})$. A diferença relativamente à teoria evidencial é que a utilidade das acções passa a ser sensível apenas à probabilidade subjectiva *incondicional* que é atribuída às *dependency hypothesis*, e não à probabilidade que era atribuída às mesmas, sob a condição de uma ou de outra acção ser realizada.

A questão consiste agora em saber como devemos interpretar *K*. Uma maneira natural é considerar *K* uma partição de mundos possíveis *M* aos quais o agente atribui uma probabilidade diferente de 0; ou, de uma maneira ainda mais específica, poder-se-á, à maneira de Gibbard e Harper, considerar cada $k \in K$ como uma conjunção de proposições contrafactuais do tipo $A \gg M$, em que as nossas crenças acerca de relações de causalidade são interpretadas em termos de dependência contrafactual. Temos, assim, de acordo com a sugestão de Stalnaker, a teoria causal de Gibbard e Harper:

$$\mathrm{UContrafactualE}\ (\mathrm{A}) = \sum_{M} pr(\mathrm{A}\ \gg M) \times u(M).$$

A utilidade causal esperada de uma acção passa a ser uma média ponderada [*weighted average*] da utilidade dos mundos possíveis em que A é verdadeira, quando esses mundos são, em parte, uma consequência causal de A; ou seja, o agente acredita que A promove M se, e somente se, $p(\mathrm{A} \gg \mathrm{M}) > p(\neg \mathrm{A} \gg \mathrm{M})$.

Mas Andy Egan (2007) apresentou dois exemplos com o intuito de mostrar que a teoria causal, tal como a teoria evidencial, pode também recomendar acções que, intuitivamente, são irracionais. O primeiro dos contra-exemplos é o seguinte:

O Botão do Psicopata

António está a reflectir sobre se há de, ou não, carregar no botão que mata todos os psicopatas. Seria muito melhor, pensa ele, viver num mundo sem psicopatas. Infelizmente, António tem quase a certeza de que apenas um psicopata carregaria em tal botão. António prefere muito mais continuar vivo a morrer deixando para trás um mundo sem psicopatas. Deve António carregar no botão?

Segundo Egan, a teoria causal da decisão, em certas circunstâncias, recomenda que se carregue no botão, contrariando a intuição de que é irracional fazê-lo. A base desta intuição consiste na ideia de que carregar no botão é a acção que, segundo as nossas expectativas mais fortes, trará piores resultados. Esta parece ser a razão oferecida por Egan para justificar a intuição de irracionalidade:

'(…) o facto de a teoria causal da decisão nos forçar a usar, com via a determinar o valor das acções, apenas as crenças incondicionais do agente nas *dependency hypotheses*, torna o seu veredicto cego relativamente aos aspectos das crenças do agente aos quais deveria ser sensível – nomeadamente, a confiança do agente em como um curso particular de acção, caso seja seguido, está condenado a falhar e a trazer consigo um resultado pior do que a sua alternativa' (Egan 2007: 101).

Mais à frente iremos analisar esta interpretação e avaliar se é, ou não, consistente com os dados do problema. Mas por agora é necessário considerar detalhadamente esses dados para compreender por que razão a teoria causal pode recomendar premir o botão.

Em primeiro lugar, o agente tem de ter uma crença forte de que não é psicopata, daí preferir bastante mais viver num mundo sem psicopatas do que no actual estado de coisas; ou seja, a probabilidade que atribui à contrafactual 'Se eu carregasse no botão, então não morreria' tem de ser elevada. Em segundo lugar, do mesmo modo que, no exemplo do fumador, fumar não provoca cancro do pulmão, premir o botão não faz com que o agente se torne psicopata. Apesar de a probabilidade condicional de ter cancro, dado fumar, ser alta, e de a probabilidade condicional de se ser psicopata, dado premir o botão, ser também elevada, tal não deve ser motivo de preocupação para o defensor da teoria causal, pois esta probabilidade não será utilizada no cálculo da utilidade esperada. A probabilidade que será utilizada é a probabilidade incondicional atribuída à contrafactual acima mencionada: 'Se eu premisse o botão, então não morreria'. Portanto, se António considerar extremamente improvável a possibilidade de ele próprio ser psicopata, e considerar suficientemente satisfatório viver num mundo sem psicopatas, a teoria causal recomendará premir o botão.

Os defensores da teoria causal podem, neste ponto, responder à acusação invocando a aplicação de um princípio ratificacionista e declarando que a acção de premir o botão não é ratificável. Se o agente desejar seguir a recomendação que resulta da aplicação da teoria

causal, então ele decidirá premir o botão. Contudo, ao tomar essa decisão, a sua crença na possibilidade de ele próprio ser psicopata aumentará. Quanto mais convencido ele estiver de que premir o botão é a coisa racional a fazer, mais convencido estará de que é um psicopata. Como tal, a probabilidade da crença em como é um psicopata chegará a um ponto em que será demasiado elevada para que a teoria causal continue a recomendar premir o botão.

Do mesmo modo, as coisas começam subitamente a sorrir para o lado dos evidencialistas, pois estes terão a certeza de que a sua teoria nunca recomendará uma acção não-ratificável. Uma forma de captar o espírito do ratificacionismo, que me parece correcta, é considerá-lo como uma garantia de que o agente não acaba por tomar uma decisão da qual mais tarde se venha a arrepender. Esse arrependimento pode surgir quando o agente ignora a informação acerca de si próprio que as suas decisões lhe oferecem. Esta é uma precaução básica que qualquer teoria da decisão deve incorporar, seja-se evidencialista ou causalista. Mas consideremos o que acontece na suposição da decisão contrária. Se é irracional premir o botão, a teoria causal recomendará que não se prima o botão. Ao decidirmos não premir o botão, a nossa crença em como não somos psicopatas torna-se mais forte. Ao tornar-se mais forte, mais confiante nos tornamos de que premir o botão é a acção que promoverá um estado melhor. Ou seja, chegará, novamente, um ponto em que a teoria causal recomendará novamente que se prima o botão. Em suma, a acção que consiste em não premir o botão também não é ratificável.

Este é um ponto da situação altamente indesejável. Se ambas as acções disponíveis fossem ratificáveis, poderia ser razoável classificar ambas como racionais, na medida em que não seria possível, de acordo com o espírito do ratificacionismo, arrepender-nos de executar qualquer uma delas. Contudo, se a teoria em causa classifica todas as acções disponíveis como não-ratificáveis, efectuar qualquer uma delas pode conduzir-nos ao arrependimento. A aplicação do ratificacionismo impedirá a teoria causal de incumprir o requisito de ser consistente, pois impedi-la-á de recomendar uma acção que não pode ser classificada como racional. Por outro lado, tal aplicação faz com que a teoria não cumpra o requisito da completude, ao impedi-la de recomendar qualquer uma das duas únicas acções disponíveis.

Independentemente do modo de interpretar a situação em termos de incumprimento de princípios de racionalidade, o que importa é qualificar de forma correcta a situação a que se chegou: uma ausência de estabilidade no processo deliberativo, tendo como resultado,

aparentemente inescapável, a impossibilidade de classificar qualquer uma das acções como racional.

Voltando um pouco atrás, avaliemos a intuição de irracionalidade que, segundo Egan, está associada à recomendação da teoria causal, nomeadamente, que premir o botão é a alternativa que trará consigo o pior resultado. James Cantwell (2010) argumenta que a intuição de Egan resulta de um equívoco. Segundo Cantwell, todos concordamos, a partir da descrição inicial do problema, que a seguinte probabilidade condicional é elevada:

(1) *pr*(António é psicopata|António prime o botão) = 0.9.

Esta é a probabilidade que a teoria evidencial irá utilizar para o cálculo da utilidade esperada, o que conduzirá a uma recomendação de não premir o botão. Somente a consideração do valor desta probabilidade, e de nenhuma outra, é que justificará a intuição de que é irracional premir o botão.

Desta probabilidade, ou da respectiva crença, segue-se:

(2) *pr*(António morre|António prime o botão) = 0.9.

Até aqui tudo bem. Contudo, para fazermos sentido da ideia de Egan, de que premir o botão conduz ao pior resultado, a seguinte probabilidade, em que assenta a teoria causal, teria também de ser alta:

(3) *pr*(se António premisse o botão, então António morreria) = 0.9.

Sabemos, no entanto, que esta probabilidade tem de ser baixa (0.1), pois um dos dados do problema é que António atribui uma alta probabilidade à crença de que ele próprio não é um psicopata. Se (3) fosse realmente verdadeira, então a recomendação da teoria causal seria não premir o botão, o que contradiz o argumento de Egan.

O diagnóstico que Cantwell faz do erro de Egan consiste em atribuir a este a crença implícita na seguinte tese:

$pr(B|A) = pr(A \rightarrow B)$.[77]

Ou seja, de (2) resultaria a seguinte condicional indicativa:

(4) *pr*(Se António premir o botão, então António morrerá) = 0.9.

As probabilidades em (2) e (4) estariam, portanto, na base da análise informal de Egan, nomeadamente, na justificação da intuição de irracionalidade associada à recomendação da teoria causal. Contudo, numa análise formal, que esteja de acordo com os princípios da teoria causal, é (3) que é usada no cálculo da utilidade causal esperada. Portanto, se Egan pretende justificar, através de (2) e (4), a sua intuição de que a recomendação da teoria causal está errada, então essa justificação não tem qualquer fundamento. Mas, por outro lado, se (3) for utilizada para o cálculo da utilidade causal esperada, então, para sermos consistentes com os dados do problema, teremos de considerá-la falsa, assim se justificando a recomendação da teoria causal, premir o botão.[78]

A meu ver não é importante especular acerca dos possíveis fundamentos da justificação de Egan para a sua intuição de irracionalidade. As intuições valem o que valem e o que se exigiria neste caso, para defender a teoria causal, seria um argumento persuasivo que mostrasse que é racional não premir o botão. Além disso, não sei até que ponto, neste problema, as intuições de irracionalidade não serão tão frequentes quanto as intuições de indecisão irresolúvel, por assim dizer. Estas últimas estariam justificadas pela análise que resulta da aplicação do princípio ratificacionista, e que se caracteriza, precisamente, pela constatação de um estado de oscilação ou de instabilidade deliberativa.

A intuição de irracionalidade talvez corresponda a algo que, do ponto de vista teórico, faz sentido. Quando decidimos atribuir uma utilidade à consequência 'premir o botão e morrer' deparamo-nos com uma dificuldade. Do ponto de vista da teoria, como será possível atribuir esse valor? Será possível fazê-lo através da nossa disposição para aceitar o *ratio* de uma aposta entre morrer e adquirir um determinado bem? Isto tem de ser

[77] Como vimos, esta tese não é necessariamente verdadeira (ver Apêndice 2). Cantwell usa a condicional material, e não a contrafactual, devido à sua semântica para as condicionais. Mas poderíamos aqui substituir uma pela outra, sem que a conclusão do seu argumento se altere.

[78] Segundo Cantwell, a teoria causal recomenda, de facto, que se prima o botão. Contudo, se quisermos ser fiéis aos dados do problema – que o agente não é psicopata – devemos ir contra a ideia de Egan, segundo a qual este é um curso de acção que conduz a uma má consequência. O problema que Cantwell realmente encontra na teoria causal consiste em esta recomendar, como vimos, uma acção não-ratificável, e não em esta recomendar que se prima o botão.

possível, caso contrário não existiriam pessoas dispostas a fumar ou a jogar roleta russa. Mas a probabilidade, por menor que esta possa ser, da ocorrência dessa terrível possibilidade parece tornar irrelevante o bem que resultará de um mundo sem psicopatas. Não pretendo com isto afirmar que, do ponto de vista do agente, a morte é simplesmente o fim de toda e qualquer utilidade, pois um agente altamente altruísta pode atribuir uma utilidade subjectiva a um determinado estado futuro do qual ele esteja ausente. A utilidade desse estado corresponde àquilo que, no momento de decisão, tem valor para o agente, e não ao valor daquilo que o agente viria a usufruir se estivesse vivo. Um agente determinado a fazer o bem a todo o custo, seja qual for esse bem, pode sempre considerar que o valor da sua vida é menor que o valor desse bem ou, o que na prática resulta no mesmo, menor que o valor da realização da própria acção. Portanto, alguém altruísta o suficiente, e que tenha uma crença bastante forte em como não é psicopata, pode, aplicando a teoria causal, obter uma recomendação de premir o botão. Mas não é preciso ir tão longe. Um agente apenas auto-interessado pode receber uma recomendação de premir o botão, desde que a utilidade subjectiva de um mundo sem psicopatas seja suficientemente elevada e a probabilidade de morrer seja suficientemente baixa. De qualquer modo, isto não se verifica com a maioria das pessoas, e a intuição de irracionalidade associada à acção de premir o botão, que resultaria de uma aversão natural a qualquer estado que inclua a possibilidade de morrer, torna-se bastante mais compreensível. Essa aversão impediria, à partida, a mínima ponderação que seja da possibilidade de provocar a nossa própria morte. Contudo, este é claramente um argumento fraco contra a recomendação de premir o botão.

De seguida iremos considerar um outro exemplo em que a análise resultante da aplicação da teoria causal, complementada com o princípio ratificacionista, também desemboca numa situação em que ambas as acções disponíveis não são ratificáveis. Este exemplo é estruturalmente semelhante ao ‘botão do psicopata’:

A Lesão Assassina

Maria está a reflectir sobre se há-de disparar, ou não, contra o seu rival António. Se disparar e acertar, tudo correrá pelo melhor. Mas, se disparar e falhar, Maria ficará numa situação desesperada. (António consegue sempre descobrir tentativas de assassinato mal-sucedidas, coisa que ele leva sempre a mal). Se Maria não disparar, tudo continuará na mesma, num estado de coisas que nem é bom, nem mau. Apesar de Maria ter quase a

certeza de que não irá realmente disparar, ela tem treinado regularmente tiro-ao-alvo, apenas para manter as suas opções em aberto. A sua pistola é completamente fiável e é mantida em condição pristina. Face a isto, ela acredita que é bastante provável, caso venha a disparar, que o seu tiro será certeiro. Até aqui não há problema. Contudo, Maria sabe também que existe uma lesão cerebral igualmente responsável por tentativas de assassinato e má pontaria no momento de disparar. Se ela tiver esta lesão, todo o seu treino de nada servirá – no momento crítico a sua mão irá de certeza tremer quando puxar o gatilho. Felizmente para a maioria de nós, e infelizmente para Maria, a maioria dos assassinos têm esta lesão e, como tal, a maioria falha os seus disparos. Deve Maria disparar?

Este é outro exemplo em que, na suposição de certos *payoffs* e de certas probabilidades, disparar é a acção com maior utilidade causal esperada. Primeiro, tal como no exemplo anterior, há que tornar a descrição consistente com este resultado. Neste caso, há que conciliar a confiança de Maria em como acertará, caso dispare, e a sua crença em como a maioria dos que disparam têm a lesão. Tal como no exemplo anterior, em que António começa com uma forte crença em como não é psicopata, aqui Maria começa com uma forte crença em como não tem a lesão. Para que ela mantenha esta crença, basta que a sua crença em como não irá disparar seja também forte.

À primeira vista, parece um tanto ou quanto arbitrário atribuir a Maria a crença em como não tem a lesão. Afinal, parece não existir qualquer informação adicional, segundo a descrição do exemplo, que permita atribuir a Maria a crença em como não tem a lesão. O caso parece ser similar ao do fumador, o qual não sabe se tem, ou não, o gene, e que parece não ter qualquer informação adicional que lho permita saber. Mas não será razoável que, seja quem for, acredite acerca de si próprio que não é um psicopata ou que não é um assassino? Afinal, estas crenças estão relacionadas com pulsões internas, tal como ter a posse do gene estava relacionada com a existência do *craving* do fumador.

