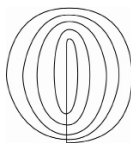


PROBLEMA DE NEWCOMB

EDIÇÃO DE 2020 do

COMPÊNDIO EM LINHA DE PROBLEMAS DE FILOSOFIA ANALÍTICA

2018-2021 FCT Project PTDC/FER-FIL/28442/2017



Editado por
Ricardo Santos e Pedro Galvão

ISBN: 978-989-8553-22-5

Compêndio em Linha de Problemas de Filosofia Analítica
Copyright © 2020 do editor
Centro de Filosofia da Universidade de Lisboa
Alameda da Universidade, Campo Grande, 1600-214 Lisboa

Problema de Newcomb
Copyright © 2020 do autor
Diogo Fernandes

DOI: <https://doi.org/10.51427/cfi.2021.0068>

Todos os direitos reservados

Resumo

O Problema de Newcomb é um *puzzle* decisório que requer a aplicação dos nossos princípios de racionalidade instrumental, tendo a sua discussão dado origem ao debate contemporâneo acerca da interpretação adequada do princípio da maximização da utilidade esperada. Como tal, na tentativa de encontrar uma solução para o mesmo, será necessário discutir as teorias em confronto que propõem interpretações distintas desse princípio. Após a apresentação do problema como um confronto entre o princípio da dominação e a teoria evidencial da decisão, apresentar-se-á a solução proposta pelos defensores da teoria causal. De seguida, apresentar-se-ão duas revisões da teoria evidencial – a *tickle defense* e a estratégia ratificacionista – concebidas com o objectivo de torná-la compatível com a solução oferecida pela teoria causal. Finalmente, apresentar-se-ão contra-exemplos à teoria causal e discutir-se-á a plausibilidade e sucesso dos mesmos.

Palavras-chave

Teoria evidencial, teoria causal, probabilidade condicional, dominação, *tickle*, ratificacionismo.

Abstract

Newcomb's Problem is a decision-theoretical puzzle that appeals to our notions of instrumental rationality. Its appearance in the literature gave rise to the confrontation between different interpretations of the expected utility maximization principle. As such, in trying to find a solution, it will be necessary to present the theories advocating those contrasting interpretations. After characterizing the problem as a confrontation between the dominance principle and evidential decision theory, I introduce the solution advocated by proponents of causal decision theory. I then present two revisionist strategies – the tickle defense and the ratificationist strategy – conceived with the purpose of making evidential decision theory compatible with the solution given by the causal theory. Finally, I present some counterexamples to the causal theory and discuss their plausibility and success.

Keywords

Evidential theory, causal theory, conditional probability, dominance, tickle, ratificationism.

Problema de Newcomb

DOI: <https://doi.org/10.51427/cfi.2021.0068>

1 Confronto entre princípios de decisão

O Problema de Newcomb (PN) é frequentemente incluído nas antologias de paradoxos (Clark 2002), normalmente agrupado em capítulos dedicados à racionalidade da acção (Sainsbury 1987). 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. No caso do PN, o que temos são duas recomendações distintas acerca do que fazer numa dada situação de escolha, em que temos de optar por executar uma de entre duas acções incompatíveis. O problema surgiu pela primeira vez em forma impressa num artigo de Robert Nozick (1969) e, portanto, a formulação que se irá adoptar é, no seu essencial, equivalente à que foi aí apresentada:¹

Temos perante nós duas caixas: uma delas é transparente e contém 1,000 Euros; a outra é opaca e contém 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 a 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 a nossa preferência, sendo que esta consiste em adquirir a maior quantia possível?

¹ 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.

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 *monocaixista* ao argumento favorável à escolha da caixa opaca; e *bicaixista* ao argumento favorável à escolha das duas caixas.

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. Essa formulação pode ser encontrada, no seu essencial, em Terence Horgan (1981):

Argumento monocaixista:

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

Argumento bicaixista:

- 1'. A caixa opaca contém 1 milhão ou está vazia.
- 2'. Se contiver 1 milhão, eu receberia 1 milhão e mil se escolhesse as duas caixas.
- 3'. Se contiver 1 milhão, eu receberia 1 milhão se escolhesse apenas a caixa opaca.
- 4'. Se estiver vazia, eu receberia 1000 se escolhesse as duas caixas.
- 5'. Se estiver vazia, eu não receberia nada se escolhesse apenas a caixa opaca.
- 6'. Das duas uma: receberia 1 milhão e mil 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, devo escolher as duas caixas.

Aceitando a verdade das premissas de ambos os argumentos, ambas as conclusões normativas, em 7 e 7', surgem-nos como perfeitamente plausíveis.

Desde Nozick, a abordagem tradicional ao 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. De acordo com esta teoria, os agentes agem racionalmente se, e somente se, agirem de forma a maximizar a utilidade condicional esperada das acções; seleccionando, portanto, de entre as acções disponíveis, aquela cuja utilidade condicional esperada é a mais elevada.

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, sendo que a acção *a* domina a acção *b* se, e somente se, 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*. Assim, na matriz abaixo a acção de escolher as duas caixas domina a acção de escolher a caixa opaca:

	Previsão duas caixas	Previsão caixa opaca
Escolher duas caixas	1,000	1,000 + 1,000,000
Escolher caixa opaca	0	1,000,000

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 aquele que é adoptado pela teoria da decisão de Richard 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 não são probabilisticamente independentes das acções, 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:

$$\text{UCE}(a) = \sum_{(i=1)}^n pr(e_i \mid a) \times u(a \wedge e_i)$$

em que pr é uma função de probabilidade subjectiva, definida numa álgebra Ω de proposições, incluindo C , A e E , contável e fechada sob negação e disjunção; $pr(x|y)$ é a probabilidade condicional de x dado y ; e u uma função de utilidade subjectiva definida em C .²

² Os tratamentos clássicos do conceito de probabilidade subjectiva são Frank Ramsey (1931) e Bruno de Finetti (1964). 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

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,

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

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

Supondo que $pr(\text{PC2}|\text{C2}) = pr(\text{PC1}|\text{C1})$, segue-se que a $\text{UCE (C1)} > \text{UCE (C2)}$ desde que o grau de fiabilidade do previsor seja superior a 0.5005. Ou seja, $pr(\text{PC2}|\text{C2}) = pr(\text{PC1}|\text{C1}) > 0.5005$. Como sabemos que o previsor é extremamente fiável, segue-se que a $\text{UCE (C1)} > \text{UCE (C2)}$.

