NOTAS SOBRE RACIONALIDADE [1]

PARTE UM

O que faz de uma teoria uma teoria da racionalidade?

A primeira coisa a constatar é que a teoria da racionalidade não é uma província filosófica demarcada. Várias disciplinas científicas utilizam e desenvolvem teorias da racionalidade e o estudo da racionalidade é hoje um assunto técnico. É certo que é uma questão delicada dizer quais são exactamente essas disciplinas. O caso não polémico, por excelência, de teoria da racionalidade é a teoria da decisão racional, incluindo a teoria dos jogos e a teoria da escolha social *(social choice theory)* [2]. A teoria tem aplicações na economia, na gestão, na elaboração da política pública *(public policy)* e mesmo na biologia evolucionista. A teoria das probabilidades é já um exemplo mais polémico – ela pode ser considerada uma teoria da racionalidade tanto quanto for tomada

[1] Estas notas constituem o trabalho preparatório da participação no Projecto de Investigação *Meaning and Rationality* apresentado pelo Gabinete de Filosofia Moderna e Contemporânea da FLUP à FCT em 2000 (embora o projecto tenha sido aprovado não existiram entretanto condições para o levar a cabo).

[2] Numa primeira caracterização a teoria da decisão apresenta modelos das situações em que um agente racional escolhe sozinho (cf. por exemplo VON NEUMANN & MORGENSTERN 1944 ou RAMSEY 1926), a teoria dos jogos apresenta modelos das situações em que um agente racional escolhe em situações em que a sua escolha depende daquilo que outros agentes escolherem, i.e. situações de co-agência (*co-agency*) ou confronto (cf. por exemplo VON NEUMANN & MORGENSTERN 1944) e a teoria da escolha social apresenta modelos para situações em que um agente racional escolhe por outros agentes racionais, i.e. em função das preferências e interesses desses outros agentes, como seu 'delegado' (a obra inicial de referência da teoria da escolha social é ARROW 1963). A teoria da escolha social é muito importante na teoria da governação (*government theory*) e na teoria da democracia.

como uma lógica da crença parcial e do argumento inconclusivo [3], na terminologia de F. Ramsey. Apontar a lógica como teoria da racionalidade é bastante polémico. Menos polémico do que indicar a lógica como investigação da racionalidade é indicar a exploração de sistemas lógicos na Inteligência Artificial, considerada como ciência experimental, como uma investigação da natureza da racionalidade, pela razão que à frente se indica.

Existem boas razões para separar cuidadosamente as questões relativas à inferência e ao raciocínio (processos psicológicos) das questões relativas a implicação, consistência, etc (relações formais). Mas se fosse possível encontrar um denominador comum às teorias da racionalidade poder-se-ia evocar o tratamento teórico de transições que supõem razões, seleccionando assim disciplinas que visam transições ou sequências mentais ou formais mais ou menos abstractamente consideradas. Propõe-se no entanto que só é estritamente legítimo falar de teorias da racionalidade se estas prevêm *agentes*, entidades cujo comportamento adaptado (ou não adaptado) a um ambiente determinado resulta da interacção entre crenças (representações de conhecimento) e desejos (estruturas de finalidades). A noção, à partida neutra e abstracta, de agente tal como é utilizada na economia, na teoria da decisão racional ou na Inteligência Artificial é assim um ponto de referência incontornável numa teoria da racionalidade. Ao longo deste texto falar-se-á por isso de agentes. A centralidade da noção de agente é de resto a razão pela qual é menos polémico considerar que existe investigação da natureza da racionalidade na Inteligência Artificial, que lida com o desenho de agentes artificiais, do que na lógica, que lida com dedução e estruturas formais abstractas.

Note-se que a noção de agente não é tão neutra como aqueles que a utilizam desejariam. É ela que impede, nomeadamente, a identificação sumária entre conhecimento e racionalidade: racionalidade não é conhecimento *tout court* mas uso teórico ou

[3] Cf. RAMSEY 1926. Esta não é obviamente a única utilização da teoria das probabilidades – pense-se na estatística e na física – nem sequer a mais importante mas apenas aquela pela qual a teoria das probabilidades pode ser considerada como uma teoria da racionalidade. Para considerar a teoria das probailidades como uma lógica da crença parcial Ramsey põe de lado, evidentemente, a concepção frequencista comum de probabilidade.

prático de conhecimento por agentes (para alguma finalidade ou objectivo, como fixar mais uma crença verdadeira ou escolher um curso de acção) de uma forma mais ou menos eficaz.

O que é ser racional? A teoria da decisão e o egoísmo do agente

A teoria da decisão racional, que foi desenvolvida no século vinte por matemáticos, estatísticos, economistas e filósofos, constitui a visão normativa actualmente dominante do estatuto e natureza da racionalidade [4]. Ela é aplicada em contextos variados e envolve formalizações complicadas. De acordo com a teoria da decisão racional o agente racional tem preferências determinadas e ordenadas e age de acordo com elas, escolhendo por entre as opções de acção que se lhe oferecem aquela que *maximiza a utilidade esperada* (essa utilidade é a utilidade atribuída pelo agente a um estado do mundo resultante da acção). A definição do agente racional na teoria da decisão supõe portanto que:

1. o agente tem uma estrutura determinada e ordenada de *preferências*
2. o agente atribui *diferentes* utilidades aos resultados (possíveis) das suas acções
3. o agente atribui (uma maior ou menor) *probabilidade* à obtenção dos estados do mundo aos quais atribui diferentes *utilidades*
4. o agente adscreve diferentes *utilidades esperadas* às opções que se lhe oferecem, conforme a probabilidade de obtenção de um dado resultado e a utilidade desse resultado
5. o agente é racional se e só se escolher de modo a *maximixar a utilidade esperada*