Seja D a proposição segundo a qual Maria dispara e A a proposição segundo a qual Maria atinge o alvo. Perceberemos melhor a compatibilidade entre as crenças acima se as traduzirmos em atribuições subjectivas de probabilidade: a $pr(D \gg A)$ pode ser alta, por exemplo, maior que 0,5, e a $pr(A|D)$ pode ser baixa, por exemplo, menor que 0,5 (vimos na análise do exemplo anterior como estas probabilidades são plausíveis). Considerem-se as seguintes utilidades:

$u(D \wedge A) = 10$; $u(D \wedge \neg A) = -10$; $u(\neg D) = 0$.[79]

De modo a calcularmos a utilidade causal de disparar, teremos de utilizar a probabilidade incondicional da hipótese 'Se Maria disparasse, então Maria acertaria', pois é esta que dá conta do nexo causal entre disparar e acertar. Teremos, portanto,

$$\text{UcausalE}(D) = pr(D \gg A) \,.\, u((D \gg A) \wedge A) + pr(D \gg \neg A) \,.\, u((D \gg \neg A) \wedge \neg A).$$

Como a utilidade de um mundo possível em que o agente dispara e acerta não é mais do que a utilidade de disparar e acertar, e a utilidade de um mundo possível em que o agente dispara e não acerta é a utilidade de disparar e não acertar, segue-se que

$$\text{UcausalE}(D) = pr(D \gg A) \,.\, u(D \wedge A) + pr(D \gg \neg A) \,.\, u(D \wedge \neg A).$$

Dada a crença de Maria em como a probabilidade de disparar e acertar é alta ($pr(D \gg A) > pr(D \gg \neg A)$), a utilidade causal esperada de disparar terá de ser maior que 0. Como Maria atribui à $u(\neg D)$ valor 0, segue-se que a utilidade causal esperada de disparar é maior que a utilidade causal esperada de não disparar. Em suma, a teoria causal aconselha o agente a executar uma acção que, intuitivamente (segundo Egan), nos parece ser a pior.
Na verdade, a moral que deve ser retirada destes dois exemplos não é a necessidade de remover a contradição entre as recomendações da teoria causal e as nossas mais fortes intuições acerca do que fazer nestes casos. O que importa é que a aplicação do princípio causal conduz a uma situação de instabilidade deliberativa, em que ambas as acções possíveis não são ratificáveis. Ao pretender seguir a recomendação de disparar, o agente modifica, ou actualiza, a sua crença inicial em como não tem a lesão. Ou seja, a forte crença inicial em como não tem a lesão diminuirá à medida que o agente considera cada vez mais provável a hipótese de disparar. Mas, à medida que aumenta a crença em como tem a lesão, menos apelativa se torna a recomendação da teoria causal. Em suma, a motivação que resulta da análise dos dois exemplos, em termos de consistência teórica da doutrina, é a de tentar resolver este impasse.

[79] Seguindo a terminologia de Lewis, a partição relevante de *dependency hypotheses* (o conjunto de mundos possíveis em que D é verdadeira) é a seguinte: $\{(D \gg A); (D \gg \neg A)\}$. Todos estes valores são os mesmos de Egan (2007).

James Joyce (2012) argumentará no sentido de mostrar que a situação deliberativa alcançada não é instável, mas que constitui antes uma espécie de equilíbrio saudável, podendo o agente executar qualquer uma das acções disponíveis, sem que isso acarrete qualquer arrependimento. Para esse efeito, Joyce irá rever o princípio causal, mostrando de que modo a consideração do resultado do cálculo inicial pode contribuir para modificar as crenças do agente, e como essa modificação deve ser teoricamente enquadrada no processo de actualização de crenças conhecido por 'condicionalização' (ver 1.2). A dinâmica deste processo, quando aplicado a este caso particular, pode ser encarada como uma tradução teórica daquele raciocínio de oscilação que é característico dos exemplos mencionados.

Comecemos por atribuir valores de probabilidade aos quatro estados do mundo que constituem a partição *K* no exemplo da Lesão Assassina - consistentes com a descrição do problema - e valores de utilidade às consequências. Seja D a acção de disparar e L a posse da lesão:

$pr(D \wedge L) = 0.16$ / $u(D \wedge L) = -30$; $pr(D \wedge \neg L) = 0.08$ / $u(D \wedge \neg L) = 10$
$pr(\neg D \wedge L) = 0.04$ / $u(\neg D \wedge L) = 0$; $pr(\neg D \wedge \neg L) = 0.72$ / $u(\neg D \wedge \neg L) = 0$.[80]

Estes valores reflectem a baixa probabilidade que Maria atribui à possibilidade de ela própria vir a disparar e revelam, desde logo, a correlação que existe entre a acção de disparar e a posse da lesão, pois $pr(L|D) = 0.8$. Quanto às utilidades, não disparar mantém o estado-de-coisas actual e, portanto, não acrescenta a esse estado qualquer valor adicional. A utilidade de (¬D) será, portanto, 0. Calcule-se a utilidade causal esperada de disparar e de não-disparar, utilizando a partição [L, ¬L]:

$$\text{UcausalE}(D) = pr(L) \times u(D \wedge L) + pr(\neg L) \,.\, u(D \wedge \neg L) =$$
$$= 0.2 \times -30 + 0.8 \times 10 = 2$$

$$\text{UcausalE}(\neg D) = pr(L) \times u(\neg D \wedge L) + pr(\neg L) \times u(\neg D \wedge \neg L)$$
$$= 0$$

[80] Como o que nos interessa aqui é salientar a probabilidade de Maria ter a lesão, utilizaremos (D & L) em vez de (D » ¬A), a interpretação contrafactual de cada *k* ε *K*. Podemos também, claro, ter uma partição [L, ¬L].

À partida, a teoria causal recomenda que se dispare, e o agente racional, de acordo com este resultado, verá a sua crença em como disparará reforçada. Isto acontece porque o agente acredita que a teoria da decisão constitui o modelo normativo de racionalidade instrumental, e porque, sendo ele livre, encarará os resultados da aplicação da teoria como razões para agir desta ou daquela maneira.

Contudo, à medida que a crença em como disparará se torna mais forte, mais forte também será a crença em como tem a lesão, pois as duas coisas encontram-se evidencialmente e causalmente relacionadas (a maioria dos assassinos tem a lesão e esta é a causa das tentativas de assassinato). Ou seja, a análise que o agente faz do resultado do cálculo contribui para a alteração da probabilidade de um estado, ter a lesão, que está intimamente relacionado com a visão causal do mundo e, portanto, com o modo através do qual as acções contribuem para modificar aspectos dessa mesma visão. De outra maneira, $pr(\mathrm{L}|u(\mathrm{D}) = 2) \neq pr(\mathrm{L})$; a probabilidade de ter a lesão, dado que a utilidade de disparar tem valor dois, é maior do que a probabilidade incondicional inicial que Maria atribui à crença em como tem a lesão. Isto obrigará Maria a actualizar as suas crenças.

De acordo com Joyce (2012), deve-se indexar o cálculo da utilidade a um momento temporal *t*, a fim de distinguir o cálculo inicial t_0, dos cálculos t_n em que se utilizam valores resultantes da actualização de crenças. Temos, assim, a pr_0, que caracteriza as crenças iniciais do agente em t_0, e a u_0 que resulta do cálculo inicial. Neste caso u_0 (D) = 2. Deixa, portanto, de ser racionalmente obrigatório agir de acordo com as crenças iniciais em t_0 ou, mais precisamente, em qualquer momento t_n no qual ainda não esteja recolhida toda a informação disponível. O seguinte princípio deixa de estar correcto:

'Se a $prob_t$ caracteriza as crenças de um agente em *t*, então ele encontra-se racionalmente obrigado, em *t*, a realizar uma acção *A* que maximiza a sua utilidade causal esperada no momento *t* (…)' (Joyce 2012: 126).

E a teoria causal passa incorporar a seguinte importante restrição ao cálculo da utilidade esperada:

'Devemos agir com base nas avaliações de utilidade no momento *t*, apenas se essas avaliações se encontrarem fundadas em crenças que incorporam toda a evidência que se

encontra livremente disponível em *t*, e que sejam relevantes para a questão de saber quais são os resultados que as nossas acções provavelmente causarão' (Joyce 2012: 127).

Em suma, quando o agente toma uma decisão, pode estar a oferecer a si mesmo evidência que vai contra a razão pela qual essa decisão é tomada, tal como nos dois exemplos apresentados.

Um crítico deste novo princípio de Joyce poderia, neste ponto, sentir-se tentado a argumentar da seguinte maneira: os defensores da teoria causal focam a sua crítica no facto de a teoria de Jeffrey ser sensível a relações evidenciais que não denotam causação e, no entanto, estão agora dispostos a aceitar este novo tipo de evidência, oferecido por um elemento teórico, a utilidade das acções, que aparentemente nada parece ter a ver com o papel causal das acções.

A meu ver, isto não constitui um problema para a teoria de Joyce, pois negar a relevância do valor obtido do cálculo é o mesmo que negar a relevância de quaisquer considerações relacionadas com a ratificabilidade das acções. O agente que aceita a força normativa da teoria, encontrará nos resultados que ela oferece uma razão para agir de uma ou de outra maneira. Quanto mais razões o agente encontrar para agir de uma certa maneira, maior é a informação que possui acerca do tipo de pessoa que é, nomeadamente, o tipo de pessoa que tenderá a carregar no botão ou a premir o gatilho; ou seja, a evidência em causa faz aumentar probabilidade da crença em como tem a lesão; por seu lado, este é um aspecto de si mesmo com relevância na estrutura causal do mundo, pois, como sabemos desde o início, ter a lesão é não só responsável pelo sangue-frio necessário para cometer homicídio, como também pela tendência para falhar o disparo na hora certa. Se quisermos ser precisos, podemos notar que a distinção entre *pr*(estado|acção) e *pr*(estado) já não identifica, respectivamente, a teoria evidencial e a teoria causal. Ambas passam a estar dependentes da evidência oferecida pelas acções/decisões, mas diferem quanto ao tipo de evidência oferecida: o quão auspiciosa é a acção ou o seu poder causal.

Para completar a análise formal do contra-argumento de Egan, é importante considerar, de um modo quantitativamente preciso, o estado de oscilação deliberativa que o caracteriza. Para que a teoria causal recomende disparar no exemplo da Lesão Assassina, várias condições terão de ser preenchidas (Joyce 2012: 130-131):

1. pr_t (L|u_t (D) = x) = pr_t(L). Como vimos, o princípio da informação completa, se assim lhe quisermos chamar, exige que a deliberação chegue a um ponto em que a utilidade esperada de (D) deixe de ser evidencialmente e causalmente relevante para a probabilidade de se ter a lesão.
2. A u(D) > u(¬ D).
3. Quando x ≠ 0, pr_t (L|u_t (D) = x) ≠ pr_0 (L). Ou seja, quando a utilidade de disparar ou não disparar aumenta ou diminui, sem que a probabilidade da lesão aumente ou diminua, tal não coloca qualquer problema à teoria causal, pois através de sucessivas iterações do ciclo de revisão de crenças, esta recomendará, respectivamente, disparar e não disparar; portanto, para que o exemplo coloque problemas à teoria causal, e desse modo seja fiel ao contra-argumento de Egan, a probabilidade da lesão tem de aumentar ou diminuir dado o cálculo em *t* da u(D).

Ora, não é possível satisfazer estas três condições. De modo a respeitar-se a condição 3, a condição 1 só pode ser satisfeita quando x = 0, o que exige pr_t (L) = 0.25.[81] Mas, para que 2 seja verdadeira, é necessário que a pr_t (L) < 0.25; como vimos, a pr_0 (L) = 0.20, quando a u_0 ((D) = 2) > u_0 ((¬ D) = 0).

Qual é a conclusão que podemos retirar destes valores? Que, após a consideração de toda a evidência disponível, a teoria causal não recomenda disparar no caso da Lesão Assasina (apesar de o fazer em t_0).

Será, então, que a teoria causal recomenda não disparar, ao contrário do que pensa Egan? Para o fazer teria de ser satisfeita uma condição 2': u_t (¬D) > u_t (D). Esta desigualdade seria verdadeira se, e somente se, pr_t (L) > 0.25; mas, como sabemos agora, a satisfação de 1 exige pr_t (L) = 0.25; conclui-se que a teoria causal também não recomenda não disparar. Mais, as condições 1 e 3 são satisfeitas, garantido que toda a informação relevante é tida em consideração, se, e somente se, u(D) = u(¬D).

Estes resultados estão em ordem com as conclusões retiradas da análise informal do problema, nomeadamente, que as duas acções disponíveis parecem ser ambas não-ratificáveis. O princípio ratificacionista diz-nos que uma acção pode ser racionalmente efectuada se, e somente se, essa acção for ratificável. Adaptando o princípio à teoria causal, sabemos que uma acção é causalmente ratificável se, e somente se, dada a decisão

[81] UcausalE (D) = pr(L) × u(D ∧ L) + pr(¬L) × u(D ∧ ¬L) = 0.25 . -30 + 0.75. 10 = 0.

de a executar, a sua utilidade causal esperada excede a utilidade causal esperada de todas as suas alternativas. Sendo d(x) a decisão de executar x, a acção de disparar seria ratificável se, e somente se,

a) $u_t(\mathrm{D}|d_t(\mathrm{D})) \geq u_t(\neg\mathrm{D}|d_t(\mathrm{D}))$;

e a acção de não disparar seria ratificável se, e somente se,

b) $u_t(\neg\mathrm{D}|d_t(\neg\mathrm{D})) \geq u_t(\mathrm{D}|d_t(\neg\mathrm{D}))$.

Para que D seja ratificável, a $u_t(\mathrm{D}|d_t(\mathrm{D})) \geq 0$, e para que ¬D seja ratificável, a $u_t(\mathrm{D}|\, d_t(\neg\mathrm{D})) \leq 0$.

Se as relações evidenciais e causais entre disparar e ter a lesão se mantiverem após considerada toda a informação, então a $pr_t(\mathrm{L}|\mathrm{D}) > pr_t(\mathrm{L}|\neg\mathrm{D})$. Como sabemos que, em equilíbrio, a $pr_t(\mathrm{L}) = 0.25$, então a $pr_t(\mathrm{L}|\mathrm{D}) > 0{,}25 > pr_t(\mathrm{L}|\neg\mathrm{D})$.[82] Do cálculo,

$u_t(\mathrm{D}|d_t(\mathrm{D})) = -30 \times pr_t(\mathrm{L}|\mathrm{D}) + 10 \times pr_t(\neg\mathrm{D}|\mathrm{L}) = 10 - 40 \times pr_t(\mathrm{L}|\mathrm{D})$,

segue-se que, quando $pr_t(\mathrm{L}|\mathrm{D}) > 0{,}25$, a $u_t(\mathrm{D}|d(\mathrm{D})) < 0$. Logo, (D) não é ratificável.[83] Por outro lado, do cálculo,

$u_t(\mathrm{D}|d_t(\neg\mathrm{D})) = -30 \times pr_t(\mathrm{L}|\neg\mathrm{D}) + 10 \times pr_t(\neg\mathrm{L}|\neg\mathrm{D}) = 10 - 40 \times pr_t(\mathrm{L}|\neg\mathrm{D})$,

segue-se que, quando $pr_t(\mathrm{L}|\neg\mathrm{D}) < 0{,}25$, a $u_t(\mathrm{D}|d_t(\neg\mathrm{D})) > 0$. Logo, ¬D também não é ratificável.

Chegamos, assim, a um ponto em que, avaliando a racionalidade das acções disponíveis, a situação se mostra pouco salutar. Em princípio, as acções não-ratificáveis são acções sub-óptimas, ou seja, acções cuja utilidade causal esperada, dada a suposição de que são escolhidas, é inferior à de outra ou outras acções disponíveis. Contudo, estamos aqui na

[82] Após a consideração de toda a informação disponível, a pr(L) apenas aumentou 0,05, daí ser plausível que em t se mantenham as relações evidenciais e causais entre ter a lesão e disparar.

[83] Neste cálculo estamos a supor que, por exemplo, a $pr_t(\mathrm{L}|\mathrm{D}) = pr_t(\mathrm{L}|\mathrm{d}(\mathrm{D}))$, ou seja, que a decisão de disparar é tão boa indicadora da posse da lesão quanto a própria acção em si, partindo do princípio que a lesão influencia o agente através das crenças e desejos deste.

presença de duas acções não-ratificáveis em que a utilidade causal ‘incondicional’ de ambas, após consideração de toda a evidência, é óptima (igual a 0).

Uma característica fundamental das acções não-ratificáveis consiste no arrependimento que resulta da decisão de escolher executá-las. A meu ver, o arrependimento e a não-ratificabilidade são dois conceitos de tal modo interligados que, do ponto de vista da teoria da decisão, um não pode ser concebido sem o outro. Existem várias situações em que, de acordo com uma maneira pouco rigorosa de falar, se considera que pode surgir arrependimento:

1. Quando o agente toma uma decisão resultante de um impulso, sem consideração adequada das suas possíveis consequências. O arrependimento que pode surgir deste tipo de decisão nada tem que ver com a teoria da decisão racional.
2. Quando o agente desdenha a possibilidade de, numa ocasião futura, poder vir a obter informação adicional sobre o estado do mundo, nada lhe custando esperar para obter essa informação. Este seria o caso de um jogador de poker que fizesse uma aposta irrevogável sem que as cartas em cima da mesa estivessem todas viradas. O arrependimento, neste caso, também nada tem que ver com a teoria da decisão racional.
3. Quando, apesar de agir de acordo com as suas crenças e desejos, maximizando a utilidade esperada, a acção tomada acaba por não ser a melhor. Esta é uma hipótese que o agente racional tem sempre de considerar. Se tal vier a suceder, provando que as suas crenças estavam erradas, ele sabe que fez tudo quanto pôde, e não existem motivos para arrependimento.