Existem infindáveis exemplos de problemas de decisão em que é recomendável colocar de lado o princípio da dominação. Trata-se de casos em que, tal como no PN, os estados do mundo não são probabilisticamente independentes das acções. Um exemplo claro é oferecido por 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

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 .

de hostilidade permanente ($50 > 0$). Contudo, um defensor da manutenção de um arsenal nuclear para efeitos de dissuasão 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 PN? A resposta de Nozick é aquela que tem vindo a ser defendida pelos defensores do bicaixismo: embora, no PN, 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, ou seja, a escolha da acção não explica (nem causa) a previsão efectuada. 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 favor do princípio da dominação.³

³ Esta caracterização do problema, embora amplamente reconhecida, tem sido contestada por alguns autores. Por exemplo, Horwich (1987) argumentou que, pelo facto de alguém jamais ter enfrentado algo remotamente parecido com o PN, nenhuma intuição acerca do mesmo se reveste de especial relevância. Levi (1975) tentou mostrar que a descrição do problema não é suficiente detalhada para podermos determinar as probabilidades a aplicar e que, por esse motivo, deveríamos optar pelo princípio *maximin*, o que resultaria numa recomendação de bicaixismo. Para uma crítica a Levi, ver Eells (1984). Bar Hillel e Margalit (1972) apresentaram uma partição de estados do mundo para o PN em que, contrariamente àquela que foi aqui apresentada, os estados do mundo não só são probabilisticamente independentes das acções, mas em que também não se verifica dominação, permitindo assim uma aplicação não-problemática da teoria de Jeffrey, favorável ao monocaixismo. O desenvolvimento em §2 de uma teoria da decisão alternativa à de Jeffrey permitirá facilmente acomodar esta última estratégia.

2 Teoria causal da decisão

O principal interesse teórico do PN reside no facto de ter estado na origem dos desenvolvimentos da teoria da decisão posteriores ao seu surgimento. Daí que o foco deste artigo esteja direccionado para esses desenvolvimentos e para a disputa entre as interpretações da teoria da decisão desenvolvidas para lidar adequadamente com o problema, ou, mais genericamente, com aqueles problemas de decisão que têm a mesma estrutura do PN. O primeiro passo consistirá na apresentação das interpretações em disputa.

A consideração em §1 do exemplo do desarmamento nuclear permitiu-nos constatar que as deliberações devem ter em conta a influência das acções nas probabilidades dos estados do mundo que determinam as possíveis consequências dessas acções. A questão que se coloca é a seguinte: de que modo deve a teoria da decisão interpretar a noção de probabilidade envolvida nas proposições condicionais utilizadas para calcular a utilidade esperada das acções? O cálculo de probabilidades, ao dar conta da noção de probabilidade condicional, oferece uma sugestão:

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

A teoria de Jeffrey (1964) adopta esta concepção de probabilidade para interpretar a probabilidade de 'E, se A'.⁴ A segunda questão que se coloca é a seguinte: será esta interpretação adequada para resolver todos os problemas de decisão em que estejam envolvidas proposições do tipo condicional acerca da probabilidade de ocorrência de estados do mundo dada a execução de uma ou outra acção? A análise do PN pode ajudar-nos a encontrar uma resposta.

Pode-se 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.

⁴A fórmula que usa probabilidades condicionais consiste, no modelo de Jeffrey, numa generalização de uma fórmula mais simples para a utilidade esperada com probabilidades incondicionais (Savage 1954). Para uma introdução aos dois modelos, ver Katie Steele e H. Orri Stefánsson (2016).

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. É certo que, por vezes, estes dois aspectos coincidem e, nesse caso, a teoria de Jeffrey pode ser aplicada sem restrições.

Robert Stalnaker (1980), numa carta 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. Tendo em conta que a noção de probabilidade condicional não permite discriminar entre os casos em que se verifica uma mera correlação estatística entre uma acção e um estado do mundo, e os casos em que, para além dessa correlação, se verifica uma causação genuína, Stalnaker sugere que, 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' — $pr(A \Box \rightarrow E)$ — a qual tem a forma lógica de uma condicional contrafactual.

As condicionais contrafactuais apresentam, de acordo com Stalnaker (1968), condições de verdade diferentes das condições de verdade das condicionais indicativas. Considere-se a seguinte definição dessas condições, de Allan Gibbard e William Harper (1978), aplicada já ao caso específico da acção: seja A uma acção que eu posso decidir executar ou não em t ; seja o mundo-A um mundo possível idêntico ao actual antes de t , no qual eu decido executar 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 $A \Box \rightarrow E$ é verdadeira em $M(A)$. Encontra-se, assim, estabelecida a nossa semântica-padrão para as contrafactuais utilizadas nos argumentos do PN.

Mas em que consiste, então, calcular a probabilidade de $A \Box \rightarrow E$, de modo a que isso seja diferente de calcular a probabilidade condicional $E|A$? Não será a seguinte identidade uma verdade lógica:

$$(1) \quad pr(A \Box \rightarrow E) = pr(E|A)?$$

Ao contrário do que a nossa intuição nos pode levar a crer, Lewis (1976) demonstrou que, sob pressupostos bastantes plausíveis, (1) não é verdadeira acerca de quaisquer duas proposições arbitrárias. Contudo, dada uma certa condição, pode ser demonstrado que (1) é verdadeira. Importa, portanto, compreender a natureza dessa mesma condição e determinar o que acontece quando a mesma 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. Identifique-se, então,

Condição X: para uma acção A e uma consequência C, a contrafactual $A \Box \rightarrow C$ é epistemicamente independente da acção A. Ou seja,

$$pr(A \Box \rightarrow C | A) = pr(A \Box \rightarrow C).$$

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. De que nos serve, então, a verdade de (1)? 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,

$$UCE(A) = \sum_j pr(E_j | A) \times u(A \wedge E_j)$$

e a outra utiliza a probabilidade da contrafactual. Esta última 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:

$$UCaE(A) = \sum_j pr(E_j \Box \rightarrow A) \times u(A \wedge E_j)$$

Podemos, agora, responder à questão. Quando a condição X se verifica, ou quando (1) é verdadeira, a $UCE(A) = UCaE(A)$. Ou seja, se a probabilidade de uma contrafactual $A \Box \rightarrow E$ 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. O passo seguinte

consiste, portanto, em encontrar uma situação em que esta coincidência não se verifica, ou seja, em que a utilidade causal esperada de uma acção seja maior ou menor do que a utilidade condicional esperada dessa mesma acção.

Consideremos o seguinte exemplo (Gibbard e Harper 1978: 163): o Rei David deseja a mulher de Uriah, Bathsheba. Contudo, a sua consciência diz-lhe que roubar Batsheba ao marido é injusto. Além disso, David 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 o grau de carisma de um rei 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, David não sabe se é carismático, embora saiba que é injusto roubar a mulher de outro homem. Que deve David 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 $pr(B \Boxrightarrow R)$ e a $pr(R|B)$. O que se pode constatar neste caso é que a Condição X não se verifica, i.e., a condicional contrafactual $B \Boxrightarrow R$ não é epistemicamente independente da acção B. Ou seja,

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

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

$$pr(B \Box \rightarrow R|B) = pr(R|B) > pr(R) = pr(B \Box \rightarrow R).$$

Ou seja, neste caso (1) é falsa, pois

$$pr(B \Box \rightarrow R) < pr(R|B).$$

Se a Condição X não se verifica, podemos então concluir que, ao contrário do que sucedia acima, a UCE (B) $\neq$ UCaE (B). No caso da UCE temos o seguinte: Se David souber que $\neg B$ é o caso (que se abstém de roubar Bathsheba), então terá uma razão para pensar que é carismático e, como tal, não-propenso a gerar revoltas. Assim, a $pr(R|B)$ é maior do que a $pr(R|\neg B)$. Com efeito, para uma dada probabilidade, que não precisa de ser muito elevada (de haver uma revolta dado o roubo de Bathsheba), a UCE ($\neg B$) 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, neste caso, o cálculo da UCaE? David 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 resultado. Neste caso, a UCaE (B) é maior do que UCaE ($\neg B$).

Verificou-se, neste caso, em que a acção B não é epistemicamente independente da condicional contrafactual $B \Box \rightarrow R$, 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 David. Conclui-se, então, que a acção que David deve executar, 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.

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, o previsor prevê-lo-ia',

na condição de eu escolher a caixa opaca, é maior do que a probabilidade incondicional dessa mesma contrafactual:⁵

$$pr(C1 \Box \rightarrow PC1|C1) > pr(C1 \Box \rightarrow PC1).$$

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 numa previsão correcta. Do que se conclui que $pr(C1 \Box \rightarrow PC1) \neq pr(PC1|C1)$. E, portanto, sabemos também que no caso do PN, $UCE(C1) \neq UCaE(C1)$. Também como no exemplo anterior, dado que a acção de escolher a caixa opaca é causalmente independente da previsão, a $pr(C1 \Box \rightarrow PC1) = pr(PC1)$. Em §1 calculou-se a UCE de C1 e C2; calcule-se agora a UCaE de ambas. Suponha-se 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, $pr(C1 \Box \rightarrow PC1) = pr(C2 \Box \rightarrow PC1) = \mu$.

$$\begin{aligned} UCaE(C1) &= pr(C1 \Box \rightarrow PC1) \times 1,000,000 + pr(C1 \Box \rightarrow PC2) \times 0 = \\ &= \mu \times 100 + (1 - \mu) \times 0 = 100\mu \end{aligned}$$

$$\begin{aligned} UCaE(C2) &= pr(C2 \Box \rightarrow PC1) \times 1,001,000 + pr(C2 \Box \rightarrow PC2) \times 1000 = \\ &= \mu \times 101 + (1 - \mu) \times 10 = 91\mu + 10 \end{aligned}$$

Assim, $UCaE(C2) - UCaE(C1) = 10 - 9\mu$, e como $\mu \leq 1$, esta diferença é sempre positiva. Logo, para qualquer valor de μ , a $UCaE(C2)$ é sempre maior do que a $UCaE(C1)$.

Pode-se, agora, reformular 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 também um conflito entre dois princípios acerca do que fazer, os quais correspondem a duas maneiras diferentes de calcular a utilidade esperada, sendo que uma delas se encontra, de facto, em conflito com o princípio da dominação. O que se constatou foi que, em determinados casos,

⁵ A fim de melhor compreender a razão pela qual estas probabilidades divergem, considere-se que, caso se escolha a caixa opaca, o mundo será tal que o previsor o previu, portanto o mundo possível mais próximo em que se escolhe a caixa opaca terá que ser um mundo em que o previsor previu essa escolha.

apenas uma dessas maneiras de calcular a utilidade de uma acção consegue seguir o rasto 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. Portanto, se a nossa intuição 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), 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 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.

A teoria de Gibbard e Harper (1978) não é a única teoria causal que existe. Brian Skyrms (1980) e David Lewis (1981) apresentaram também as suas próprias versões. As três têm em comum a exigência de incorporação de nexos causais no cálculo da utilidade esperada. Apesar de existirem diferenças entre elas, Lewis (1981) argumentou também que as três constituem basicamente a mesma. Por esta razão, é possível apresentar uma definição formal da teoria que consiste numa generalização das várias teorias causais existentes. A ideia central é a de que o valor causal das acções consiste em utilidades esperadas do tipo incondicional (à maneira de Savage 1954), 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 hypothesis*. Para uma dada acção A ,

$$\text{UCaE}(A) = \sum_K pr(K) \times u(A \wedge K)$$

As proposições em K mantêm fixa a nossa perspectiva 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. 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 outra acção ser realizada.

A distinção entre as diferentes teorias causais depende da maneira como cada uma delas interpreta K . Uma maneira natural de o fazer é considerar K uma partição de mundos M aos quais o agente atribui uma probabilidade diferente de 0; ou seja, poder-se-á, à maneira de Gibbard e Harper, considerar cada $k \in K$ como uma proposição contrafactual do tipo $A \Box \rightarrow M$, em que as nossas crenças acerca de relações de causalidade são interpretadas em termos de dependência contrafactual. A utilidade causal esperada de uma acção passa a ser uma média ponderada 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, $pr(A \Box \rightarrow M) > pr(\neg A \Box \rightarrow M)$.⁶

Segundo Skyrms (1980), a partição K poderá consistir numa especificação daqueles factores que se encontram fora da influência das nossas acções, 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 – 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 a fórmula de Skyrms:

⁶ Existe um problema com a caracterização contrafactual de K . Para que a função $pr(A \Box \rightarrow B)$ 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 $A \Box \rightarrow B$ satisfaça, entre outras, a *lei do terceiro excluído condicional*, ou seja, desde que $(A \Box \rightarrow B) \vee (A \Box \rightarrow \neg B)$ seja uma verdade lógica. Uma maneira de garantir a caracterização pretendida consiste em pressupor que existe realmente um, e apenas um, mundo possível mais próximo em que B é verdadeira (Stalnaker 1968). Contudo, Lewis (1973) mostrou que a lei e o pressuposto em causa são insustentáveis. 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. Lewis (1976) concebeu um método para calcular a probabilidade de contrafactuais, designado por visualização [*imaging*], que pode ser generalizado (Gärdenfors 1988) de maneira a lidar com os casos em que o pressuposto acima não se verifica.

$$\text{UCaE}(A) = \sum_{ij} \text{pr}(K_i) \text{pr}(C_j | K_i \wedge A) u(C_j \wedge K_i \wedge A)$$

Esta fórmula oferece o resultado intuitivamente correcto no caso de David e Batsheba e uma recomendação bicaixista no PN. Se tivermos uma preferência pela análise contrafactual da causalidade (Lewis 1973b), então esta opção surge como natural. A contrafactual capta naturalmente, e descreve adequadamente, a forma de raciocínio que os agentes empregam nas suas deliberações acerca do que fazer. Por outro lado, se se preferir uma interpretação do conceito de causalidade em termos de dependência probabilística (Suppes 1970), então a opção pela fórmula de Skyrms talvez seja a mais apropriada.⁷

3 Teoria evidencial da decisão: Propostas revisionistas

Em §2 viu-se que existem maneiras diferentes de interpretar a noção de independência probabilística. A primeira pode ser denominada como independência evidencial (ou epistémica) 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 ser o caso. 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; como no caso do PN, pode haver independência causal sem que haja independência evidencial. Esta distinção encontra-se na base das designações das teorias em disputa: a teoria evidencial da decisão (Jeffrey 1964) e as várias versões da teoria causal da decisão (Gibbard e Harper 1978; Skyrms 1980; Lewis 1981).

⁷ Na literatura conhecida (Lewis 1980, Skyrms 1988, Gaifman 1988), existe ainda uma outra maneira de substituir K na fórmula da utilidade causal, interpretando as *dependency hypothesis* 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 . 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.