Finalmente, a única forma de arrependimento que se encontra relacionada com a aplicação da teoria da decisão é a seguinte:

4. Quando o agente decide executar uma acção não-ratificável, vindo mais tarde a saber que uma acção alternativa tinha uma maior utilidade esperada. Ou, de outro modo, quando o agente ignora a evidência, acerca da utilidade de todas as acções disponíveis, que a sua própria decisão lhe pode oferecer.

Joyce admite que, na Lesão Assassina, ambas as acções, embora não-ratificáveis, são permissíveis. Admitindo que, neste caso, a não-ratificabilidade modela teoricamente o raciocínio oscilatório de uma acção para outra, ele classifica essa mesma situação como um equilíbrio saudável, pois toda a informação disponível está já a ser considerada. Ou seja, essa oscilação estará presente até se atingir um estado epistémico em que o ímpeto de disparar é contrabalançado pelo ímpeto contrário de não disparar. Esse equilíbrio é atingido, precisamente, quando *pr*(L) = 0.25:

'Com efeito, uma vez alcançado o equilíbrio no qual toda a informação disponível acerca dos efeitos das nossas acções é tida em consideração, as nossas utilidades esperadas causais, incondicionadas, incorporam todas as considerações ratificacionistas relevantes' (Joyce 2012: 138).

O racional desta posição é o seguinte: apenas se chega a *pr*(L) = 0.25, após se ter em conta, como foi visto acima, a evidência que a decisão de disparar em t_0 nos oferece. Se *pr*(L) fosse superior a 0.25, não disparar seria a única acção recomendada; se *pr*(L) fosse inferior a 0.25, disparar seria a única acção recomendada. O raciocínio oscilatório, que caracteriza a não-ratificabilidade das acções, está já implicitamente contido em equilíbrio, quando *pr*(L) = 0.25.

Aquilo que aparentemente não se pode negar é que, seja qual for a decisão tomada, a possibilidade de arrependimento encontra-se sempre presente, o que lança sérias dúvidas sobre a permissibilidade de exectuar qualquer uma das acções disponíveis. Segundo Joyce, a indecisão acerca de qual a decisão a tomar é transferida para a incerteza acerca da possibilidade de se vir a sentir, legitimamente, arrependimento: uma probabilidade de 0.25 de se vir a sentir arrependimento, caso se decida disparar, e uma probabilidade de 0.75 de se vir a sentir arrependimento, caso se decida não disparar. Mas o essencial do argumento de Joyce consiste, a meu ver, em chamar a atenção para as possibilidades contrárias:

'Temos a certeza do seguinte: "Existe uma acção *S* ou ¬*S*, não sabemos qual delas, tal que, se decidirmos executá-la, a legitimidade do subsequente arrependimento é garantida". Contudo, também temos a certeza do seguinte: "Existe uma acção *S* ou ¬*S*, não sabemos qual delas, tal que, se decidirmos executá-la, não temos legitimidade para vir a sentir arrependimento"' (Joyce 2012: 140).

Isto, de certa forma, voltaria a empatar o jogo, pois o agente teria uma probabilidade de de 0.75 de não vir a sentir arrependimento, caso decidisse disparar, e uma probabilidade de 0.25 de não vir a sentir a sentir arrependimento, caso decidisse não disparar. Esta incerteza, embora garantindo que não há uma razão decisiva para se optar por qualquer uma das acções, pois cada uma delas tem uma probabilidade positiva de vir a ser aquela da qual o agente não se arrependerá, tornaria ambas as acções racionalmente permissíveis. Ora, esta observação de Joyce não me parece muito convincente – alguns talvez a qualifiquem com meramente retórica - pois afinal uma das acções disponíveis apresenta uma probabilidade muito maior do que a outra de gerar arrependimento. Penso, contudo, que Joyce tem outras razões para se sentir optimista. Considere-se o seguinte: esta antecipação do arrependimento que supostamente se poderá vir a sentir, tanto com respeito a uma, como a outra das acções, não tem qualquer influência sobre a utilidade de qualquer uma das acções disponíveis. Temos, assim, todas as condições necessárias para escaparmos à situação descrita em 4 (da qual depende a legitimidade desse arrependimento), pois, em condições de equilíbrio, a decisão de executar qualquer uma das acções não faz com que a acção contrária tenha uma maior utilidade esperada (sabemos, aliás, que essa utilidade é idêntica). De acordo com esta interpretação, dever-se-ia sustentar que nenhuma das acções disponíveis provoca arrependimento, e que essa antecipação de remorsos não tem qualquer legitimidade. Na verdade, se lermos 4 com atenção, constatamos que não passa de uma definição de ratificabilidade, na qual não se faz qualquer referência às consequências efectivas das decisões tomadas. Portanto, se, no contexto da teoria da decisão, o conceito de arrependimento for definido através do conceito de ratificabilidade, e se se verificar que as condições para o surgimento de arrependimento não estão preenchidas, seja qual for a acção, então, para efeitos práticos, afirmar que ambas as acções não são ratificáveis não é diferente de dizer que ambas o são. Considerando, assim, que ambas as acções podem ser legitimamente consideradas como ratificáveis, não só a teoria causal escapa a qualquer acusação de inconsistência, como também escapa à acusação de incompletude.

11.2. Visualização de mundos-possíveis

Qual é a conclusão que se deve retirar das conclusões a que chegámos no que respeita à normatividade da teoria? A permissão de qualificar ambas as acções como ratificáveis, quando em equilíbrio, fará com que possamos realmente considerá-las como permissíveis? Ou será que, pelo contrário, a confiança no princípio da ratificabilidade, enquanto guia fiável para averiguar da racionalidade das acções, deve ser posta em causa?

Creio que Joyce tem razão ao afirmar que quaisquer considerações relacionadas com a ratificabilidade estão já contidas no equilíbrio alcançado ao condicionalizar-se sobre a utilidade esperada inicial. Ou seja, quando a $pr(\mathrm{L}) = 0.25$, já se está a incluir a revisão da crença inicial acerca da probabilidade de se ter a lesão. Contudo, neste ponto da discussão, as razões que justificariam a permissibilidade de executar qualquer uma das acções não me parecem mais fortes do que aquelas que desautorizam essa possibilidade. Uma outra maneira de expressar esta incerteza poderá passar por colocar em causa a adequação da condicionalização, enquanto método para revisão de crenças. Afinal, trata-se de um método que reproduz, ao nível teórico, um raciocínio oscilatório, ratificacionista, que, embora natural, nos conduz aparentemente a um beco sem saída. Existirá algum outro método que nos permita sair deste impasse e que, ao mesmo tempo, reproduza igualmente uma forma natural de raciocinar?

Esse método existe e pressupõe uma semântica de contrafactuais, tal como a que foi considerada anteriormente (ver §7.1): $x \gg y$ é verdadeira se, e somente se, y for verdadeira no mundo possível w mais próximo em que x é verdadeira. Esta semântica pressupõe que existe, de facto, um e um único mundo w que é o mais semelhante ao mundo actual, e nisto difere da semântica de Van Fraassen. David Lewis (1976) provou que é possível calcular a probabilidade de uma proposição contrafactual $x \gg y$, com um antecedente possível x, através da igualdade $x \gg y = pr^{x}(y)$; ou, segundo a terminologia de Lewis, a probabilidade de $x \gg y$ é igual à probabilidade do consequente, após a *visualização* [*imaging*] do antecedente. Segundo Lewis, esta maneira de calcular a probabilidade de contrafactuais pode também ser interpretada como um método de actualização de crenças:[84]

[84] Considere-se: 'The only way to have a probability-revision conditional is to interpret the conditional in Stalnaker's way and revise by imaging. (…) The requirements were given by Stalnaker for revision of worlds, but they can carry over *mutatis mutandis* to revision of probability functions' (Lewis 1976: 313).

$pr^x(\mathrm{y}) = \sum_{\mathrm{w} \in W} pr(\mathrm{w}) \times pr(\mathrm{y}|\mathrm{w}_x)$,

em que *W* é a partição dos mundos possíveis relevantes para o problema em questão, *pr* é uma distribuição de probabilidades sobre esses mundos possíveis, e w_x é o mundo-*x* mais semelhante a *w*.[85]

No caso da Lesão assassina, temos os seguintes quatro mundos possíveis:[86]

$pr(\mathrm{w}_1) = \mathrm{D} \gg \mathrm{L} / (\mathrm{D} \wedge \mathrm{L}) = 0.16$

$pr(\mathrm{w}_2) = \mathrm{D} \gg \neg\mathrm{L} / (\mathrm{D} \wedge \neg\mathrm{L}) = 0.08$

$pr(\mathrm{w}_3) = \neg\mathrm{D} \gg \mathrm{L} / (\neg\mathrm{D} \wedge \mathrm{L}) = 0.04$

$pr(\mathrm{w}_4) = \neg\mathrm{D} \gg \neg\mathrm{L} / (\neg\mathrm{D} \wedge \neg\mathrm{L}) = 0.72$

Calcule-se, então, as probabilidades destas quatro contrafactuais:

[85] Lewis demonstrou exactamente o seguinte: dado o pressuposto de Stalnaker, de que existe realmente um, e apenas um, mundo possível mais próximo, a $pr^x(\mathrm{y})$ será sempre igual a $pr(\mathrm{x} \gg y)$. Isto significa que, podendo fazer-se uma distribuição de probabilidades sobre mundos possíveis, a probabilidade não tem de ser toda transferida para um único desses mundos. Existem, contudo, alguns casos com os quais é difícil lidar; por exemplo, o mundo actual pode não se encontrar suficientemente determinado para que eu possa atribuir uma probabilidade diferente de zero à contrafactual 'se eu chocasse com outro carro, este seria um Seat Ibiza'. Ou seja, a relação de semelhança entre mundos possíveis pode ser demasiado vaga para me permitir identificar um qualquer mundo possível que torne esta proposição verdadeira. Contudo, uma interpretação correcta da *visualização* (Joyce 1999: 172-176) permite-nos contornar este problema. Devemos fazer uma distinção entre duas operações epistémicas: uma delas consiste em atribuir uma probabilidade subjectiva incondicional a $\mathrm{x} \gg y$, a qual pode, como no caso acima, ter valor 0; e outra que consiste em determinar o quão provável é y, quando suponho contrafactualmente, ou *visualizo*/*imagino*, que x é verdadeira; neste último caso, eu posso atribuir à mesma proposição $\mathrm{x} \gg y$ uma probabilidade diferente de 0, ainda que, possivelmente, bastante diminuta. Isto significa que, de acordo com esta interpretação, a $pr^x(\mathrm{y}) \geq p(\mathrm{x} \gg y)$. Quando x é verdadeiro no mundo actual, então as duas probabilidades serão sempre idênticas e terão valor 1. Este problema inerente à caracterização *contrafactual* das utilidades causais é importante e tem de ser resolvido pelos defensores da teoria causal (ver §9.2, n. 67). Contudo, em todos os exemplos considerados neste trabalho, incluindo o PN, o pressuposto de Stalnaker é sempre verdadeiro, e, portanto, nunca se verifica qualquer discrepância entre as duas operações epistémicas acima consideradas.

[86] Este modelo pode ser utilizado para o cálculo de contrafactuais, seja qual for a interpretação que é feita das mesmas: standard ou retroactiva. Conforme uma ou outra, as relações entre os vários mundos possíveis *y* e o mundo-*x* mais próximo serão distintas. Na Lesão Assassina, dado que as acções, disparar ou não disparar, não têm influência causal sobre a posse, ou a falta dela, da lesão, a interpretação das contrafactuais é a que segue a natural causalidade, ou a ausência dela, do antecedente para o consequente; ou seja, é a relação standard que se verifica neste caso.

$$pr^{D}(L) = pr(w_1) \times pr(L|w_1) + pr(w_2) \times pr(L|w_2) +$$
$$+ pr(w_3) \times pr(L|w_1) + pr(w_4) \times pr(L|w_2) =$$
$$= 0.16 \times 1 + 0.08 \times 0 +$$
$$+ 0.04 \times 1 + 0.72 \times 0 = 0.2$$

$$pr^{D}(\neg L) = pr(w_1) \times pr(\neg L|w_1) + pr(w_2) \times pr(\neg L|w_2) +$$
$$+ pr(w_3) \times pr(\neg L|w_1) + pr(w_4) \times pr(\neg L|w_2) =$$
$$= 0.16 \times 0 + 0.08 \times 1 +$$
$$+ 0.04 \times 0 + 0.72 \times 1 = 0.8$$

$$pr^{\neg D}(L) = pr(w_1) \times pr(L|w_3) + pr(w_2) \times pr(L|w_4) +$$
$$+ pr(w_3) \times pr(L|w_3) + pr(w_4) \times pr(L|w_4) =$$
$$= 0.16 \times 1 + 0.08 \times 0 +$$
$$+ 0.04 \times 1 + 0.72 \times 0 = 0.2$$

$$pr^{\neg D}(\neg L) = pr(w_1) \times pr(\neg L|w_3) + pr(w_2) \times pr(\neg L|w_4) +$$
$$+ pr(w_3) \times pr(\neg L|w_3) + pr(w_4) \times pr(\neg L|w_4) =$$
$$= 0.16 \times 0 + 0.08 \times 1 +$$
$$+ 0.04 \times 0 + 0.72 \times 1 = 0.8$$

Sabendo-se as utilidades das quatro consequências possíveis,

$$u(D \wedge L) = -30;\ u(D \wedge \neg L) = 10;\ u(\neg D \wedge L) = 0;\ u(\neg D \wedge \neg L) = 0,$$

podemos calcular a utilidade causal esperada das duas acções, à qual podemos chamar de ‘utilidade contrafactual esperada’, correspondendo esta a uma actualização de crenças a partir das probabilidades iniciais:

$$\text{UctfE}(D) = pr(D \gg L) \times u(D \wedge L) + pr(D \wedge \neg L) \times u(D \wedge \neg L) =$$
$$= 0.2 \times -30 + 0.8 \times 10 =$$
$$= 2$$

$$\begin{aligned}\text{UctfE}\,(\neg D) &= pr(\neg D \gg L) \times u(\neg D \wedge L) + pr(\neg D \gg \neg L) \times u(\neg D \wedge \neg L) = \\ &= 0.2 \times 0 + 0.8 \times 0 = \\ &= 0\end{aligned}$$

Constata-se, assim, sem surpresa, que a aplicação do método de Lewis para a actualização de crenças resulta numa recomendação de disparar. Este resultado, ao questionar a legitimidade da intuição prevalecente de que não disparar é a acção racional, não parece, à primeira vista, ter muito a seu favor. Mas, tal como já foi mencionado, não é tanto a intuição de irracionalidade que está em causa e que deve ser justificada. Afinal, as intuições valem o que valem. O que está verdadeiramente em causa é a instabilidade deliberativa que resulta da aplicação do princípio ratificacionista. A ideia de que existe um impasse, resultante da aplicação da teoria, faz com que esta saia desacreditada e com que esse resultado assuma contornos de irracionalidade. Ao permitir resolver esse impasse, impedindo o bloqueio da 'máquina de decisão' – ou repondo o funcionamento do mecanismo – a *visualização* dissipa essa aura de irracionalidade.

Poder-se-ia questionar este resultado afirmando que, ao contrário do que acontecia com o *botão do psicopata*, a desutilidade da consequência negativa, falhar o disparo, não é suficientemente elevada para justificar a intuição de irracionalidade associada à acção de disparar. Contudo, o exemplo pode ser modificado, supondo-se que a consequência da falta de pontaria é a condenação à morte e, desde que o incentivo seja suficientemente grande, e as probabilidades suficientemente favoráveis, a teoria causal pode sempre recomendar que se dispare. Portanto, desde que não atribuamos à mais fatal das consequências uma desutilidade infinita – e não há grandes razões para acreditar que as pessoas em geral o façam – a teoria causal recomenda uma acção que, afinal, é a única que permite obter o resultado *óptimo*.

Para que este resultado possa ser adoptado como resolução do problema, é necessário justificar a razão pela qual se deve, neste caso, actualizar as crenças através do método de Lewis, e não através da condicionalização. Desde logo, é necessário explicar por que motivo a *visualização* consiste num método de revisão de crenças.

O raciocínio subjacente a este método é, como sabemos, do tipo hipotético. Isto significa que funciona através de suposições: supõe-se a verdade do antecedente e verificamos de que maneira, num mundo minimamente alterado para incluir essa modificação, essa suposição afecta a verdade do consequente. Uma suposição pode, então, ser interpretada

como uma forma de revisão de crenças provisória. Ao supormos D, estamos a agir *como se* acreditássemos que D é verdadeiro. Existem, contudo, vários tipos de suposição, assim como existem vários tipos de crença condicional: podemos supor de forma indicativa ou de forma conjuntiva/contrafactual. Considere-se o seguinte exemplo:

'Se Oswald não matou Kennedy, então alguém o matou',
'Se Oswald não tivesse matado Kennedy, então alguém o teria feito'.