Ellery Eells (1982, 1984a, 1984b) apresentou uma defesa da teoria 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 (1985a: 195), para pôr a funcionar a teoria evidencial são apenas necessárias as conectivas lógicas, as operações de adição e multiplicação, e os conceitos de probabilidade e utilidade subjectivas. A teoria causal, por seu lado, necessita de tudo isto e ainda da condicional contrafactual; ou seja, temos de ser capazes de atribuir probabilidades subjectivas a condicionais contrafactuais. Este último empreendimento enfrenta enormes dificuldades, pois o nosso entendimento do conceito de causalidade é imperfeito e sujeito a análises distintas. Mesmo adoptando a análise de Lewis (1973b), dada a vagueza inerente à determinação do valor de verdade destas proposições, a nossa capacidade de lidar com o conceito de causalidade permanece imperfeita. 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.⁸

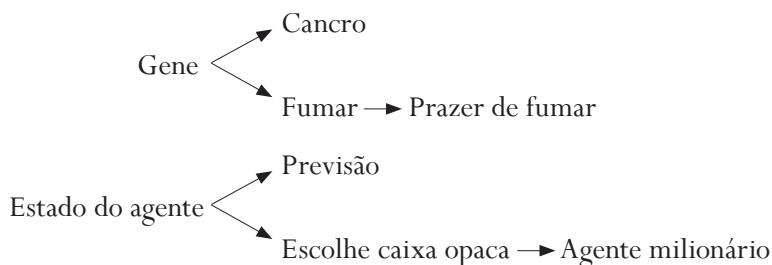
Estes casos têm todos a mesma estrutura e existe uma fórmula para os produzir. Todos eles podem ser designados como *Problemas de Newcomb de Causa Comum* (PNCC). Ao revelar-se esta estrutura, pode-se mais facilmente compreender como funcionam as relações de dependência probabilística e os motivos pelas quais a teoria evidencial oferece uma solução errada para estes problemas. 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

⁸ Até à data do argumento de Eells existiam duas outras vantagens da teoria evidencial. Em primeiro lugar, não existia um teorema da representação para a teoria causal, ao contrário do que acontecia com a teoria de Jeffrey (ver Bolker 1966). 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 acontecia com a teoria evidencial. Contudo, existem hoje teorias causais às quais isto já não se aplica, 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.

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.

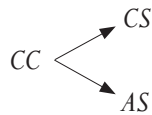
Supondo que um determinado agente sente uma vontade irresistível de fumar e retira daí enorme prazer, o que deve ele fazer? Como se pode facilmente constatar, fumar domina a opção contrária: fumar e ter cancro é melhor do que não fumar e ter cancro, e fumar e não ter cancro é melhor do que não ter cancro e não fumar. Não havendo qualquer relação causal entre fumar e o desenvolvimento da doença, fumar é a opção racional. Contudo, dada a relação de dependência probabilística em causa, a teoria evidencial recomenda que não se fume.

Este caso e o PN partilham algo 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.⁹ A estrutura causal dos dois problemas pode ser representada da seguinte maneira:



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 da seguinte maneira:

⁹ Na verdade, este é um pressuposto da análise de Eells; importa, pois, notar que existem versões do PN em que a previsão não é causada por um estado do agente.



Gibbard e Harper referem-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 tem qualquer eficácia na produção da consequência desejada. Skyrms (1980) 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.

A revisão da teoria evidencial que faz uso destas distinções é conhecida como *tickle defense* (Eells 1982). Em *o 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 espelhada na relação de dependência condicional que justifica a aplicação do princípio evidencial de Jeffrey. Mas, como foi visto, o cerne do problema encontra-se no facto de, entre acções sintomáticas e consequências sintomáticas, não se verificar qualquer relação de causalidade. Assim, há que quebrar de algum modo esse vínculo evidencial. Uma maneira de alcançar este objectivo passa por 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. Existirá, assim, algum elemento da deliberação do agente que o possa ajudar a identificar a presença da causa comum? A análise da estrutura geral destes problemas oferece-nos uma resposta. No caso do fumador, a acção sintomática é acompanhada pelo desejo de fumar – o *tickle*. Ou seja, 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 uma defesa da teoria evidencial para todos os casos possíveis. 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 a acção maximizadora: a fórmula utilizada para o cálculo da utilidade esperada, as crenças do agente e os seus desejos. Como se supõe 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 os 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: condicionarizar as probabilidades dos estados não apenas às acções sintomáticas, mas também à existência de eventuais *tickles*. Se antes tínhamos a desigualdade

$$pr(\text{cancro}|\text{fumar}) > pr(\text{cancro}|\text{ñ fumar}),$$

agora, acrescentando os *tickles* aos condicionantes, passamos a ter a seguinte igualdade:

$$pr(\text{cancro}|\text{tenho tickle} \wedge \text{fumo}) = pr(\text{cancro}|\text{tenho tickle} \wedge \text{ñ fumo}).$$

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: tenhamos ou não o gene, fumar é sempre melhor. Esta noção de bloqueio/tapamento [*screening*] constitui a 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 ou físicas.¹⁰

10 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 (não fumar). 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.

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. Jeffrey (1983) fez um importante acrescento à sua teoria, designado por *ratificacionismo*, que permite contornar este problema. Passamos, então, a incluir o seguinte princípio na nossa teoria da decisão: 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* já não é um estado mental associado à acção sintomática, e que o agente reconhece como tal, mas sim a sua própria decisão de escolher uma ou outra acção. Isto implicar aceitar que a decisão é conceptualmente distinta da acção final e que, para qualquer decisão d e quaisquer duas acções alternativas a e b ,

$$pr(d(a) \wedge b) > 0.$$

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 um.

A ideia fundamental do ratificacionismo 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 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.¹¹ Se quisermos, podemos definir a

Eells (1985b) respondeu que a maneira correcta de abordar um problema enfrentado por um terceiro consiste em determinar o que eu faria (dadas as minhas crenças), caso estivesse na posição desse agente (dados os seus desejos). Nesse caso, as minhas crenças iriam certamente gerar *tickles* (eventualmente diferentes dos dele), aos quais a minha deliberação deverá prestar a devida atenção.

¹¹ Na verdade, a decisão é aquilo que se designa por *metatickle*. Dada a possibilidade mencionada de não existir um *tickle*, i.e., de não existir uma manifestação

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. Nas palavras de Jeffrey (1983: 16),

'(...) escolhe para a pessoa que esperas vir a ser quando tiveres escolhido.'

Aplique-se a análise ratificacionista ao PN. Considere-se 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

$$\begin{aligned} & pr\ d(opaca) \text{ (prevê opaca|escolhe opaca)} = \\ & = pr\ d(opaca) \text{ (prevê opaca|escolhe duas)}. \end{aligned}$$

Isto significa que 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

$$\begin{aligned} & pr\ d(2 \text{ caixas}) \text{ (prevê opaca|escolhe opaca)} = \\ & = pr\ d(2 \text{ caixas}) \text{ (prevê opaca|escolhe duas)}. \end{aligned}$$

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.