No primeiro caso, supomos como as coisas *são* se *for* verdade que Oswald não matou Kennedy; no segundo, supomos como as coisas *seriam* se *fosse* verdade que Oswald não tivesse matado Kennedy. A diferença entre estas duas suposições torna-se notória quando consideramos que a primeira delas é claramente verdadeira, enquanto a segunda, se acreditarmos que não existiu uma conspiração, pode ser falsa.
Como deveremos interpretar, em termos probabilísticos, a relação entre a nova crença, ou suposição, e o modo como a estrutura da realidade é por ela alterada? Consideremos o exemplo de §1.2: uma das consequências da Teoria Geral da Relatividade era a de que a trajectória da luz encurvaria perto do Sol. Em 1919, foi possível observar este fenómeno. A constatação deste facto deverá, então, constituir evidência favorável à confirmação da teoria de Einstein. Ou seja, a nossa crença na verdade da teoria deverá ser revista e, neste caso, reforçada. É assim que funciona o processo confirmatório bayesiano, utilizado para a revisão das nossas crenças, o qual constitui um requisito de racionalidade no contexto bayesiano. Disto se segue que a relação probabilística que se verifica entre a suposição indicativa - neste caso a verificação de um facto - e o modo como essa suposição modifica a nossa visão do mundo, é do tipo evidencial. A evidência que o novo facto constitui faz aumentar a probabilidade subjectiva que atribuímos à teoria. Segundo Lewis:

'Se o antecedente fosse realmente adicionado, dever-se-ia (se possível) rever através da condicionalização. As razões a favor de uma resposta à nova evidência baseada na condicionalização são igualmente razões contra uma resposta baseada na visualização' (Lewis 1976: 312).

O que se pode dizer, então, acerca da suposição contrafactual? Neste caso, aquilo que o agente deverá ter em consideração, de modo a avaliar o impacto da suposição na sua visão

do mundo, são as semelhanças e as diferenças entre as situações possíveis nas quais essa suposição é temporariamente aceite como verdadeira. Ora, todos os defensores da teoria causal defendem que o único aspecto relevante da suposição contrafactual, quando esta se refere à realização de uma acção, é a eficácia causal dessa acção sobre o mundo:

'Após completada a operação epistémica que consiste em supor *A* [neste caso, D] desta maneira conjuntiva, o agente calcula, então, a utilidade de *A* [neste caso D], com base nesta nova crença provisória, de modo a obter o valor respeitante à sua eficácia' (Joyce 1999: 175).

Portanto, ao calcularmos a probabilidade de contrafactuais - como D ≫ L – através da visualização – supor D e visualizar o mundo-D mais próximo - estamos já a ter em conta tudo o que é relevante para determinar a eficácia causal de D sobre o estado do mundo L – neste caso nenhuma. A relação probabilística entre a suposição e o modo como o mundo se modifica não pode ser do tipo evidencial; essa relação evidencial consiste precisamente naquilo que desejamos eliminar do cálculo da utilidade causal da acção suposta. Essa relação probabilística tem, portanto, de ser do tipo causal.

Não se trata aqui, como antes, de calcular a utilidade causal esperada e, depois, averiguar que tipo de evidência nos oferece o resultado desse cálculo para determinarmos com mais rigor a probabilidade do estado do mundo relevante. Aqui, a revisão temporária de crenças já está a ser incluída no próprio cálculo da utilidade causal esperada, tornando todo o processo mais coeso e natural. Também aqui são dispensáveis quaisquer considerações relacionadas com a ratificabilidade das opções, pois a evidência acerca da lesão, que é oferecida pela própria decisão, não é relevante em termos de eficácia causal (mesmo tendo em conta que a lesão é causa do disparo falhado), ao contrário do que acontecia na condicionalização. Ou seja, apesar de a probabilidade condicional de se ter a lesão, dado o disparo, ser elevada, a probabilidade de se ter a lesão, na suposição contrafactual do disparo, é baixa, pois a crença de Maria é a de que ela própria não é uma assassina. Em suma, não parece, assim, existir uma razão forte para aderirmos a um princípio de decisão, e a um tipo subjacente de raciocínio, que conduz a situações de instabilidade deliberativa e a aparentes becos sem saída. [87]

[87] É necessário mencionar que Teller (1976) concebeu um argumento pragmático favorável à condicionalização, com o objectivo de estabelecer solidamente este método como um requisito de

Importa também considerar as preferências do agente no caso da lesão assassina e, dessa maneira, ensaiar um argumento pragmático a favor da revisão por *visualização*: dado que disparar é a única acção que, à partida, permite ao agente obter a sua consequência preferida, então este deverá optar pelo método de actualização de crenças que lhe permitirá satisfazer as suas preferências. Isto não é contraditório com os dados do problema, nomeadamente, com a forte crença de Maria em como não é uma assassina. Afinal, nada nos indica que Maria atribui qualquer especial utilidade ao facto de agir moralmente. Não ter fibra para ser assassino é perfeitamente compatível com o desejo de que uma morte favorável aconteça.

Este argumento pragmático parece, à primeira vista, incorrer em circularidade: afinal uma teoria não deve ser julgada pelo facto de recomendar uma acção que augura uma consequência positiva ou que satisfaz a nossa intuição de racionalidade. A teoria não é verdadeira porque a acção é racional; a acção é racional porque a teoria é verdadeira. Contudo, uma teoria da decisão pode perfeitamente estar sujeita a restrições pragmáticas. Se, em certas situações, uma determinada versão da teoria não permite alcançar aquilo para o qual foi concebida – seleccionar uma acção como a única que é racional de entre um conjunto exaustivo de acções mutuamente exclusivas – existe, então, um bom motivo para duvidar da sua correcção, ainda mais quando existe uma versão alternativa capaz de produzir resultados.

Vimos atrás que a teoria bayesiana da decisão não exige dos agentes que estes ajam calculando a sua utilidade esperada (ver §3.1). Este não era, portanto, um requisito de racionalidade da teoria. O que esta exigia era que os agentes ordenassem as suas preferências de acordo com certas restrições estruturais, cujo apelo intuitivo era bastante forte, daí a sua designação como axiomas. Esta parece ser a única maneira de salvaguardar o valor normativo da teoria e de impor requisitos de racionalidade minimamente realistas. Ora, do mesmo modo que uma teoria da decisão pode ser avaliada de acordo com o seu

racionalidade. Mais precisamente, foi possível construir um *dutch book* diacrónico, estruturalmente equivalente aos argumentos *money-pump* atrás considerados, mostrando que um agente que não reveja as suas crenças através da condicionalização, encontra-se sujeito a ser explorado. De forma a acomodarmos este facto, é preciso conhecer o contexto em que esse *dutch book* está a ser aplicado. Se a informação que o agente tem de incorporar na sua revisão de crenças for do tipo evidencial, então é um requisito de racionalidade actualizar por condicionalização; se for do tipo causal, então é um requisito de racionalidade actualizar por *visualização*. Existe, de qualquer modo, uma outra razão, sugerida por Cantwell (2010), para continuar a defender a revisão de crenças através da visualização: a utilidade da consequência favorável da acção recomendada irá compensar as eventuais perdas sofridas pelo *dutch book*. Seria certamente bem-vinda uma prova de que este tipo de compensação é suficiente para contrabalançar as eventuais perdas incorridas no *dutch book*, em todos aqueles casos em que se revê através da visualização.

valor normativo, o mesmo acontece com uma teoria para a actualização de crenças. Como se viu, a teoria bayesiana da confirmação diz-nos que novo grau de crença é racional adoptar quando obtemos novos dados observacionais que confirmam as nossas hipóteses, tendo em conta os nossos graus de crença anteriores. Trata-se, pois, de uma teoria que nos permite atribuir um valor quantitativo preciso ao modo como um determinado corpo de informação observacional serve para confirmar racionalmente as nossas hipóteses científicas. O seu contexto de aplicação por excelência é, portanto, o da investigação científica.

Todavia, o contexto em que se coloca o nosso problema é o da decisão e da actualização das nossas crenças acerca de qual é a acção que melhor satisfará os nossos interesses. Seria errado apresentar o método utilizado por Joyce, a condicionalização sobre o resultado do cálculo da utilidade, como uma descrição da maneira como os agentes de facto raciocinam ou como uma exigência de racionalidade para a deliberação acerca do que fazer. Estamos, isso sim, perante um modelo teórico do tipo de raciocínio que caracteriza o princípio ratificacionista. Do mesmo modo, o método da visualização constitui um modelo teórico de outra forma comum de raciocínio, empregue num contexto deliberativo, no qual os indivíduos ponderam as consequências das suas acções. A questão que se pode colocar é a de saber qual dos dois métodos constitui, neste contexto, a exigência de racionalidade mais adequada. Enquanto a condicionalização, como vimos, procura evidência acerca do tipo de pessoa que somos – tal como indica o conceito de *self-signaling* - a visualização procura, fundamentalmente, descobrir a eficácia causal das nossas acções quanto à produção das melhores consequências. Se num contexto de deliberação moral, a primeira pode assumir uma considerável importância, num contexto de deliberação instrumental a segunda parece ser mais prevalecente, ainda para mais quando António já possui uma crença forte em como não é psicopata, e Maria já possui uma crença forte em como não é assassina. Em suma, uma teoria causal da decisão apresentará como exigência de racionalidade a suposição contrafactual e o método de actualização de crenças através da visualização de mundos possíveis. Os argumentos num problema de decisão encontram-se aliás formulados de uma maneira que capta esta forma natural de raciocínio: 'o que aconteceria se eu fizesse x? E o que aconteceria se fizesse y'?

Resumindo, temos a seguinte generalização do método da condicionalização:

se pr_t corresponde à minha crença em *t*, e se pr_{t1} corresponde à minha crença em t_1, e se entre *t* e t_1 adquiri a crença *x*, então para qualquer *y*, $pr_{t1}(y) = pr_t(y|x)$.

E uma generalização do método de actualização de crenças através da *visualização*:

se pr_t corresponde à minha crença em *t*, e se pr_{t1} corresponde à minha crença em t_1, e se entre *t* e t_1 adquiri a crença *x*, então para qualquer *y*, $pr_{t1}(y) = pr_t(x \gg y) = pr^x(y)$.[88]

Um modelo adequado da actualização de crenças deverá representar adequadamente as relações de evidência epistémica e de evidência causal. Em muitos casos, os dois métodos obtêm o mesmo resultado. Isto acontece pelo seguinte motivo: dada uma acção *x* e um estado *y*, quando a contrafactual $x \gg y$ é epistemicamente independente de *x*, então a $pr(x \gg y)$ é igual a $pr(y|x)$. Contudo, no caso da Lesão Assassina, a $pr(x \gg y|x) \neq pr(x \gg y)$, nomeadamente, a primeira é maior do que a segunda, pois ao tomar conhecimento da minha decisão de disparar, a probabilidade que atribuo à crença em como tenho a lesão aumenta.

Portanto, embora a condicionalização seja capaz de captar as nuances das relações de evidência epistémica, apenas a *visualização* é sensível tanto às relações de correlação sem causação, como também aos casos em que *x* é a causa de *y*. Tal como o *rationale* de Stalnaker para introduzir as contrafactuais no modelo de decisão era o de imunizar o cálculo da utilidade contra as relações de correlação sem causação, também aqui o intuito é o mesmo: imunizar a revisão de crenças contra as relações de correlação sem causação. Segundo Joyce (1999: 175):

'A visualização é a operação correcta para calcular os valores de eficácia. Obtemos o valor correcto para U(*A*) identificando $P(A \gg B)$ com $P^A(B)$ na teoria causal da decisão. Um decisor que avalia o valor de eficácia de *A* irá modificar provisoriamente as suas opiniões, supondo que *A* é verdadeira e redistribuindo a sua probabilidade subjectiva pelos mundos ¬*A* e *A*, de uma maneira que esteja de acordo com os seus juízos acerca da similitude geral, relevante, entre mundos. (…) o requisito básico consiste em atribuir um menor peso adicional aos mundos tidos como "menos semelhantes ao mundo actual" do

[88] James Cantwell (2010) defende também que a actualização deve ser feita por *imaging*; contudo, ele defende que este método deve utilizar a probabilidade de condicionais indicativas e não a probabilidade de contrafactuais. As suas razões prendem-se com a sua própria semântica das condicionais em geral.

que aos mundos "mais próximos da actualidade". (…) Após completada a operação epistémica de supor *A* deste modo conjuntivo, o agente calcula a utilidade de *A*, relativamente a este novo estado de crença provisório, de modo a obter o seu valor de eficácia'.

Foi precisamente este o procedimento que se efectuou acima e cujo resultado consistiu numa recomendação de disparar, permitindo-nos sair do impasse da (não-)ratificabilidade de ambas as acções. Joyce encontra-se a descrever a maneira como ele acha que se deve calcular a utilidade causal de uma acção em geral (ver n. 85), e não qualquer modo que ele considere adequado para a actualização de crenças num problema de decisão. Como vimos, ele é um defensor da condicionalização. Contudo, a contenção aqui é a de que o mesmo pode ser dito acerca da actualização de crenças, a qual deverá funcionar através do método da suposição contrafactual. Ou seja, aquando do cálculo da utilidade causal de uma acção, estão já a ser tidas em conta as nossas crenças acerca do poder causal das nossas acções sobre os aspectos do mundo relevantes, na suposição de que estas seriam realizadas. Como se viu, isto constitui, como afirma Joyce, uma actualização provisória de crenças.

Portanto, desde que as relações entre as acções e estados num dado problema de decisão sejam adequadamente representadas por condicionais contrafactuais, o método de revisão de crenças deverá ser a *visualização*. Por outro lado, se a preservação da teoria evidencial dependia da completa fiabilidade do princípio ratificacionista, passamos agora a ter boas e melhores razões para declaramos a superioridade da teoria causal sobre aquela.

12. Agir mudando o passado

12.1. Tipos de dominação

Agora que estamos na posse de um método para calcular a utilidade esperada das acções que não só é sensível às relações de causalidade entre acções e estados – e, por conseguinte, à ausência das mesmas – mas que também captura de forma natural, através da suposição contrafactual, o processo de revisão de crenças, podemos agora aplicá-lo ao PN. A grande vantagem, convém notar, consiste em poder atribuir-se probabilidades a

contrafactuais de uma maneira que não nos parece forçada ou arbitrária, mas sim através de um procedimento que modela o raciocínio natural empregue no processo deliberativo. A existência deste procedimento milita a favor da teoria causal, contra aqueles que a criticam pela desnecessária inserção de um conceito de tão difícil análise, como o é o de causalidade, no contexto da avaliação da racionalidade das acções.

Considerem-se, então, as probabilidades dos mundos possíveis que constituem a partição *K* - as *dependency hypothesis* de Lewis - os quais representam a nossa visão da estrutura causal do mundo. Essas probabilidades reflectem a elevada taxa de sucesso do previsor. Seja novamente C1 a acção de escolher a caixa opaca, C2 a acção de escolher as duas caixas, PC1 a previsão de caixa opaca e PC2 a previsão de duas caixas:

$pr\ (w_1) = C1 \gg PC1\ /\ (C1 \wedge PC1) = 0.45$
$pr\ (w_2) = C1 \gg PC2\ /\ (C1 \wedge PC2) = 0.05$
$pr\ (w_3) = C2 \gg PC1\ /\ (C2 \wedge PC1) = 0.05$
$pr\ (w_4) = C2 \gg PC2\ /\ (C2 \wedge PC2) = 0.45$

Como sabemos, estas contrafactuais podem ser interpretadas de duas maneiras diferentes para efeitos do cálculo da sua probabilidade: de acordo com a resolução standard ou de acordo com a resolução retroactiva. Em primeiro lugar, vamos interpretá-las de maneira standard, aquela que, segundo o bicaixista, é a mais apropriada pragmaticamente; ou seja, aquela que tem como critérios fundamentais de resolução da vagueza, ou de proximidade entre mundos possíveis, a necessidade de evitar grandes violações das leis da natureza e a maximização da região do espaço-tempo em que os factos particulares coincidem. No contexto do PN, estes critérios estão relacionais com o valor de verdade do consequente, ou a correcção da previsão, dada a realização de uma ou outra acção. Como numa interpretação standard, seja qual for a acção realizada, o passado se mantém idêntico, um certo mundo-C2 tem como mundo-C1 mais semelhante aquele em que o consequente de C2 é idêntico ao consequente de C1; ou, de outra maneira, aquele em que o passado se mantém idêntico.