Calcule-se, agora, para a pessoa que o agente é, após ter tomado a sua decisão, a utilidade condicional esperada de cada acção:

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. A crítica que será aqui apresentada à estratégia ratificacionista equivale, portanto, a uma crítica à *metatrickle defense* de Eells.

$$\begin{aligned} \text{UCE}_{d(C1)}(C1) &= pr_{d(C1)}(PC1|C1) \times u(PC1 \wedge C1) + \\ &+ pr_{d(C1)}(PC2|C1) \times u(PC2 \wedge C1) = \\ &= 1 \times 1.000.000 + 0 \times 0 = 1.000.000 \end{aligned}$$

$$\begin{aligned} \text{UCE}_{d(C1)}(C2) &= pr_{d(C1)}(PC1|C2) \times u(PC1 \wedge C2) + \\ &+ pr_{d(C1)}(PC2|C1) \times u(PC2 \wedge C1) = \\ &= 1 \times 1.001.000 + 0 \times 1000 = 1.001.000 \end{aligned}$$

$$\begin{aligned} \text{UCE}_{d(C2)}(C1) &= pr_{d(C2)}(PC1|C1) \times u(PC1 \wedge C1) + \\ &+ pr_{d(C2)}(PC2|C1) \times u(PC2 \wedge C1) = \\ &= 0 \times 1.000.000 + 1 \times 0 = 0 \end{aligned}$$

$$\begin{aligned} \text{UCE}_{d(C2)}(C2) &= pr_{d(C2)}(PC1|C2) \times u(PC1 \wedge C2) + \\ &+ pr_{d(C2)}(PC2|C2) \times u(PC2 \wedge C2) = \\ &= 0 \times 1.001.000 + 1 \times 1000 = 1000 \end{aligned}$$

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. Para tornar esta estratégia 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.

Mas será que a estratégia ratificacionista coloca realmente a teoria evidencial em pé de igualdade com a teoria causal? Existe pelo menos uma razão para duvidar disto. 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 PNCC. Considere-se um problema de decisão em que o previsor não só prevê com elevado grau de probabilidade a nossa acção, como também prevê 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 $C1 \wedge d(C1)$, como também o é a prever $C2 \wedge d(C1)$. Considere-se a seguinte matriz com os *payoffs* desta nova versão do problema:

	$P(C2 \wedge d(C2))$	$P(C1 \wedge d(C2))$	$P(C2 \wedge d(C1))$	$P(C1 \wedge d(C1))$
$C2 \wedge d(C2)$	1000	1,001,000	1000	1,001,000
$C1 \wedge d(C2)$	"	"	"	"
$C2 \wedge d(C1)$	0	1,000,000	0	1,000,000
$C1 \wedge d(C1)$	"	"	"	"

De modo a aplicarmos a teoria evidencial, temos de saber quais são as probabilidades condicionais em causa, i.e., 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 quatro previsões possíveis):

	$P(C2 \wedge d(C2))$	$P(C1 \wedge d(C2))$	$P(C2 \wedge d(C1))$	$P(C1 \wedge d(C1))$
$pr(* C2 \wedge d(C2))$	1000	1,001,000	1000	1,001,000
$pr(* C1 \wedge d(C2))$	"	"	"	"
$pr(* C2 \wedge d(C1))$	0	1,000,000	0	1,000,000
$pr(* C1 \wedge d(C1))$	"	"	"	"

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 acção final como a decisão inicial.

Se considerarmos, por exemplo, que a UCE ($C2 \wedge d(C2)$) = $UCE_{d(C2)}(C2)$, 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:

$$\begin{aligned} \text{UCE } (C1 \wedge d(C2)) &> \text{UCE } (C2 \wedge d(C2)), \\ \text{UCE } (C1 \wedge d(C1)) &> \text{UCE } (C2 \wedge d(C1)). \end{aligned}$$

Conclui-se que nesta versão do PN, a acção ratificável é a acção monocaixista. Outros exemplos têm sido oferecidos com o intuito de demonstrar que a estratégia ratificacionista não é um guia fiável para a acção. Tem-se argumentado que esta estratégia não só oferece, por vezes, recomendações contrárias às da teoria causal da decisão (Sobel 1986), mas também que a sua aplicação tem como resultado a identificação de mais do que uma acção óptima [*optimal*] (Jacinto 2011). Joyce (2012) chegou também à conclusão de que a aplicação do princípio, em certos problemas de decisão, conduz a uma situação em que nenhuma das acções disponíveis é dada como ratificável. Estes resultados têm certamente provocado danos consideráveis na credibilidade da teoria evidencial.¹²

¹² Uma outra linha de disputa quanto à solução do PN, que vale a pena mencionar, consiste em analisar as premissas dos argumentos originais (ver secção §1) e disputar o seu valor de verdade consoante a adopção de uma das resoluções para a vagueza da semelhança entre mundos possíveis. Segundo Lewis (1973b, 1979), existe uma resolução *standard* da vagueza que depende da noção de assimetria da dependência contrafactual e que nos faz aceitar 'se estivesse mais frio, o termómetro indicaria uma temperatura menor', mas rejeitar 'se o termómetro indicasse uma temperatura menor, estaria mais frio'. De acordo com a resolução *standard*, as premissas 1 e 4 do argumento monocaixista não podem ser ambas verdadeiras. Contudo, Horgan (1981) argumentou que no PN, devido a razões pragmáticas, devemos aceitar o que Lewis designa como uma resolução *retroactiva* [*backtracking*] da vagueza, a qual viola a assimetria da dependência contrafactual; adoptando esta resolução, são agora as premissas 2' e 5' do argumento bicaixista que não podem ser ambas verdadeiras. Note-se, contudo, que não é obrigatório adoptar-se a semântica de Lewis (1973a) (para contra-exemplos a esta ver Tichy 1976); se se adoptar, por exemplo, uma semântica de tipo probabilístico (van Fraassen 1976), em que as contrafactuais não constituem discurso acerca de factos, os argumentos originais podem ser novamente colocados em pé de igualdade.

4 Contra-exemplos à teoria causal

Andy Egan (2007) apresentou dois problemas de decisão em que a aplicação da teoria causal dá origem a recomendações que, segundo o próprio Egan, são intuitivamente erradas. Tendo ambos os problemas a mesma estrutura, basta apresentar um deles:

O Botão do Psicopata

Maria está a reflectir sobre se há-de, ou não, carregar no botão que mata todos os psicopatas. Seria muito melhor, pensa ela, viver num mundo sem psicopatas. Infelizmente, Maria tem quase a certeza de que apenas um psicopata carregaria em tal botão. Maria prefere muito mais continuar viva a morrer deixando para trás um mundo sem psicopatas. Deve Maria carregar no botão?

Podemos ilustrar este problema através da seguinte tabela de decisão:

	Psicopata	Não Psicopata
Disparar	Péssimo	Ótimo
Não disparar	<i>Status quo</i>	<i>Status quo</i>

Este é um exemplo em que, na suposição de certas crenças e certos desejos, a teoria causal recomenda, de facto, que Maria carregue no botão, acção que, segundo Egan, trará consigo os piores resultados. Ou seja, dado um grau de crença suficientemente elevado em como não se é psicopata, e dada uma utilidade subjectiva igualmente elevada de viver num mundo sem psicopatas, a teoria causal recomenda que se prima o botão. Por exemplo, seja C a acção de premir o botão e P o estado do mundo em que Maria é psicopata,

$$pr(C \wedge P) = 0.16 / u(C \wedge P) = -30; pr(C \wedge \neg P) = 0.08 / u(C \wedge \neg P) = 10 \\ pr(\neg C \wedge P) = 0.04 / u(\neg C \wedge P) = 0; pr(\neg C \wedge \neg P) = 0.72 / u(\neg C \wedge \neg P) = 0.$$

Portanto,

$$U_{caE}(C) = pr(P) \times u(C \wedge P) + pr(\neg P) \times u(C \wedge \neg P) = \\ = 0,2 \times -30 + 0,8 \times 10 = 2$$

$$U_{caE}(\neg C) = pr(P) \times u(\neg C \wedge P) + pr(\neg P) \times u(\neg C \wedge \neg P) = 0.$$

Perante este exemplo, o defensor da teoria causal parece ter duas maneiras de sustentar a sua posição: defender, contrariamente à sugestão de Egan, que não é verdade que $\neg C$ é a acção racionalmente superior, ou encontrar uma estratégia que lhe permita defender que a teoria causal não recomenda C . Joyce (2012) adoptou a segunda destas estratégias, embarcando num projecto revisionista da teoria causal, com o intuito de acomodar a teoria àquelas que, supostamente, seriam as intuições de racionalidade associadas a este tipo de exemplos. Para o efeito, Joyce (2012: 127) propôs que se acrescentasse à teoria um *princípio de informação completa*: 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.

O *botão do psicopata* seria, portanto, um exemplo de uma situação em que o agente não deverá agir apenas com base no cálculo inicial da utilidade esperada, pois o resultado aí obtido pode ter uma influência decisiva na posterior deliberação do agente, alertando-o para aspectos do seu carácter que são causalmente relevantes para avaliar o problema em causa. Concretamente, pressupondo a racionalidade de Maria, ela encarará o resultado do cálculo como uma razão para agir de determinada maneira, sentindo-se motivada para carregar no botão. Contudo, quanto mais decidida a fazê-lo ela se encontra, maior será a probabilidade que ela atribui à possibilidade de ser psicopata, o que, por sua vez, lhe dará uma razão para, afinal, não carregar no botão. Ou seja, a análise que o agente faz do resultado do cálculo contribui para a alteração da probabilidade de um estado, ser psicopata, 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, $pr(P|u(C) = 2) \neq pr(P)$: a probabilidade de ser psicopata, dado que a utilidade de disparar tem valor dois, é maior do que a probabilidade incondicional inicial que Maria atribui à crença em como não é psicopata. Isto significa que se deverá 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. 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.

Para que a teoria causal recomende disparar no *botão do psicopata*, várias condições terão de ser preenchidas (Joyce 2012: 130-131):

1. A deliberação tem de chegar a um ponto em que a utilidade esperada de C deixe de ser causalmente relevante para a probabilidade de P; ou seja, $pr_t(P|u_t(C) = x) = pr_t(P)$. Isto é exigido pelo *princípio da informação completa*.
2. $u(C) > u(\neg C)$.
3. A probabilidade de P tem de aumentar ou diminuir dado o cálculo em t de $u(C)$; i.e., quando $x \neq 0$, $pr_t(P|u_t(C) = x) \neq pr_0(P)$.¹³

Ora, não é possível satisfazer simultaneamente estas três condições no problema em causa. Os dados do problema garantem-nos que a condição 3 é satisfeita. Mas, nesse caso, a condição 1 só pode ser satisfeita quando $x = 0$, o que exige $pr_t(P) = 0.25$.¹⁴ Mas, para que 2 seja verdadeira, é necessário que a $pr_t(P) < 0.25$, pois, como vimos, a $pr_0(P) = 0.20$, quando a $u_0((C) = 2) < u_0((\neg C) = 0)$.

Qual é a conclusão que podemos retirar destes valores? Que, após a consideração de toda a evidência disponível, a teoria causal não recomenda premir o botão, ao contrário do que pensa Egan? Para o fazer teria de ser satisfeita uma condição 2', diferente da condição 2, nomeadamente, que $u_t(\neg C) > u_t(C)$. Esta desigualdade seria verdadeira se, e somente se, $pr_t(P) > 0.25$; mas, dado que a satisfação de 1 exige $pr_t(P) = 0.25$, conclui-se que a teoria causal também não recomenda não-premir o botão. Mais, as condições 1 e 3 são satisfeitas, garantindo

¹³ Quando a utilidade de disparar ou de não-disparar aumenta ou diminui, sem que a probabilidade de ser psicopata 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, a probabilidade de ser psicopata tem de aumentar ou diminuir dado o cálculo em t da $u(C)$.

¹⁴ $U_{caE}(C) = 0.25 \times -30 + 0.75 \times 10 = 0$.

que toda a informação relevante é tida em consideração, se, e somente se, $u(C) = u(\neg C)$.

Joyce acredita que o resultado obtido é salutar e favorável às aspirações dos defensores da teoria causal, pois atingido esta espécie de equilíbrio deliberativo, a teoria não recomendaria qualquer uma das duas acções disponíveis. Contudo, Joyce conclui a partir daqui que qualquer uma das acções é racionalmente permissível, dado que é possível saber *ex ante* que, embora haja uma certa probabilidade de nos virmos a arrepender, seja qual for a nossa decisão (0.25 caso primamos o botão e 0.75 caso contrário), existe igualmente uma certa probabilidade de não nos virmos a arrepender, seja qual for a decisão tomada (0.75 caso primamos o botão e 0.25 caso contrário).

Mas será esta uma situação com que os agentes racionais se sintam confortáveis? Uma teoria da racionalidade instrumental que mostra ser incompleta pode realmente ser um guia fiável para a acção? Uma alternativa poderá passar por admitir que, tendo em conta os valores das probabilidades e utilidades subjectivas do agente em causa, disparar é realmente a acção recomendada, na medida em que esses valores não são classificáveis de acordo com os critérios de racionalidade da teoria. Ou seja, admitindo que existe alguém que atribui uma utilidade de tal modo elevada a um mundo sem psicopatas e que se encontra disposto a jogar com a sua própria vida, por que não aceitar que, para essa pessoa, premir o botão é acção correcta?¹⁵

5 O futuro da discussão: Regresso da teoria evidencial?

Tendo em conta aquilo que foi aqui apresentado, não é abusivo afirmar que até há pouco tempo o resultado do confronto entre a teoria evidencial e a teoria causal tinha vindo a ser favorável a esta última. Uma das razões para isso terá certamente que ver com a solução para o PN. Apesar de o conflito entre intuições ser um fenómeno bastante comum entre aqueles que se vêem confrontados com o problema, é inegável que o número de bicaixistas, principalmente entre os filósofos e os especialistas em teoria da decisão, é claramente superior ao dos

¹⁵ Para outras críticas à posição de Joyce, ver Cantwell (2010) e Ahmed (2014, Cap. 3).

monocaixistas.¹⁶ Este facto não é certamente alheio à força do argumento a partir da dominação, e ao carácter aparentemente óbvio da noção de que algum tipo de conhecimento causal é necessário para avaliar *ex ante* a correcção das nossas decisões acerca do que fazer.