Considere-se, por exemplo, o cálculo da $pr\ (w_1) = C1 \gg PC1$, imediatamente abaixo. Dos dois mundos-C1, aquele que é mais semelhante a w_4 é aquele em que o consequente é idêntico ao consequente de w_4, ou seja, w_2. Logo, temos como última parcela da soma

abaixo: $pr\ (w_4) \times pr(PC1|w_2)$. Calcule-se, então, a probabilidade dos quatro mundos possíveis de acordo com uma interpretação standard/causal das respectivas contrafactuais:

$$pr^{C1}(PC1) = pr\ (w_1) \times pr(PC1|w_1) + pr\ (w_2) \times pr(PC1|w_2) +$$
$$+ pr\ (w_3) \times pr(PC1|w_1) + pr\ (w_4) \times pr(PC1|w_2) =$$
$$= 0.45 \times 1 + 0.05 \times 0 +$$
$$+ 0.05 \times 1 + 0.45 \times 0 = 0.5$$

$$pr^{C1}(PC2) = pr\ (w_1) \times pr(PC2|w_1) + pr\ (w_2) \times pr(PC2|w_2) +$$
$$+ pr\ (w_3) \times pr(PC2|w_1) + pr\ (w_4) \times pr(PC2|w_2) =$$
$$= 0.45 \times 0 + 0.05 \times 1 +$$
$$+ 0.05 \times 0 + 0.45 \times 1 = 0.5$$

$$pr^{C2}(PC1) = pr\ (w_1) \times pr(PC1|w_3) + pr\ (w_2) \times pr(PC1|w_4) +$$
$$+ pr\ (w_3) \times pr(PC1|w_3) + pr\ (w_4) \times pr(PC1|w_4) =$$
$$= 0.45 \times 1 + 0.05 \times 0 +$$
$$+ 0.05 \times 1 + 0.45 \times 0 = 0.5$$

$$pr^{C2}(PC2) = pr\ (w_1) \times pr(PC2|w_3) + pr\ (w_2) \times pr(PC2|w_4) +$$
$$+ pr\ (w_3) \times pr(PC2|w_3) + pr\ (w_4) \times pr(PC2|w_4) =$$
$$= 0.45 \times 0 + 0.05 \times 1 +$$
$$+ 0.45 \times 0 + 0.45 \times 1 = 0.5$$

Como seria de esperar, os quatro mundos possíveis têm a mesma probabilidade. Isto acontece porque as acções não têm qualquer eficácia causal sobre as previsões, de onde se segue que a probabilidade de cada uma das previsões é independentemente das acções realizadas: a previsão já esta feita e acção nenhuma pode alterá-la.

Como já sabemos, utilizando estas probabilidades para o cálculo da utilidade contrafactual esperada de C1 e C2, a utilidade de C2, a escolha das duas caixas, supera a utilidade de C1, a escolha da caixa opaca:

UctfE (C1) = pr(C1 ≫ PC1) × u(C1 ∧ PC1) + pr(C1 ≫ PC2) × u(C1 ∧ PC2) =
= 0.5 × 1.000.000 + 0.5 × 0 =
= 500.000

UctfE (C2) = pr(C2 ≫ PC1) × u(C2 ∧ PC1) + pr(C2 ≫ PC2) × u(C2 ∧ PC2) =
= 0.5 × 1.001.000 + 0.5 × 1000 =
= 501.000

Considere-se, agora, a probabilidade dos quatro mundos possíveis, de acordo com uma interpretação retroactiva das contrafactuais. Aqui, o critério mais importante de resolução da vagueza será diferente. Agora, de acordo com o argumento monocaixista, o critério pragmaticamente apropriado para determinar a semelhança entre mundos possíveis é a fiabilidade da previsão. Para a salvaguardar, há que sacrificar alguns aspectos do funcionamento normal do mundo, nomeadamente, a preservação do passado. Isto significa que devemos interpretar as contrafactuais como se, por vezes, o antecedente, ou as acções realizadas, tivessem o poder de alterar a verdade do consequente ou da previsão feita anteriormente. Considere-se, como acima, o cálculo da $pr(w_1)$ = C1 ≫ PC1. Dos dois mundos-C1, aquele que é mais semelhante a w_4 é aquele em que, tal como em w_4, a previsão está correcta, ou seja, w_1. Logo, temos como última parcela da soma abaixo: $pr(w_4) \times pr(PC1|w_1)$. Calcule-se, então, a probabilidade dos quatro mundos possíveis relevantes de acordo com uma interpretação retroactiva das respectivas contrafactuais:

pr_r^{C1} (PC1) = pr_r (w_1) × pr(PC1|w_1) + pr_r (w_2) × pr(PC1|w_2) +
+ pr_r (w_3) × pr(PC1|w_1) + pr_r (w_4) × pr(PC1|w_1) =
= 0.45 × 1 + 0.05 × 0 +
+ 0.05 × 1 + 0.45 × 1 = 0.95

pr_r^{C1} (PC2) = pr_r (w_1) × pr(PC2|w_1) + pr_r (w_2) × pr(PC2|w_2) +
+ pr_r (w_3) × pr(PC2|w_3) + pr_r (w_4) × pr(PC2|w_1) =
= 0.45 × 0 + 0.05 × 1 +
+ 0.05 × 0 + 0.45 × 0 = 0.05

$$pr_r^{C2}(PC1) = pr_r(w_1) \times pr(PC1|w_4) + pr_r(w_2) \times pr(PC1|w_4) +$$
$$+ pr_r(w_3) \times pr(PC1|w_3) + pr_r(w_4) \times pr(PC1|w_4) =$$
$$= 0.45 \times 0 + 0.05 \times 0 +$$
$$+ 0.05 \times 1 + 0.45 \times 0 = 0.05$$

$$pr_r^{C2}(PC2) = pr_r(w_1) \times pr(PC2|w_4) + pr_r(w_2) \times pr(PC2|w_4) +$$
$$+ pr_r(w_3) \times pr(PC2|w_3) + pr_r(w_4) \times pr(PC2|w_4) =$$
$$= 0.45 \times 1 + 0.05 \times 1 +$$
$$+ 0.45 \times 0 + 0.45 \times 1 = 0.95$$

Como seria de esperar, as probabilidades dos mundos possíveis em que acção e previsão coincidem são muito superiores às probabilidades dos mundos em que acção e previsão divergem. Isto porque, ao contrário do que sucede com maior frequência, o passado não é aqui contrafactualmente independente do futuro. Assim, como já sabemos, a utilidade contrafactual esperada de C1, escolher a caixa opaca, é bem superior à utilidade contrafactual esperada de C2, escolher as duas caixas:

$$UctfE_r(C1) = pr_r(C1 \gg PC1) \times u(C1 \wedge PC1) + pr_r(C1 \gg PC2) \times u(C1 \wedge PC2) =$$
$$= 0.95 \times 1.000.000 + 0.05 \times 0 =$$
$$= 950.000$$

$$UctfE_r(C2) = pr_r(C2 \gg PC1) \times u(C2 \wedge PC1) + pr_r(C2 \wedge PC2) \times u(C2 \wedge PC2) =$$
$$= 0.05 \times 1.001.000 + 0.95 \times 1000 =$$
$$= 51.000$$

Até agora não se encontraram razões para desautorizar o diagnóstico inicial do problema como um confronto entre princípios de racionalidade: a dominação e a maximização da utilidade condicional esperada. Torna-se, portanto, necessário enquadrar, na problemática deste confronto, os resultados obtidos e mostrar como diversas soluções a resolvem de maneira diferente.

A teoria evidencial encontra-se comprometida com um tipo de dominação que se pode denominar ‘evidencial’ (dominação – *e*):

Seja [*E1*, *E2*, *E3*, …] uma partição de estados do mundo evidencialmente irrelevantes para a escolha do agente entre *A* e *B*. Se o agente preferir fracamente *A* a B, dado E_i e para cada E_i, então ele deve escolher fazer *A*.

Este é um princípio consideravelmente forte, pois para que se possa aplicar não é suficiente que os estados do mundo sejam causalmente independentes das acção. É também necessário que sejam evidencialmente independentes. Como tal, a dominação – *e* não se aplica ao PN.
A dominação que se aplica ao PN, e aos casos de causa comum em geral, pode ser denominada 'causal' (dominação – *c*):

Seja [*E1*, *E2*, *E3*, …] uma partição de estados do mundo causalmente irrelevantes para a escolha do agente entre *A* e *B*. Se o agente preferir fracamente *A* a *B*, dado E_i e para cada E_i, então ela deve escolher fazer *A*.

Este é o tipo de dominação que se encontra teoricamente comprometida com o princípio causal.
Ao definirmos o princípio da dominação destas duas maneiras, estamos a oferecer às teorias evidencial e causal os meios para tornarem os seus princípios compatíveis com o raciocínio a partir da dominação. Contudo, o problema estava inicialmente do lado da teoria evidencial. Já se sabia que não era possível aplicar o princípio da dominação a algumas situações em que os estados não eram probabilisticamente independentes das acções. A teoria evidencial surge, precisamente, para integrar esse elo probabilístico no cálculo da utilidade esperada. Se se for bicaixista, como era o caso de Nozick, a solução para o problema consistia em seguir a dominação em casos de conflito entre esta e o PMUCE. Se se fosse monocaixista, o princípio evidencial mantinha-se verdadeiro, embora o tipo de dominação compatível com este seja a dominação – *e*, a qual não se aplica ao PN.
Esta diferença entre tipos de dominação tem como fundamento a relação que existe entre correlação estatística/evidencial e causação. Esta última implica a primeira, mas não o inverso: pode haver correlação sem causação. Portanto, no que respeita aos estados do mundo:

causalmente relevantes → evidencialmente relevantes.

Daqui conclui-se que

dominação – *e* → dominação – *c*.

Ou seja, não é possível que os estados sejam, ao mesmo tempo, causalmente relevantes e evidencialmente irrelevantes. Ou seja, a dominação – *e* é um princípio mais forte do que a dominação – *c*.
Mas, se formos simultaneamente defensores do monocaixismo e da teoria causal, parece faltar-nos um tipo de dominação compatível com a interpretação retroactiva, algo que poderíamos designar como dominação 'retroactiva' (dominação – *r*):

Seja [*E1, E2, E3*, …] uma partição de estados do mundo retroactivamente irrelevantes para a escolha do agente entre *A* e *B*. Se o agente preferir fracamente *A* a *B*, dado E_i e para cada E_i, entre ela deve escolher fazer *A*.

Este é um princípio que não se aplica obviamente ao PN, pois, como vimos, os estados não são contrafactualmente irrelevantes, no sentido retroactivo, das acções (aplica-se, por exemplo, ao sonho do fumador). Mas este é um tipo de dominação que o bicaixista não aceitará, pois para ele é suficiente que os estados sejam contrafactualmente irrelevantes, no sentido standard, para que exista dominação de uma acção sobre a outra. Não é que o monocaixista defenda uma outra versão da teoria causal, para a qual tenha de encontrar um tipo de dominação compatível. O monocaixista defende apenas que, em certos casos, as contrafactuais utilizadas para o cálculo da utilidade esperada devem ser interpretadas de maneira diferente. Isto terá consequências no que respeita à determinação das probabilidades dessas contrafactuais, mas não quanto à escolha de teoria. Contudo, o monocaixista necessitará sempre de um tipo de dominação compatível com uma interpretação retroactiva dessas contrafactuais. Nessa medida, apesar de concordar com o bicaixista, em que a dominação – *c* se aplica ao PN, isso para ele não é suficiente.
A dominação – *r* é um princípio mais forte do que a dominação – *c*, pois não só exige que os estados sejam causalmente irrelevantes no sentido standard, mas que também o sejam no sentido retroactivo. Por outro lado, é mais fraco do que a dominação – *e*, pois não é

possível um estado ser retroactivamente relevante, sem que também o seja evidencialmente. Causação 'retroactiva' implica correlação, mas não o contrário. Disto segue-se que

$$\text{dominação} - e \rightarrow \text{dominação} - r \rightarrow \text{dominação} - c.$$

A dominação – *r* coincidirá, na esmagadora maioria dos casos, com a dominação – *c*, pois, normalmente, quando os estados são causalmente irrelevantes, são-no também retroactivamente. Contudo, casos como o PN, segundo o monocaixista, constituem excepções.

12.2. Um novo Problema de Newcomb

Consideremos, agora, três problemas de decisão distintos que, embora possam ser representados pela mesma tabela de decisão, exigem, de acordo com as suas diferentes resoluções, a aplicação de princípios da dominação diferentes. A estrutura geral do problema já é nossa conhecida (ver exemplo de Jeffrey em §6.1): José convida António para jantar e, como é de bom tom, espera-se que António forneça o vinho. Contudo, António não consegue contactar José e, portanto, não sabe se vai ser servida carne ou peixe. António prefere beber vinho tinto com carne a vinho branco com peixe, e vinho branco com carne a vinho tinto com peixe. A seguinte é a tabela de decisão do problema, incluindo as utilidades de António:

	Peixe	Carne
Comprar tinto	50	100
Comprar Branco	40	90

A única coisa que, à partida, se pode deduzir desta tabela é que o agente gosta muito mais de carne do que de peixe. Contudo, sem mais informação, fica por apurar o tipo de relação existente entre os estados e as acções. Suponhamos, em primeiro lugar, que não existe

qualquer maneira de António adivinhar o que irá ser servido, nem de conseguir, com as suas acções, influenciar José a servir peixe ou carne. Neste caso, os estados não só são causalmente independentes das acções, como também o são evidencialmente. Como tal, aplica-se a este caso a dominação – *e* e, *a fortiori*, a dominação – *r* e *c*. Como escolher tinto é sempre melhor, seja qual for o estado do mundo que se venha a verificar, a acção racional consiste em comprar tinto.

Suponhamos agora outra situação, na qual a acção do agente pode contribuir para que se verifique, com uma certa probabilidade, um ou outro dos estados do mundo. Nem sempre José decidiu já o que servir antes de António chegar, podendo o tipo de vinho comprado influenciar a sua decisão. Contudo, como José gosta mais de peixe do que de carne, a probabilidade de ele decidir servir peixe, dada a escolha de branco, é maior do que a probabilidade de ele servir carne, dada a escolha de tinto. Ou seja,

$pr(\mathrm{T} \gg \mathrm{C}) - pr(\mathrm{T} \gg \mathrm{P}) < pr(\mathrm{B} \gg \mathrm{P}) - pr(\mathrm{B} \gg \mathrm{C})$.

Neste caso, os estados do mundo não são apenas evidencialmente dependentes das acções, como também o são contrafactualmente. Ou seja, aplicam-se aqui tanto o princípio evidencial, como o princípio causal. Como neste caso tinto com peixe é sempre preferido a branco com peixe, tanto a utilidade condicional esperada de comprar tinto, como a utilidade contrafactual esperada de comprar tinto, excedem as respectivas utilidades de comprar branco. Ou seja, mesmo que $pr(\mathrm{B} \gg \mathrm{P}) = 1$, a utilidade de comprar branco nunca excede a de comprar tinto. Neste caso, como no anterior, não existe confronto entre dominação e maximização da utilidade esperada, apesar de não se verificar qualquer tipo de dominação.

Consideremos, finalmente, um terceiro caso: José tem uma habilidade especial para adivinhar que vinho vai António comprar; contudo, José é um brincalhão e gosta de fazer desfeitas ao amigo; portanto, se ele prevê que António vai comprar tinto, ele decide cozinhar peixe, e se ele prevê que António vai comprar branco, ele decide cozinhar carne. O que se pode constatar é que, neste caso, existe dominação – *c*, pois os estados são contrafactualmente independentes das acções. Faça António o que fizer, tal não alterará a decisão de José: a refeição já está pronta. Ou seja, a previsão de José foi feita antes da decisão tomada por António. Portanto, aplicando-se o princípio causal, obtém-se o mesmo resultado que raciocinando a partir da dominação: dado que $pr(\mathrm{T} \gg \mathrm{C}) = pr(\mathrm{T} \gg$

P) e *pr*(B ≫ P) = *pr*(B ≫ C), a utilidade causal esperada (no sentido standard) de comprar vinho tinto excederá sempre a de comprar branco.

Contudo, tal como no PN, os estados não são nem retroactivamente independentes das acções, nem evidencialmente independentes. Não só comprar tinto constitui evidência da decisão de servir peixe e comprar branco da decisão de servir carne, como também *pr*(T ≫ P) > *pr*(T ≫ C) e *pr*(B ≫ C) > *pr*(B ≫ P). Portanto, não existindo nem dominação – *e*, nem dominação – *r*, aplica-se tanto o princípio evidencial como o princípio retroactivo, os quais, neste caso, coincidem um com o outro no resultado. Supondo que *pr*(T ≫ P)/*pr*(P|T) ≥ 0,7; e que *pr*(T ≫ C)/*pr*(C|T) ≤ 0,3, temos as seguintes utilidades, evidencial e retroactiva:

U(T) = 0,7 × 50 + 0,3 × 100 = 65
U(B) = 0,3 × 40 + 0,7 × 100 = 82

Tal como no PN, quanto maior for a capacidade de José de prever que vinho irá António comprar, mais António se convence de que existem apenas duas consequências possíveis: beber tinto com peixe ou beber branco com carne. Como António prefere muito mais a segunda do que a primeira, a acção que é racional executar, de acordo com os argumentos evidencial e retroactivo, é a de comprar vinho branco.