Contudo, Ahmed (1914, 1914b, 2018) apresentou um conjunto de argumentos bastante persuasivos a favor da teoria evidencial, os quais alteraram os pratos da balança e deram 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 a teoria evidencial, e não a causal, que parece oferecer os resultados intuitivamente correctos. Um desses problemas (Ahmed 2014: Cap. 5), relacionado com apostas em eventos já ocorridos, envolve inúmeros 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.¹⁷

Quanto ao PN, Ahmed (2014: Cap. 7) mantém-se fiel ao argumento monocaixista mais popular: Por que não és rico? [*Why Ain'cha Rich?*] Se a nossa crença na fiabilidade do previsor reflectir as estatísticas reais da sua *performance*, então a seguinte consequência é inevitável: quando dois agentes se confrontam numa sessão de jogadas repetidas do PN, o monocaixista irá acrescentar sucessivamente milhão sobre milhão à sua fortuna pessoal, enquanto o bicaixista permanecerá relativamente pobre. Ou seja, ninguém põe em causa o seguinte: o bicaixista está comprometido com uma política de lidar com o PN

16 Acerca deste ponto pode ser consultado o inquérito *online* realizado pela base de dados bibliográfica *PhilPapers*, 19 de Novembro de 2013, a partir de <http://philpapers.org/surveys/results.pl>. Entre aqueles que se designam a si próprios como especialistas, 61% tende para o bicaixismo, contra 21.5% de monocaixistas.

17 Os detalhes de ambos estes exemplos são demasiado complexos e intrincados para que a sua exposição, desenvolvimento e discussão possa caber nos limites deste artigo. Contudo, a sua análise é essencial para uma discussão dos mais recentes desenvolvimentos acerca da disputa entre as teorias evidencial e causal.

que lhe garantirá o pior resultado que alguém confrontado com o PN pode alcançar.¹⁸

A objecção tradicional a este argumento foi apresentada por Joyce (1999: 152-154): 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. Contudo, tal não constitui 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.¹⁹

O problema deste contra-argumento é que a estratégia empregue, que consiste numa elucidação das circunstâncias dos agentes, altera significativamente a natureza do problema. Dado que o tipo de personalidade do agente pode ser entendido como uma causa comum do estado do mundo e da acção empreendida, o mesmo funciona para o agente como um *tickle* (ver §3), 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 problema é que, neste caso, já não nos encontramos perante a versão *standard* do PN.

18 Importa notar que Arntzenius (2008) apresentou um problema de decisão em que, de acordo com a sua análise, a teoria evidencial incorre neste mesmo problema: recomendar sempre a acção que, com elevadíssima probabilidade, virá a mostrar-se sub-ótima.

19 Poderá, contudo, ser implausível que, em todas as situações, o agente possua sempre, no decorrer da sua deliberação, 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), não partilhada por Joyce, 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?'.

Finalmente, Ahmed (2014: 194-201) apresentou um problema de decisão com o objectivo de minar a autoridade do argumento a partir da dominação. A teoria evidencial, embora não deixe de reconhecer a validade destes argumentos, exige um princípio de dominação consideravelmente mais forte: para que se aplique este tipo de raciocínio, não só não deve existir qualquer relação de dependência causal entre acções e estados, mas também não deverá verificar-se qualquer relação de dependência evidencial. Como sabemos, tal não acontece no PN e, portanto, o princípio da dominação não se aplica. A característica inesperada deste novo problema consiste, segundo Ahmed, em verificar-se nele 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 (99%). O que devemos fazer? Gastar os trinta florins numa noite de pândega e diversão, ou passar uma noite de sóbria inquietude a contemplar a nossa esplêndida, embora inútil, armadura?

Ora, dada a 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.

Parece, pois, verificar-se que os estados do mundo são evidencialmente independentes das acções disponíveis, pois o modo como gastaremos o dinheiro em nada modifica a nossas crenças, dada a confiança que temos na fiabilidade do previsor. Por outro lado, parece existir realmente uma relação de dependência causal entre a possibilidade de sobrevivência e o tipo de equipamento militar que envergaremos na batalha.

Mas qual é, afinal, a razão que torna este caso relevante para uma eventual solução para o PN? Sabendo-se que a teoria evidencial e a teoria causal defendem tipos diferentes de dominação, a avaliação de qualquer argumento que recorra a este raciocínio acaba por depender da opção prévia por uma destas teorias. Como tal, a estratégia de Ahmed consiste em encontrar um caso em que, apesar de existir independência evidencial, o raciocínio por dominação continua a ser persuasivo. Se este for realmente o caso, então isso pode constituir evidência de que o que torna os argumentos por dominação adequados não é a independência causal dos estados relativamente às acções, mas sim a sua independência evidencial.

Devemos, então, colocar-nos a seguinte questão: aplicar-se-á a este caso o argumento da dominação? Ahmed não só acredita que sim, mas também que o exemplo não deve ser descartado, pois para além de não ser mais irrealista do que o PN, é também, como aquele, intuitivamente claro. Contudo, ambas estas crenças podem ser postas em causa. Primeiro, não é de todo óbvio que a razão pela qual o princípio da dominação se aplica se deve apenas à independência evidencial: afinal, parece verificar-se não apenas independência evidencial, mas também causal, pois se estamos certos (99%) de que não vamos sobreviver, envergar a armadura não pode ser causalmente relevante para a nossa sobrevivência. Com efeito, nada nos dados do problema sugere a existência de qualquer possibilidade de a nossa acção poder influenciar, seja de que maneira for, a previsão (e *a fortiori* as nossas crenças na probabilidade dos estados), e isso significa que a probabilidade de o previsor estar correcto envergando eu a armadura só pode ser idêntica à probabilidade da contrafactual 'se eu envergassem a armadura, então o previsor estaria correcto'.

É possível argumentar que este raciocínio é falacioso, da mesma maneira que seria falacioso afirmar que no PN a crença na infalibilidade do previsor implica que a acção de escolher ambas as caixas é causalmente irrelevante para determinar se a previsão está ou não correcta. Mas também não é claro que esta analogia seja adequada: é certo que a nossa acção tem influência causal naquilo que obteremos (e daí ser plausível afirmar que ela é a causa da correcção de ambas as previsões), mas o que importa considerar é se a acção tem influência causal na previsão, e não nas consequências obtidas. E a resposta a esta questão parece, de facto, ser negativa. De tudo isto

parece seguir-se também que, afinal, o exemplo de Ahmed não é tão intuitivamente claro quanto o PN.

Estou em crer que, de futuro, a discussão acerca do PN e do confronto entre teorias da decisão passará certamente pela discussão deste exemplo e dos outros identificados nesta secção.²⁰

Diogo Fernandes
Grupo LanCog, Centro de Filosofia da Universidade de Lisboa
diogo.fernandes@campus.ul.pt