Este problema é estruturalmente idêntico ao PN que já conhecemos. Mas será que este tipo de argumento nos parece agora mais persuasivo? Ou será que o bicaixista estará disposto a comer peixe, apesar de preferir muito mais comer carne. Pessoalmente, enquanto apreciador de carne – apesar de reconhecer a moral superior do pescetariano/vegetariano/vegan – acredito que não ganho nada comprando uma garrafa de tinto, apesar de a refeição já estar pronta quando tiver chegado à casa do meu amigo. Do mesmo modo, não vejo que ganhe seja o que for escolhendo as duas caixas e, por isso, prefiro escolher a caixa opaca. Afinal, se estamos dispostos a aceitar, em certas circunstâncias, a verdade de contrafactuais interpretadas de maneira retroactiva (ou, pelo menos, a adoptar uma crença na sua elevada probabilidade, de acordo com a semântica de Van Fraassen), por que não estaremos igualmente dispostos a agir de acordo com essa mesma interpretação? O que está em causa nos problemas de decisão em geral é a questão de saber de que modo as nossas acções contribuem para tornar verdadeiros os estados do mundo. Ora, não existindo qualquer razão de princípio para descartar a possibilidade de

esses estados poderem ser representados por contrafactuais ‘retroactivas’, por que não aceitar que as nossas acções, oferecendo o antecedente, contribuem para ‘tornar esses estados verdadeiros’?

Observações finais

O conceito de causalidade assumiu um papel preponderante na análise, explicação e tentativa de resolução dos problemas considerados na segunda e terceira parte deste trabalho. Mesmo que o objectivo central não tenha consistido directamente numa análise deste conceito, e que, por esse mesmo motivo, não se tenha decidido quanto à correcção definitiva de uma ou outra perspectiva acerca da metafísica e epistemologia da causação, o facto é que o apelo ao poder explicativo de tal conceito é essencial para a mais rudimentar compreensão dos problemas considerados e para a elaboração das teorias propostas para os resolver. Não só esteve em causa a consideração do poder causal das nossas acções, mas também o modo como a *desejabilidade* de certos bens ou estados de coisas, juntamente com as nossas crenças (nomeadamente, crenças acerca de poderes casuais), determinam as nossas escolhas e explicam a nossa maneira de agir. Não é, portanto, de estranhar, nem de todo inesperado, que as conclusões que apresentámos acerca do poder explicativo de versões behavioristas da teoria da decisão (ver §2) estejam de acordo com a ideia de que **a rejeição da noção de causalidade não nos permite oferecer explicações compatíveis com a hipótese de que os nossos desejos e crenças são causas das nossas acções.** A deficiência desta abordagem científica é confirmada de várias maneiras pela forma como a mera observação do comportamento dos indivíduos é insuficiente para se obterem conclusões adequadas acerca dos seus desejos e crenças.

A discussão dos fundamentos da teoria, na Parte 1, também não revelou conclusões que possamos considerar inesperadas (ver §3). **A transitividade revelou-se um princípio sólido**, ainda que o fenómeno da vagueza, relacionado com a formulação dos nossos juízos, possa 'contaminar' as nossas ordenações de preferências. Por outro lado, mesmo que se aceite a possibilidade dessa contaminação, tal **não é suficiente para salvar o axioma da completude**. Quanto ao axioma da independência, optou-se por não se discutir os problemas habitualmente considerados quando está em causa o ataque à validade normativa da teoria da utilidade esperada em geral: os 'paradoxos' de Allais (1953) e de Ellsberg (1961), casos em que se revelam preferências supostamente justificáveis, mas que contrariam o princípio da maximização da utilidade esperada. Não só existem várias e boas soluções há muito avançadas para lidar com estes desafios, como também a opção por se considerar um problema de escolha sequencial permitiu analisar a questão da inconsistência dinâmica em geral. Portanto, **se a escolha resoluta for uma**

exigência de racionalidade, tal como foi defendido, existem então boas razões para se rejeitar o axioma da independência. Convém também notar que a consideração do axioma independência não tem apenas interesse para os que estão interessados na teoria de Savage. Allan Gibbard (1984) demonstrou um teorema da representação para a teoria causal utilizando o teorema de Savage, suplementando os axiomas deste último com certas restrições sobre crenças em contrafactuais – de modo a garantir a independência contrafactual dos estados - permitindo, desse modo, seleccionar uma partição aceitável de estados do mundo em relação à qual é possível calcular utilidades causais. É claro que as desvantagens da teoria de Savage repercurtir-se-ão nos resultados obtidos por Gibbard. Uma das vantagens da teoria de Jeffrey, relativamente à de Savage, consiste no facto de a sua fórmula permitir obter resultados que não variam consoante a partição dos estados do mundo. Ou seja, a utilidade esperada das acções não varia consoante a partição adoptada para representar o problema de decisão em causa. Esta é uma característica altamente desejável de uma teoria da decisão, pois só se pode realmente afirmar que o princípio da maximização da utilidade esperada oferece uma solução para um problema de decisão, se a acção com maior utilidade não variar consoante a sua representação. Felizmente, existem vários resultados que demonstram que a fórmula do cálculo da utilidade causal esperada é independente da partição escolhida (Joyce 1999).

Já na Parte 2, em §5.2, ao analisar-se o argumento da dominação, considerou-se uma reformulação da partição dos estados do mundo no PN que dispensava a aplicação do princípio da dominação. Foi aí avançado um argumento monocaixista, admitidamente fraco, segundo o qual o único princípio de racionalidade que nos restava, eliminada do cenário a dominação, seria o princípio de Jeffrey, o que resultaria numa recomendação monocaixista. Nesse ponto não tínhamos ainda à nossa disposição o aparato da teoria causal. Agora que já o temos, podemos não só mostrar que a sua aplicação resulta numa recomendação bicaixista, mas também colocar em evidência que este resultado está de acordo com essa característica altamente desejável das teorias da decisão. Considere-se (C representa a correcção da previsão e ¬C a sua incorrecção, E1 representa a escolha da caixa opaca e E2 a escolha das duas caixas):

$$\begin{aligned} \text{UcausalE}\,(E1) &= pr\,(E1 \gg C) \times 1{,}000{,}000 + pr\,(E1 \gg \neg C) \times 0 = \\ &= pr\,(C) \times 1{,}000{,}000 \end{aligned}$$

$$\text{UcausalE (E2)} = pr\,(\text{E2} \gg \text{C}) \times 1{,}000{,}000 + pr\,(\text{E2} \gg \neg\text{C}) \times 1{,}001{,}000 =$$
$$= pr\,(\text{C}) \times 1{,}000{,}000 + pr\,(\sim\text{C}) \times 1{,}001{,}000$$

Ou seja, UcausalE (E2) > UcausalE (E1).

Na Parte 2 apresentou-se também uma reformulação da teoria de Savage, de acordo com uma sugestão de Gilboa/Jacinto (ver §6.2), e, dada essa reformulação, considerou-se que **a teoria causal pode ser interpretada como uma extensão natural da teoria de Savage.** Esta interpretação pode ser reforçada mediante a observação de um aspecto importante dos argumentos considerados, nomeadamente, o facto de Gilboa defender uma posição monocaixista. Ou seja, **no contexto da teoria de Savage dá-se a mesma cisão que, de acordo com o que foi visto, se verifica no seio da teoria causal entre as duas interpretações das contrafactuais, standard e retroactiva**. Daí surgir como natural a atribuição a Gilboa de uma interpretação retroactiva das contrafactuais envolvidas na reformulação dos estados do mundo na teoria de Savage. Ou seja, não me parece possível defender o monocaixismo nesse contexto, sem que se esteja, pelo menos implicitamente, a defender as duas seguintes atribuições de probabilidade: $pr\,(\text{C1} \gg \text{PC1}) = 1$ e $pr\,(\text{C2} \gg \text{PC2}) = 1$.

A conclusão respeitante à teoria de Savage é a de que esta é neutra quanto à solução do PN, estando essa solução dependente do modo como se interpretam as contrafactuais acima. Esta tensão, que é automaticamente transplantada para a teoria causal, é incómoda no que respeita à economia e valor normativo da teoria. Ainda que possamos ter confiança em como a teoria é adequada para dar conta do modo como deve funcionar a nossa racionalidade instrumental, ou de que temos o instrumento necessário para lidar com todos os problemas decisórios que podemos enfrentar, verificamos, afinal, que esse instrumento é de certa forma rombo, não possuindo a precisão necessária à operação que temos pela frente. Em suma, parece ser necessária uma extensão da teoria que nos permita destrinçar, em cada circunstância, que interpretação devemos aplicar. Horgan (1981), possivelmente sentindo esta tensão, defende que a interpretação retroactiva é pragmaticamente a mais adequada em todas as circunstâncias e que, desse modo, se deve aplicá-la a todos os problemas de decisão. **Quando confrontado com um típico PNCC, o Sonho do Fumador, as observações de Horgan podem ser interpretadas como um esboço da *tickle defense* de Eells**:

‘Resumindo, a falha do exemplo encontra-se no pressuposto de que o agente tem de *agir* antes de possuir a informação relevante para determinar a probabilidade de ter cancro do pulmão. Ele não tem de o fazer, pois as suas desejabilidades (e comportamento passado) oferecem-lhe já as melhores notícias. Este mesmo problema surge na maioria dos putativos contra-exemplos à maximização-*V* [utilidade condicional esperada], oferecida pelos bicaixistas revisionistas’ (Horgan 1981: 179).

Contudo, tendo em conta que a *tickle defense*, e a estratégia ratificacionista em particular, não são guias fiáveis da acção, resta-nos a opção definitiva pela teoria causal.

Em suma, mesmo no contexto da teoria causal, não temos ao nosso dispor um método adequado, baseado na estrutura dos casos em discussão, que nos permita discriminar entre os ‘casos standard’ e os ‘casos retroactivos’. Isto não implica, contudo, que não possamos ter algumas ideias esclarecidas acerca daquilo em que, de caso para caso, pode pesar em favor de uma ou outra interpretação. Por exemplo, parece não existir nada de estruturalmente distinto entre o PN e o Sonho do Fumador, o que poderá levar alguém a argumentar no seguinte sentido: se, no PN, a fiabilidade do previsor é o elemento circunstancial decisivo que nos leva a optar pela caixa opaca, obscurecendo o facto de não existir influência causal entre decisão e previsão, então, no Sonho do Fumador, a estreita correlação entre o desenvolvimento do cancro e a acção de fumar deveria obscurecer o facto de não existir influência causal desta para aquele, levando-nos a optar por deixar de fumar. Afinal, em ambos os casos, trata-se de maximizar o valor das notícias que a acção nos oferece.

Ora, este não me parece ser um argumento forte. Se no PN é correcto afirmar que a fiabilidade do previsor ‘determina’ fortemente o sucesso da nossa acção, levando-nos a obter a consequência mais desejada das duas possíveis (ignorando a possibilidade de ganhar 1,001,000), no Sonho do Fumador não se pode afirmar que a fortíssima correlação estatística (o equivalente à fiabilidade do previsor) determine de alguma maneira qualquer uma das consequências disponíveis: fumar com cancro e não fumar sem cancro. Neste caso, a ausência de causalidade entre fumar e desenvolver cancro não se encontra obscurecida por qualquer factor ‘determinador’ do sucesso da acção. Já no dilema do prisioneiro, tal como no PN, a identidade entre os dois jogadores encontra-se fortemente associada à segunda consequência mais desejável, enquanto factor que a ‘determina’,

obscurecendo o facto inegável de que não existe causalidade entre as acções dos dois prisioneiros. Além disso, existe um elemento claro de frustração associado à acção de não fumar, o qual não parece existir nos casos do monocaixismo e da cooperação: 1,000,000 é muito melhor do que 1000, e 1 ano de prisão é muito 'melhor' do que 5. Por outro lado, ainda que 'evitar' o cancro seja uma consequência altamente desejável, tal não eliminará o desejo de fumar.

Se isto nos soar como uma repetição do argumento pragmático favorável ao monocaixismo (ver §7.2), tal não deve constituir surpresa. Acredito que o máximo que podemos concluir, a este respeito, é que **a opção por uma ou outra interpretação deverá funcionar mais ou menos de acordo com o método do equilíbrio reflectido** (Rawls 1971): tanto nos domínios dedutivo e indutivo, como no domínio prático, devemos tentar encontrar um equilíbrio razoável entre os nossos juízos ponderados e os resultados das nossas teorias, tentando ajustar num todo coerente as nossas intuições mais fortes acerca do que é racional e as consequências da aplicação dos princípios, revendo por vezes os nossos juízos, mas também, quando necessário, fazendo ajustes nos detalhes das teorias.

Ainda na Parte 2, avaliaram-se as premissas dos argumentos mono e bicaixista de acordo com uma semântica de contrafactuais que atribuía a estas proposições um valor de verdade indeterminado, equivalente a um grau de crença no consequente, dada a verdade do antecedente. Ora, dado que teremos sempre de interpretar essas contrafactuais de uma forma causal, para efeitos do cálculo da utilidade esperada, talvez seja necessário procurar uma análise da causalidade que seja mais adequada a essa semântica, nomeadamente, uma análise probabilística (Suppes 1970) ou uma análise que recorra ao conceito de chance objectiva [*objective chance*].[89] Se se adoptar uma destas análise não é descabido afirmar que a formúla da utilidade esperada que lhe está naturalmente associada não é a de Gibbard e Harper, mas talvez a de Skyrms (1980) ou talvez a do próprio Lewis (1981). Considere-se, por exemplo, a interpretação de Skyrms: a partição *K* deve consistir numa especificação daqueles factores que se encontram fora da influência das nossas acções,

[89] Existe uma maneira de substituir *K* na fórmula da utilidade causal, interpretando as *dependency hypotheses* como especificações da 'probabilidade' condicional objectiva [*objective conditional chance*] de um determinado estado do mundo E, dada uma acção A ou uma acção alternativa B (há aqui uma dificuldade relacionada com a tradução de *chance*), em que as estimativas dos agentes quanto a essa 'probabilidade' correspondem a diferenças quanto às suas crenças nos poderes causais de A e B para produzir E (Lewis 1980, Skyrms 1988). De acordo com esta interpretação, deve-se ter em conta, por exemplo, que a 'probabilidade' objectiva de contrair cancro do pulmão é maior caso se fume do que no caso de não se fumar, e que este risco, ou o valor desta 'probabilidade', é independente daquilo que as pessoas pensam. Existe, claro, uma forma de reduzir esta *chance objectiva* ao grau de crença racional numa certa partição de eventos (de Finetti 1964).

mas que são causalmente relevantes para determinar as suas consequências. A ideia fundamental consiste em condicionar as crenças do agente - acerca da influência causal das suas acções - relativamente à partição *K*, a qual estabelece o pano-de-fundo ou a estrutura causal do mundo em que executamos as nossas acções. Portanto, um determinado factor *X* é uma causa mais eficiente de *C* do que um outro factor *Y*, apenas quando, tendo em conta a estrutura causal do mundo, a probabilidade condicional de *C* dado *X* é maior do que a probabilidade condicional de *C* dado *Y*. Representado *C* os factores que podem ser influenciados pela nossa acção, temos, assim, a fórmula de Skyrms:

$$\text{UcausalE (A)} = \sum_{ij}\ pr(Ki)\ pr(Cj\ |Ki \wedge A)\ u(Cj \wedge Ki \wedge A).$$

Esta fórmula oferece-nos os resultados correctos em casos como o Sonho do Fumador e uma recomendação bicaixista no PN. Contudo, existe sempre a necessidade de interpretar *C* e de determinar, no caso do PN, se o conteúdo da caixa opaca é um dos factores que pode ser influenciado causalmente pelas nossas acções. Um argumento monocaixista formulado no contexto da teoria de Skyrms seria um argumento que nos tentaria convencer de que a interpretação pragmaticamente adequada de *C* consiste em incluir nesta partição o conteúdo da caixa opaca. Portanto, um argumento deste tipo fará certamente eco do mesmo tipo de considerações avançadas pelo meta-argumento monocaixista da Parte 2. Em suma, para efeitos práticos, a teoria de Gibbard e Harper e a de Skyrms são inter-substituíveis.[90]

De qualquer modo, **existe um motivo pelo qual é preferível optar pela interpretação contrafactual das *dependency hypotheses.* Esse motivo está relacionado com o tipo de premissas utilizadas nos argumentos-objecto, ou seja, com o facto de estes se encontrarem formulados através de contrafactuais. Estas proposições constituem a base do raciocínio comum empregue pelos indivíduos em circunstâncias deliberativas e, tanto quanto consigo conceber, parece-me difícil encontrar outros e melhores argumentos/meta-argumentos para defender uma solução pré-teórica para o PN. Mais, a suposição contrafactual é o método de revisão de crenças que**

[90] A menos que se considere que a falsidade do pressuposto de Stalnaker implica a impossibilidade de se interpretar contrafactualmente as *dependency hypothesis*. Contudo, como se viu (§9.2, n. 67), existe uma interpretação da *visualização* que é capaz de ultrapassar esse problema (Joyce 1999; Gardenfors 1988: 108-18).

deve ser empregue quando está em causa a determinação da influência causal das nossas acções sobre os estados do mundo, tal como foi defendido na Parte 3.

Se a condicionalização é o método de revisão a ser empregue aquando da adição de uma nova crença à nossa visão do mundo, a *visualização* de mundos possíveis é o método a ser empregue quando se tem de ter em conta a alteração do mundo produzida pela nossa acção. Ambos os métodos consistem, a meu ver, em formas diferentes de aprendizagem; à primeira chamaríamos aprendizagem 'bayesiana' e à segunda aprendizagem 'contrafactual'.

Poderá haver quem negue este estatuto ao método da visualização, com base na ideia de que, ao contrário da condicionalização, regrada pela aplicação do Teorema de Bayes, a visualização é um procedimento demasiado vago e impreciso para produzir resultados aceitáveis. Contudo, como Lewis afirma, a revisão da nossa função de probabilidade não é gratuita: para um qualquer mundo *M* e um antecedente possível *A*, a probabilidade do mundo *M'* mais próximo em que *A* é verdadeira difere da probabilidade de *M* o mínimo possível, apenas o suficiente para permitir a verdade de *A*, e não mais do que isso. Verifica-se uma transferência de probabilidade dos mundos não-*A* para os mundos-*A*, mas sem que se crie ou destrua qualquer probabilidade, ou seja, a soma das probabilidades dos mundos-*A* mais próximos tem de continuar a ser 1. É neste sentido que a visualização constitui um método de revisão de crenças que oferece uma revisão *mínima* da nossa função de probabilidade.

Mas não estaremos afinal comprometidos com uma análise contrafactual do fenómeno da causação? Se tivermos as ambições de Lewis – que não temos - então diremos que sim. Contudo, **para efeitos do cálculo da probabilidade de contrafactuais, basta-nos defender que estamos comprometidos com uma análise em que a determinação da nossa função de probabilidade sobre mundos possíveis é determinada através da análise da similitude entre esses mundos**. Se se fala de uma 'interpretação causal' das contrafactuais, é porque a nossa percepção da causalidade é fundamental para avaliarmos os problemas em causa e para determinar a racionalidade das nossas acções. Assim sendo, podemos simplesmente falar, por um lado, de uma resolução standard dessa similitude, e de uma resolução retroactiva por outro. A definição dessa função de probabilidade serve um empreendimento bastante específico, no contexto da tomada de decisões, que consiste em atribuir probabilidades a contrafactuais e em calcular valores de utilidade, dada uma certa interpretação do princípio da utilidade esperada.

Por outro lado, **a semântica de Van Fraassen, aquela com que acabámos por nos comprometer** (ver §8.2) – e aquela que, de acordo com a nossa visão intuitiva das coisas, nos permite considerar simultaneamente verdadeiras as premissas 3 e 6 do argumento monocaixista original e validar a conclusão desse mesmo argumento – **também não nos compromete com qualquer análise específica do fenómeno da causalidade**. Esta semântica encontra-se na base de um empreendimento distinto: atribuir valores de verdade a proposições condicionais em geral, independentemente do modo como essas proposições serão interpretadas, evidencialmente ou causalmente, para efeitos do cálculo de utilidade. Se adoptássemos a versão da teoria causal de Skyrms, em vez da de Gibbard e Harper, tal não nos ofereceria motivo para rejeitar a semântica adoptada. No contexto do Sonho do Fumador, por exemplo, parece-me razoável, de acordo com as nossas intuições e com a nossa visão das relações evidenciais em causa, apresentar um elevado grau de crença na contrafactual 'Se eu fumasse, então desenvolveria cancro' (ainda que fumar não provoque cancro), e, no entanto, atribuir-lhe uma probabilidade baixa para efeitos do cálculo da utilidade da respectiva acção.

Em suma, **defendeu-se que as teorias causais, como a de Gibbard e Harper, são as únicas capazes de lidar de forma adequada com os problemas de decisão em que as relações de dependência evidencial, entre estados e acções, não espelham as relações de causalidade que se verificam entre esses mesmos estados e acções. Defendeu-se que a teoria evidencialista de Jeffrey não se encontra apta para resolver de forma adequada os PNCC, incluindo as suas versões revisionistas que incluem estratégias como a *tickle defense* e o ratificacionismo. Além disso, defendeu-se que certos contra-exemplos não se mostraram eficazes na tentativa de desacreditar a teoria causal, não obtendo sucesso na tentativa de mostrar que algumas recomendações desta teoria vão contra as nossas intuições mais fortes acerca de como devemos agir em determinadas circunstâncias.**

Quanto ao PN propriamente dito, defendeu-se que, mesmo no contexto da teoria causal, é possível verificarem-se conflitos acerca de qual é a solução adequada. Isto resulta da constatação de que, para efeitos do cálculo da probabilidade das contrafactuais envolvidas, é possível interpretar essas contrafactuais de duas maneiras incompatíveis: a maneira *standard* e a maneira retroactiva. Verificou-se também que a defesa do monocaixismo se encontra historicamente associada à opção pela interpretação retroactiva, enquanto interpretação universal das

contrafactuais - em qualquer problema de decisão - e a uma opção pela versão evidencialista da teoria. Contudo, defendeu-se que não tem de existir uma interpretação padrão, que essa interpretação depende de considerações de relevância pragmática, e que a defesa do monocaixismo – aqui empreendida - pode ser feita no contexto da própria teoria causal. Isto terá obviamente consequências no que respeita à maneira como encaramos o valor normativo da teoria, ainda que os casos que originam o conflito de interpretações – o PN e o dilema - representem uma fracção relativamente pequena da totalidade dos problemas de decisão que enfrentamos ao longo da vida.

Parte da força do argumento apresentado, favorável a uma solução monocaixista, teve como base a adopção de uma semântica de contrafactuais que nos permitiu aceitar a validade do argumento monocaixista original, sem que essa validade dependesse de qualquer interpretação controversa das proposições em causa. Mais, essa semântica permitiu-nos colocar em destaque, na estrutura do argumento original, considerações probabilísticas essenciais, relevantes para a determinação do curso de acção mais vantajoso. **A desvantagem deste procedimento consiste, todavia, na necessidade de se efectuar uma divisão do trabalho na avaliação das contrafactuais envolvidas: num primeiro passo determina-se o seu valor de verdade, num contexto que ainda não faz parte dos procedimentos técnicos da teoria, e depois, para efeitos do cálculo da probabilidade dessas mesmas proposições, opta-se por uma das interpretações disponíveis - causal ou não-causal - com base em considerações pragmáticas, das quais faz parte o modo e a facilidade com que, em primeiro lugar, determinámos a sua verdade**.

Mais recentemente, Arif Ahmed (1914, 1914b, 2018) apresentou um conjunto de argumentos a favor da teoria evidencial, os quais vieram dar novo ímpeto e interesse à discussão em redor do PN e acerca do confronto entre teorias da decisão. Entre esses argumentos podem encontrar-se problemas de decisão em que, contrariamente à maioria dos casos aqui analisados, é aparentemente a teoria evidencial, e não a causal, que oferece os resultados intuitivamente correctos. Um desses problemas (Ahmed 2014: Cap. 5), relacionado com apostas em eventos ocorridos no passado, envolve vários pressupostos relacionados com uma perspectiva determinista específica, designada como *determinismo suave*, encontrando-se, como tal, aberto a fortes objecções teóricas. Contudo, o outro problema em causa (Cap. 6) apresenta-nos um cenário plausível envolvendo fenómenos

estatísticos de natureza quântica que, embora algo rebuscado (à semelhança do PN), não envolve o mesmo tipo de pressupostos teóricos do caso anterior.

Ahmed (2014: 196-99) apresentou ainda um problema de decisão com o objectivo de minar a autoridade do argumento a partir da dominação. A característica inesperada deste novo problema consiste em verificar-se aí, supostamente, uma relação de dependência causal, sem existir, contudo, qualquer relação de dependência evidencial, contrariando, assim, a ideia de que pode existir correlação sem causação, mas não o inverso. Considere-se:

Somos um soldado medieval na véspera de uma grande batalha e temos a opção de gastar 30 florins numa armadura de excelente qualidade. Contudo, um previsor altamente fiável diz-nos que amanhã vamos ser mortalmente feridos. O que devemos fazer? Gastar os trinta florins numa noite de pândega e diversão, ou passar uma noite de sóbria inquietude contemplando a nossa esplêndida, embora praticamente inútil, armadura?

Dada uma preferência por uma noite de embriaguez, gastar os trinta florins em bebida domina a opção de comprar a armadura. Se morrermos, é preferível embriagarmo-nos, e se não morrermos, é ainda preferível embriagarmo-nos. Contudo, apesar da fiabilidade da previsão, supomos que amanhã as leis da Física continuarão em vigor e que lutar com a armadura aumentará as nossas hipóteses de sobrevivência. Portanto, na suposição de certos *payoffs* e de certas probabilidades, a teoria causal recomenda que se compre a armadura.

Estou em crer que, de futuro, a discussão acerca do PN e do confronto entre teorias da decisão passará pela discussão deste e dos outros exemplos oferecidos por Ahmed.

Apêndice 1

Simplificação do lado esquerdo da segunda desigualdade:

pr(PC1|C1).u(PC1∧C1) + (1 – pr(PC1|C1).u(PC2∧C1)) =
pr(PC1|C1).u(PC1∧C1) + u.(PC2∧C1) – pr(PC1|C1).u(PC2∧C1) =
pr(PC1|C1).(u(PC∧|C1) – u(PC2∧C1)) + u(PC2∧C1)).

Simplificação do lado direito da segunda desigualdade:

pr(PC1|C2).u(PC1∧C2) + (1 – pr(PC1|C2)).u(PC2∧C2) =
pr(PC1|C2).u(PC1∧C2) + u.(PC2∧C2) – pr(PC1|C2).u(PC2∧C2) =
pr(PC1|C2).(u(PC1∧C2) – u(PC2∧C2)) + u(PC2∧C2).

Logo, a desigualdade é verdadeira se, e somente se,

pr(PC1|C1).(u(PC∧C1) – u(PC2∧C1)) + u(PC2∧C1) > pr(PC1|C2).(u(PC1∧C2) – u(PC2∧C2)) + u(PC2∧C2)

se, e somente se,

pr(PC1|C1).(u(PC1∧C1) – u(PC2∧C1)) > pr(PC1|C2).(u(PC1∧C2) – u(PC2∧C2)) + u.(PC2∧C2) – u(PC2∧C1)

se, e somente se,

$$\frac{pr(PC1|C1).(\mathrm{u}(\mathrm{PC1}\wedge\mathrm{C1})-u(PC2\wedge C1)}{u(PC1\wedge C1)-u(\mathrm{PC2}\wedge\mathrm{C1})} > \frac{pr(PC1|C2).\big(u(\mathrm{PC1}\wedge\mathrm{C2})-u(PC2\wedge C2)\big)+u(PC2\wedge C2)-u(PC2\&C1)}{u(PC1\wedge C1)-u(PC2\wedge C1)}$$

se, e somente se,

$pr(\text{PC1}|\text{C1}) >$

$$\frac{pr(\text{PC1}|\text{C2}).(u(PC1 \wedge C2) - u(PC2 \wedge C2))}{u(PC1 \wedge C1) - u(PC2 \wedge C1)} + \frac{u(PC2 \wedge C2) - u(PC2 \wedge C1)}{u(PC1 \wedge C1) - u(PC2 \wedge C1)}$$

se, e somente se,

$$pr(\text{PC1}|\text{C1}) > pr(\text{PC1}|\text{C2}) \cdot \left(\frac{u(PC1 \wedge C2) - u(PC2 \wedge C2)}{u(PC1 \wedge C1) - u(PC2 \wedge C1)} \right) + \frac{u(PC2 \wedge C2) - u(PC2 \wedge C1)}{u(PC1 \wedge C1) - u(PC2 \wedge C1)}.$$

Apêndice 2

Assim como em relação a proposições incondicionais podemos ter diferentes graus de crença, o mesmo acontece relativamente a proposições do tipo condicional. Podemos ter a certeza de que B, se A – por exemplo, se tem quatro ângulos iguais é um rectângulo – podemos ter uma crença forte em como ele passará no exame, caso estude arduamente, ou podemos ter uma crença fraca em como ele passará no exame, caso estude arduamente. Como as expressões 'se A' ou 'caso A' parecem ser equivalentes a 'na condição de A', surge como natural a seguinte igualdade:

pr(B, se A) = *pr*(B|A) = *pr*(A∧B)/*pr*(A).

Como os graus de crença (*c*) podem ser representados por valores probabilísticos, temos a seguinte hipótese:

c (B, se A) = *c* (A∧B)/*c* (A).

A postulação desta hipótese constitui nada mais do que uma tentativa de encontrar condições de verdade para as proposições condicionais, condições essas que dependeriam das relações lógicas existentes entre os nossos graus de crença em 'B, se A', 'A∧B' e 'A'; ou seja, dada a probabilidade x que atribuímos a A ser verdadeira, tentamos determinar a probabilidade y de um mundo possível em que A∧B seja verdadeira. A ideia central consiste em tentar determinar um X, tal que a *c* (X) = *c* (B, se A). Ora, X tem de ser algo que seja objecto de crença, uma proposição. Uma candidata natural a ocupar o lugar de X é a condicional de Stalnaker, cujas condições de verdade são idênticas para condicionais indicativas e contrafactuais:

'Considere-se um mundo possível em que A é verdadeira, e que, no entanto, difere minimamente do mundo actual. "*S A então B*" *é verdadeira (falsa) apenas no caso de B ser verdadeira (falsa) nesse mundo possível*' (Stalnaker, 1968, pp. 33-4).

(Mesmo que a hipótese acima se aplique em primeiro lugar a condicionais indicativas, é possível estendê-la plausivelmente às contrafactuais: acreditar que uma contrafactual é

verdadeira significa acreditar que *seria* provável que B, quando A tinha uma probabilidade não-zero de ocorrer. Por exemplo, não só é elevada a probabilidade de eu adoecer, caso beba este leite estragado, como também, quando eu o deitar fora, *teria sido* elevada a probabilidade de adoecer, caso o tivesse bebido.)

Edgington (1995: 274) oferece uma ilustração da prova de Lewis. O objectivo da prova consiste em demonstrar a seguinte possibilidade: $p(X) = p(A \gg B) \neq p(B|A)$. Uma distribuição de probabilidades consiste numa atribuição de valores entre zero e um aos membros de uma partição. Por exemplo, {A∧B, A∧¬B, ¬A} é uma partição. Neste caso, cada elemento da partição que nos interessa - {A∧B, A∧¬B, ¬A∧B, ¬A∧¬B} – constituiu um mundo possível.

Temos, portanto, duas distribuições de probabilidades, *d*1 e *d*2. Ambas concordam entre si quando A é verdadeira, mas discordam quando A é falsa:

	A	B	A ≫ B	*d*1	*d*2
1.	V	V	V	0,4	0,4
2.	V	F	F	0,1	0,1
3.	F	(V∨F)	V	0,4	0,1
4.	F	(V∨F)	F	0,1	0,4

No mundo possível 1, em que A e B são verdadeiras, A ≫ B é verdadeira; ou seja, no mundo possível mais próximo do actual em que A é verdadeira e tudo o resto permanece igual, B também é verdadeira; por exemplo, se eu *tivesse* comido a maçã podre, *teria ficado* doente. Um raciocínio do mesmo tipo aplica-se ao mundo possível 2. Nos mundos possíveis 3 e 4, nos quais A é falsa, fica por saber se o mundo ¬A mais próximo é um mundo em que B é verdadeira ou é um mundo em que B é falsa. É certamente estranho atribuir os valores de verdade V e F à proposição A ≫ B, quando estamos a atribuir à *probabilidade da sua verdade* um valor máximo de 0,4. Contudo, o que importa é que,

para qualquer distribuição de probabilidades, a probabilidade da *verdade* de A ≫ B seja idêntica a *p*(B|A). Os valores à direita, abaixo de *d*1 e *d*2, cuja soma é igual a um, consistem na distribuição de probabilidade para os quatro mundos possíveis à esquerda. Por exemplo, *d*1 e *d*2 atribuem ambas uma probabilidade de 0,4 ao mundo possível em que tanto A como B são verdadeiras.