Referências

- Ahmed, Arif. 2014. *Evidence, Decision and Causality*. Cambridge University Press.
- Ahmed, Arif. 2014b. Dicing with death. *Analysis* 74: 587-92.
- Ahmed, Arif. 2018. Self-Control and hyperbolic discounting. In *Self-Control & Rationality: Interdisciplinary Essays*, ed. por J. L. Bermúdez. Cambridge University Press.
- Ahmed, Arif. (ed.) 2018b. *Newcomb's Problem*. Cambridge University Press.
- Arntzenius, Frank. 2008. No regrets, or: Edith Piaf revamps decision theory. *Erkenntnis* 68: 277-297.
- Bar-Hillel, Maya e Margalit, Avishai. 1972. Newcomb's Paradox Revisited. *The British Journal for the Philosophy of Science* 23(4): 295-304.
- Bolker, Ethan D. 1966. Functions Resembling Quotients of Measures. *Transactions of the American Mathematical Society* 124: 292-312.
- Cantwell, John. 2010. On an alleged counter-example to causal decision theory. *Synthese* 173: 127-152.
- de Finetti, Bruno. 1964. Foresight: Its Logical Laws, Its Subjective Sources. In *Studies in Subjective Probability*, ed. por H. Kyburg e H. Smokler. Nova Iorque: John Wiley, 93-158.
- Clark, Michael. 2002. *Paradoxes from A to Z*. Routledge.
- Eells, Ellery. 1982. *Rational Decision and Causality*. Cambridge: Cambridge University Press.
- Eells, Ellery. 1984a. Newcomb's Many Solutions. *Theory and Decision* 16: 59-105.
- Eells, Ellery. 1984b. Metatrickles and the Dynamics of Deliberation. *Theory and Decision* 17: 71-95.
- Eells, Ellery. 1985a. Causality, Decision, and Newcomb's Paradox. In *Paradoxes of Rationality and Cooperation, Prisoner's Dilemma and Newcomb's Problem*, ed. por Richmond Campbell e Lanning Sowden. Vancouver: The University of British Columbia Press, 183-213.
- Eells, Ellery. 1985b. Reply to Jackson and Pargetter. In *Paradoxes of Rationality and Cooperation, Prisoner's Dilemma and Newcomb's Problem*, ed. por Richmond Campbell e Lanning Sowden. Vancouver: The University of British Columbia Press, 219-23.
- Egan, Andy (2007) Some counterexamples to causal decision theory. *Philosophical Review* 116(1): 93-114.
- Gaifman, Haim. 1988. A Theory of Higher Order Probability. In *Causality, Chance and Choice*, ed. por Brian Skyrms e W. Harper. Dordrecht: Kluwer, 191-219.
- Gärdenfors, Peter. 1988. *Knowledge in Flux*. Cambridge, MA: MIT Press.

²⁰ Agradeço a Bruno Jacinto (CFCUL e LanCog) as correcções e valiosas sugestões.

- Gibbard, Allan e Harper, William H. [1978]/1980. Counterfactuals and Two Kinds of Expected Utility. In *Ifs*, ed. por W. L. Harper, R. Stalnaker, e G. Pearce. D. Reidel Publishing Company, 153-190.
- Horgan, Terence. [1981]/1985. Counterfactuals and Newcomb's Problem. In *Paradoxes of Rationality and Cooperation, Prisoner's Dilemma and Newcomb's Problem*, ed. por Richmond Campbell e Lanning Sowden. Vancouver: The University of British Columbia Press, 159-182.
- Horgan, Terence. 1985. Newcomb's Problem: A stalemate. In *Paradoxes of Rationality and Cooperation, Prisoner's Dilemma and Newcomb's Problem*, ed. por Richmond Campbell e Lanning Sowden. Vancouver: The University of British Columbia Press, 223-34.
- Horwich, Paul. 1987. *Asymmetries in Time: Problems in the Philosophy of Science*. Cambridge, MA: MIT Press.
- Jacinto, Bruno. 2011. 'Newcomb's Problem, Savage's Decision Theory and Ratificationism, manuscrito não publicado, produzido para o projecto Paradoxos: Dedutivos, Indutivos e Práticos (PTDC/FIL/67039/2006).
- Jackson, Frank e Pargetter, Robert. 1983. Where the Tickle Defense Goes Wrong. *Australasian Journal of Philosophy* 61: 95-299.
- Jeffrey, Richard. [1964]/1983. *The Logic of Decision*, 2ª ed. Chicago: University of Chicago Press.
- Joyce, James M. 1999. *The Foundations of Causal Decision Theory*. New York: Cambridge University Press.
- Joyce, James M. 2012. Regret and instability in causal decision theory. *Synthese* 187: 123-145.
- Joyce, James M. 2018. Deliberation and Stability in Newcomb's Problem and Pseudo-Newcomb Problems. In *Newcomb's Problem*, ed. por Arif Ahmed. Cambridge University Press, 138-159.
- Levi, Isaac. 1975. Newcomb's Many Problems. *Theory and Decision* 6: 161-175.
- Lewis, David. 1973a. *Counterfactuals*. Oxford: Basil Blackwell.
- Lewis, David. 1973b. Causation. *The Journal of Philosophy* 70(17): 556-567.
- Lewis, David. 1976. Probabilities of Conditionals and Conditional Probabilities. *Philosophical Review* 85: 297-315.
- Lewis, David. 1979. Counterfactual Dependence and Time's Arrow. *Noûs* 13(4): 455-476.
- Lewis, David. 1980 A Subjectivist's Guide to Objective Chance. In *Studies in Inductive Logic and Probability* (2), ed. by R. Jeffrey. Berkeley: University of California Press, 263-94.
- Lewis, David. 1981. Causal Decision Theory. *Australasian Journal of Philosophy* 59: pp. 5-30.
- 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*, ed. por Richmond Campbell e Lanning Sowden. Vancouver: The University of British Columbia Press, 107-33.
- Price, Huw. 2012. Causation, chance, and the rational significance of supernatural evidence. *Philosophical Review* 121: 483-538.
- Ramsey, Frank. 1931. Truth and Probability. In *The Foundations of Mathematics and Other Logical Essays*, ed. por R. Braithwaite. London: Kegan Paul: 156-98.
- Sainsbury, Mark. 2009. *Paradoxes*. Cambridge University Press.
- Savage, Leonard J. [1954]/1972. *The Foundations of Statistics*, 2ª ed. revista. New York: Dover.
- Skyrms, Brian. 1980. *Causal Necessity*. New Haven, CT: Yale University.
- Skyrms, Brian. 1988. Conditional Chance. In *Probability and Causality*, ed. por J. Fetzer. Dordrecht: Kluwe: 161-78.

- Sobel, John Howard. 1986. Metatrickles and Ratificationism. *Proceedings of the Biennial Meeting of the Philosophy of Science Association* (1), contributed papers.
- Stalnaker, Robert C. 1968. A Theory of Conditionals. In *Studies in Logical Theory*. *American Philosophical Quarterly*, Monograph Series 2. Oxford: Blackwell, 98-112.
- Stalnaker, Robert C. (1980) Letter to David Lewis, May 21, 1972. In *Ifs*, ed. por W. L. Harper, R. Stalnaker, e G. Pearce. D. Reidel Publishing Company, 151-152.
- Steele, Katie e Stefánsson, H. Orri. Decision Theory. In *The Stanford Encyclopedia of Philosophy*, ed. por E. Zalta (Winter 2016 Edition), URL = <http://plato.stanford.edu/archives/win2016/entries/decision-theory/>
- Suppes, Patrick. 1970. *A Probabilistic Theory of Causality*. Amsterdam: North Holland.
- Tichy, Pavel. 1976. A Counterexample to the Stalnaker-Lewis Analysis of Counterfactuals. *Philosophical Studies* 22: 271-3.
- Van Fraassen, Bas. 1976. Probabilities of Conditionals. In *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science* (1), ed. por W. Harper e C. Hooker. D. Reidel Publishing Company, 261-308.