Consideremos, primeiro, os valores de *p*(B|A). Em ambas as distribuições, *p*(B|A) é idêntica; ou seja, *p*(B|A) = *p*(A∧B)/*p*(A); a *p*(A∧B) é a probabilidade do mundo possível em que A e B são verdadeiras, 0,4; já *p*(A) é a soma das probabilidade dos mundos possíveis em que A é verdadeira, 0,5; logo, *p*(B|A) = 0,4/0,5 = 0,8. Consideremos, agora, a *p*(A ≫ B) em *d*1; esta é idêntica à soma das probabilidades dos mundos possíveis em que A ≫ B é verdadeira em *d*1; logo p_{d1} (A ≫ B) = 0,4 + 0,4 = 0,8. Até aqui tudo bem, ou seja, p_{d1} (A ≫ B) = *p*(B|A). Mas consideremos a *p*(A ≫ B) em *d*2; esta continua a ser idêntica à soma das probabilidades dos mundos possíveis em que A ≫ B é verdadeira em *d*2; mas como *d*2 é diferente de *d*1 quando A é falsa, a p_{d2} (A ≫ B) = 0,4 + 0,1 = 0,5. Ou seja, p_{d2} (A ≫ B) ≠ *p*(B|A). Logo, fica demonstrado que nem sempre é verdade que *pr*(A ≫ B) = *p*(B|A).

Referências

Ahmed, Arif (2014) *Evidence, Decision and Causality*. Cambridge University Press.

_____ (2014b) Dicing with death. *Analysis* 74: 587-92.

_____ (2018) Self-Control and hyperbolic discounting. In *Self-Control & Rationality: Interdisciplinary Essays*, (ed.) J. L. Bermúdez. Cambridge University Press, 2018.

_____ (ed.) (2018b) *Newcomb's Problem*. Cambridge University Press.

Adams, Ernest W. (1975) *The Logic of Conditionals*. Dordrecht: D. Reidel Publishing Company.

Aldred, Jonathan (2007) Intransivity and Vague Preferences. *The Journal of Ethics* 11: 377-403.

Allais, M. (1953) Le comportement de l'homme rationnel devant le risque: critique des postulats et axioms de l'école americaine. *Econometrica* 21: 503-46.

Arnauld, Antoine and Pierre Nicole [1662]/(1996) *Logic or the Art of Thinking*, 5th ed., translated and edited by Jill Vance Buroker. Cambridge: Cambridge University Press.

Arrow, K. J. (1970) The Theory of Risk Aversion. In *Essays in the Theory of Risk-Bearing*. North-Holland Publishing Company.

Axelrod, R. [1984]/(1990) *The Evolution of Co-Operation*. London: Penguin Books.

Bar-Hillel, Maya and Margalit, Avishai (1972) Newcomb's Paradox Revisited. *The Bristish Journal for the Philosophy of Science*, Vol. 23, Nº. 4: 295-304.

Bernoulli, Daniel [1738]/(1954) Expositions of a new theory on the measurement of risk. *Econometrica* 22: 23-36.

Bolker, Ethan D. (1966) Functions Resembling Quotients of Measures. *Transactions of the American Mathematical Society* 124: 292-312.

Broome, John (1989) An Economic Newcomb Problem. *Analysis*, 1989, Vol. 49: 220-22.

____ (2004) Is incommensurability vagueness? In *Incommensurability, incomparability, and practical reason*, (ed.) R. Chang, pp. 67-89. Harvard University Press.

Cantwell, John (2010) On an alleged counterexample to causal decision theory. *Synthese* 173: 127-52.

Carlson, Erick (2004) Broome's argument against value incomparability. *Utilitas* 16: 220-24.

Carnap, Rudolf (1953) Testability and Meaning. In *Readings in the Philosophy of Science*, (eds.) Feigl and Brodbeck, pp. 47-92. New York: Appleton-Century-Crofts

Clark, Michael (2002) *Paradoxes from A to Z*. Routledge.

Davidson, Donald (1974) Actions, Reasons and Causes. In his *Essays on Actions and Events*, pp. 3-19. Oxford: Clarendon Press.

Davidson, D., Suppes and P., Siegel S. (1957) Decision-Making: An experimental Approach. In *Decision-Making: Selected Readings*, (eds.) Edwards e Tversky, pp. 170-208. Harmondsworth: Penguin Books.

Davis, Lawrence H. [1977]/(1985) "Prisoners, Paradox, and Rationality" in *Paradoxes of Rationality and Cooperation, Prisoner's Dilemma and Newcomb's Problem*, (eds.) Richmond Campbell and Lanning Sowden, pp. 45-59. Vancouver: The University of British Columbia Press.

Davis, Morton D. (1997) *Game Theory, A Nontechnical Introduction*. Mineola, New York: Dover Publications.

de Finetti, Bruno (1964) Foresight: Its Logical Laws, Its Subjective Sources. In *Studies in Subjective Probability*, (eds.) H. Kyburg and H. Smokler, pp. 93-158. New York: John Wiley.

Edgington, Dorothy (1995) On Conditionals. *Mind*, Vol. 104. 414: 235-329.

Eells, Ellery (1984a) Newcomb's Many Solutions. *Theory and Decision* 16: 59-105.

____ (1984b) Metatickles and the Dynamics of Deliberation. *Theory and Decision*, 17: 71-95.

____ (1984c) Causal Decision Theory. *Proceedings of the Biennial Meeting of the Philosophy of Science Association*, Vol. Two: Symposia and Invited Papers, pp. 177-200. University of Chicago Press.

____ (1985) Causality, Decision, and Newcomb's Paradox. In *Paradoxes of Rationality and Cooperation, Prisoner's Dilemma and Newcomb's Problem*, (eds.) Richmond Campbell and Lanning Sowden, pp. 183-213. Vancouver: The University of Bristish Columbia Press.

Egan, Andy (2007) Some counterexamples to causal decision theory. *Philosophical Review* 116 (1): 93-114.

Ellsberg, D. (1961) Risk, ambiguity, and the Savage axioms. *Quarterly Journal of Economics* 75: 643-69.

Espinoza, Nicolas (2008) The small improvement argument. *Synthese* 165: 127-39.

Fine, Kit (1975) Vagueness, truth and logic. *Synthese* 30: 265-300.

Gardenfors, Peter (1988) *Knowledge in Flux*. Cambridge: MIT Press.

Gibbard, Allan (1984) Decision Matrices and Instrumental Expected Utility. In *Second International Congress on the Foundations of Utility*, Venice, Italy, June.

Gibbard, Allan and Harper, William H. [1978]/(1980) "Counterfactuals and Two Kinds of Expected Utility" in *Ifs*, (eds.) W. L. Harper, R. Stalnaker, and G. Pearce, pp. 153-90. D. Reidel Publishing Company.

Gilboa, Itzhak (2009) *Theory of Decision under Uncertainty*. Cambridge: Cambridge University Press.

Goodman, Nelson (1947) The Problem of Counterfactual Conditionals. *The Journal of Philosophy* 44: 113-28.

Gustafsson, Johan E. (2010) A Money-Pump for Acyclic Intransitive Preferences. *Dialectica* 64 (2): 251-57.

Hempel, C. G. (1965) Aspects of Scientific Explanation. In his *Aspects of Scientific Explanation and other Essays in the Philosophy of Science*, pp. 331-496: New York: The Free Press.

Hofstadter, D. R. (1983) Metamagical Themas: The calculus of cooperation is tested through a lottery. *Scientific American, Inc.*, September.

Holton, Richard (2016) Addiction, Self-Signaling and the Deep Self. *Mind and Language*, 31 (3): 300-313.

Horgan, Terence [1981]/(1985) Counterfactuals and Newcomb's Problem. In *Paradoxes of Rationality and Cooperation, Prisoner's Dilemma and Newcomb's Problem*, (eds.) Campbell and Lanning Sowden, pp. 159-82. Vancouver: The University of British Columbia Press.

___ (1985) Newcomb's Problem: A stalemate. In *Paradoxes of Rationality and Cooperation, Prisoner's Dilemma and Newcomb's Problem*, (eds.) Richmond Campbell and Lanning Sowden, pp. 223-34. Vancouver: The University of British Columbia Press.

Jacinto, Bruno, Newcomb's problem, Savage´s decision theory and ratificationism, manuscrito não-publicado, produzido em 2011 para o Projecto Paradoxos (PTDC/FIL/67039/2006).

Jackson, Frank and Pargetter, Robert (1983) Where the Tickle Defense Goes Wrong. *Australasian Journal of Philosophy*, 61: 295-99.

Jeffrey, Richard [1964]/(1983) *The Logic of Decision*, 2nd ed. Chicago: University of Chicago Press.

____ (1991) Matter of Fact Conditionals. *Proceedings of the Aristotelian Society*, 81: 125-37.

Joyce, James M. (1999) *The Foundations of Causal Decision Theory*. New York: Cambridge University Press.

____ (2012) Regret and instability in causal decision theory. *Synthese* 187: 123-45.

Krantz, D. T., Luce, R. D., Suppes, P., Tversky, A. (1971) *Foundations of Measurement*, Vol. 1. Academic Press Inc.

Kreps, David (1988) *Notes on the Theory of Choice*. Boulder: Westview Press.

Kripke, Saul (1963) Semantical Considerations on Modal Logic. *Acta Philosophica Fennica*, 16: 83-94.

Lewis, David (1973a) *Counterfactuals*. Oxford: Basil Blackwell.

____ (1973b) Causation. *The Journal of Philosophy*, Vol. 70, Nº. 17: 556-67.

____ (1976) Probabilities of Conditionals and Conditional Probabilities. *Philosophical Review* 85: 297-315.

____ (1979a) Counterfactual Dependence and Time's Arrow. *Noûs*, Vol. 13, Nº. 4: 455-76.

____ (1979b) Prisoners' Dilemma is a Newcomb Problem. *Philosophy and Public Affairs,* Vol. 8, Nº. 3: 235-40.

____ (1980) A Subjectivist's Guide to Objective Chance. In *Studies in Inductive Logic and Probability* Vol. 2, (ed.) R. Jeffrey, pp. 263-94. Berkeley: University of California Press.

____ (1981) Causal Decision Theory. *Australasian Journal of Philosophy*, 59: 5-30.

Levi, Isaac (1975) Newcomb's Many Problems. *Theory and Decision* 6: 161-75.

Luce, Duncan R. and Raiffa, Howard (1957) *Games and Decisions*. New York: John Wiley and Sons.

Maher, Patrick (1993) *Betting on Theories*. Cambridge: Cambridge University Press.

McClennen, E. (1990) *Rationality and Dynamic Choice*. Cambridge: Cambridge University Press.

Mill, J. S. [1843]/(1911) *A System of Logic*: *Ratiocinative and Inductive*. London: Longmans Green.

Nozick, Robert [1969]/(1985) Newcomb's Problem and Two Principles of Choice. In *Paradoxes of Rationality and Cooperation*: *Prisoner's Dilemma and Newcomb's Problem*, (eds.) Richmond Campbell and Lanning Sowden, pp. 107-33. Vancouver: The University of Bristish Columbia Press.

____ (1993) *The Nature of Rationality*. Princeton: Princeton University Press.

Peterson, Martin (2009) *An Introduction to Decision Theory*. New York: Cambridge University Press.

Pettit, P. and Sugden, R. (1989) The Backwards Induction Paradox. *The Journal of Philosophy* 86: 169-82.

Price, Huw (2012) Causation, chance, and the rational significance of supernatural evidence. *Philosophical Review* 121: 483-538.

Rabinowicz, Wlodek (1995) To Have One's Cake and Eat It Too: Sequential Choice and Expected-Utility Violations. *The Journal of Philosophy* Vol. 92, Nº. 11: 586-620.

____ (2000) Money Pump with Foresight. In *Imperceptible Harms and Benefits*, (ed.) M. J. Almeida, pp. 123-54. Kluwer Academic Publishers.

Ramsey, Frank (1931) Truth and Probability. In *The Foundations of Mathematics and Other Logical Essays*, (ed.) R. Braithwaite, pp. 156-98. London: Kegan Paul.

____ (1931b) General Propositions and Causality. In *The Foundations of Mathematics and Other Logical Essays*, (ed.) R. Braithwaite, pp. 237-56. London: Kegan Paul.

Raz, Joseph (1986) *The Morality of Freedom*. Oxford: Oxford University Press.

Raws, John (1971) *A Theory of Justice*. Oxford: Oxford University Press.

Resnik, Michael (1987) *Choices: An Introduction to Decision Theory*. Minneapolis: University of Minnesota Press.

Ryle, Gilbert (1949) *The Concept of Mind*. London: Hutchinson.

Samuelson, Paul (1938) A note on the pure theory of consumer's behavior. *Economica* 5: 61-71.

Savage, Leonard J. [1954]/(1972) *The Foundations of Statistics*, 2nd revised edition. New York, Dover.

Skinner, B. F. (1953) *Science and Human Behavior*. New York: Macmillan.

Skyrms, Brian (1980) *Causal Necessity*. New Haven CT: Yale University Press.

____ (1988) Conditional Chance. In *Probability and Causality*, (ed.) J. Fetzer, pp. 161-78. Dordrecht: Kluwe.

Sobel, John Howard (1985) Not Every Prisoners' Dilemma is a Newcomb Problem. In *Paradoxes of Rationality and Cooperation, Prisoner's Dilemma and Newcomb's Problem,* (eds.) Richmond Campbell and Lanning Sowden, pp. 263-74. Vancouver: The University of British Columbia Press.

Stalnaker, Robert C. (1968) A Theory of Conditionals. In *Studies in Logical Theory,* (ed.) N. Rescher, *American Philosophical Quarterly, Monograph Series* 2.

____ (1980) Letter to David Lewis, May 21, 1972. In *Ifs*, (eds.) W. L. Harper, R. Stalnaker, and G. Pearce, pp. 151-52. D. Reidel Publishing Company.

____ (1984) *Inquiry*. Cambridge MA: MIT Press.

Stalnaker R., and Jeffrey, R. (1994) Conditionals as Random Variables. In *Probability and Conditionals*: *Belief Revision and Rational Decision,* (eds.) E. Eells e B. Skyrms, pp. 31-46. Cambridge University Press.

Steele, Katie and H. Orri Stefánsson, "Decision Theory", in E. Zalta, *The Stanford Encyclopedia of Philosophy* (*Winter 2016 Edition*), URL =

<http://plato.stanford.edu/archives/win2016/entries/decision theory/>

Suppes, Patrick (1970) *A Probabilistic Theory of Causality*. Amsterdam: North Holland.

Teller, P. (1976) Conditionalization, observation, and change of preferences. In *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science,* Vol. 1, (eds.) W. Harper and C. Hooker, pp. 205-54. D. Reidel Publishing Company.

Temkin, L. S. (2012) *Rethinking the Good*. Oxford University Press.

Tichy, Pavel (1976) A Counterexample to the Stalnaker-Lewis Analysis of Counterfactuals. *Philosophical Studies* 22: 271-3.

Tversky, Amos (1969) Intransitivity of Preferences. *Psychological Review,* 1969, Vol. 76, Nº. 1: 31-48.

Van Fraassen, Bas (1976) Probabilities of Conditionals. In *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science,* Vol. 1, (eds.) W. Harper and C. Hooker, pp. 261-308. D. Reidel Publishing Company.

von Neumann, John and Otto Morgenstern (1953) *Theory of Games and Economic Behavior*, 3rd ed. Princeton: Princeton University Press.

von Mises, Richard [1936]/1957 *Probability, Statistics, and Truth*. London: Allen & Unwin.

Weirich, Paul, "Causal Decision Theory" in E. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2016 Edition), URL =

<https://plato.stanford.edu/archives/win2016/entries/decision-causal/>

Williamson, Timothy (1994) *Vagueness*. London: Routledge.

Zilhão, António (2001) Teoria da decisão. In *Enciclopédia de Termos Lógico-Filosóficos*, (eds.) João Branquinho e Desidério Murcho, pp. 686-92. Lisboa: Gradiva.

____ (2007) Hempel e a Explicação da Acção. In *Do Círculo de Viena à Filosofia Analítica Contemporânea*, (coord.) António Zilhão, pp. 159-88. Livros de Areia.

____ (2010)a *Pensar com Risco, 25 Lições de Lógica Indutiva*. Lisboa: Imprensa Nacional-Casa da Moeda.

____ (2010)b *Animal Racional ou Bípede Implume*? Lisboa: Guerra & Paz.

____ (2013) Filosofia da Ciência. In *Filosofia, Uma Introdução por Disciplinas*, (ed.) Pedro Galvão, pp. 249-81. Lisboa: Edições 70.