A ideia de decisão racional está intimamente ligada com a forma de conceber o comportamento de agentes no mercado e do próprio mercado utilizada nomeadamente na teoria económica clássica [5]. O comportamento do mercado resultaria da interacção entre agentes completamente racionais cada um prosseguin-

---

[4] Cf. por exemplo NOZICK 1993: 41

[5] Como notam por exemplo Shafir e Tversky (SHAFIR&TVERSKY 1995: 77), que trabalham exactamente numa avaliação experimental da racionalidade de agentes, esta é uma concepção perfeitamente apriorista e não baseada em quaisquer resultados experimentais, que tem por isso mesmo muito mais pertinência como pretensão normativa do que como descrição. R. Nozick nota ainda que a caracterização da teoria da decisão é uma caracterização da melhor acção e não exactamente da acção mais racional (NOZICK 1993: 65).

do os seus interesses egoístas e decidindo-se por aquelas acções às quais atribuem a maior probabilidade de originar as melhores consequências para si de acordo com os seus objectivos e com a informação de que dispõem. Espera-se que cada agente racional maximize a utilidade esperada, i.e. vá pelo máximo, opte pela acção que com grande probabilidade conduza ao resultado ao qual atribui a maior utilidade esperada.

Como se vê, basicamente, a teoria da decisão racional analisa condições definidas sobre as preferências de agentes assumindo que utilidades e probabilidades estão implícitas nas preferências demonstradas. Para que seja possível manejar teoricamente as situações de escolha a teoria da decisão atribui valores numéricos às utilidades e probabilidades e propõe um cálculo sobre esses valores [6]. A teoria da decisão lida portanto com medidas e comparações das preferências dos agentes, das utilidades dos resultados *(outcomes)* e da probabilidade de obtenção dos resultados dadas as acções.

Assume-se que os valores envolvidos dizem alguma coisa acerca da estrutura da escolha racional mesmo se sua interpretação é problemática: sem a utilização de valores numéricos não seria sequer possível uma análise da racionalidade da escolha. O problema da noção de utilidade que aparece na definição de acordo com a qual a decisão racional é aquela que maximiza a utilidade esperada reside obviamente no facto de a sua *quantificação* ser controversa (embora largamente praticada na literatura

[6] Diferentes axiomatizações são propostas. Cf. por exemplo VON NEUMANN & MORGENSTERN, 1944, *Theory of Games and Economic Behavior.* Os autores apresentam uma teoria matemática da decisão que se aplica quando um agente está perante opções exclusivas de acção. A axiomatização da utilidade apresentada por John Von Neumann e Oskar Morgenstern, à qual os estudos da decisão racional usualmente se reportam, toma as probabilidades como dadas. Cada acção tem resultados (*outcomes*) aos quais são atribuídas utilidades, i.e. valores. Atribui-se probabilidades condicionais a cada resultado possível R em relação com uma acção, por exemplo a acção A. A utilidade esperada de um dado resultado R de uma acção A é dada pela seguinte fórmula: u (R) x P (R | A). De acordo com a teoria, a racionalidade do agente envolve praticar o acto com a maior utilidade esperada. Antes de Von Neumann e Morgenstern já F. Ramsey tinha antes elaborado uma teoria da crença pacial e do raciocínio inconclusivo baseada numa teoria subjectivista da probabilidade, e nas noções de preferência e expectativa matemática do agente quanto a resultados das suas acções, apresentando propostas para a manipulação numérica desta lógica (cf RAMSEY 1926, *Truth and Probability*).

económica e filosófica). Ao que acresce o facto de as utilidades não serem aparentemente *transferíveis* entre agentes. Este é de qualquer modo um problema comum à economia e à filosofia moral e política utilitarista, cujas histórias de resto sempre foram próximas [7]. A alternativa comum tem sido a utilização do cálculo de utilidades e probabilidades nas teorizações filosóficas e económicas da escolha racional.

A teoria da decisão enuncia ainda determinados constrangimentos sobre a estrutura do agente racional, por exemplo e existência de uma listagem ordenada e hierarquizada *(ranking)* de preferências, a transitividade de preferências (se o agente prefere *a* a *b* e *b* a *c*, deve preferir *a* a *c*), a coerência pelo menos mínima das crenças em qualquer corte temporal, a cada instante, e ao longo do tempo [8] e a capacidade de fazer inferências válidas.

Um agente é racional se escolhe por entre os meios de que dispõe aqueles que são adequados para atingir os fins que deseja atingir mediante um cálculo de utilidades e probabilidades. Essas escolhas são evidentemente feitas mediante cenários de futuro, que configuram os resultados esperados da acções possíveis, e que portanto antecipam as consequências dos cursos de acção alternativos.

Quando se fala de escolhas racionais pensa-se obviamente em situações em que existem opções de acção para agentes, alguma coisa que os agentes pensam que podem fazer. Opções são acções possíveis relativamente às quais o agente ainda não decidiu e por entre as quais, supostamente, ele pode escolher com liberdade. Fazer uma escolha é decidir por uma dessas opções,

---

[7] É um problema importante na filosofia da economia a adequação empírica da ideia de decisor racional, cujas preferências estão ordenadas, são transitivas e que escolhe de modo a maximizar as preferências, ou o resultado das preferências. Nomeadamente, se o comportamento de agentes não é conforme a estas descrições é muito difícil sustentar que as descrições económicas revelam leis causais que governam o comportamento económico. Como alguma ideia de maximização das preferências parece ser imprescindível, o que acontece é que não será defensável pretender que existe algo de análogo à explicação física que seria a explicação económica. É evidente que o mesmo vale para todas as disciplinas intencionais, nomeadamente para a psicologia.

[8] Embora o desconto temporal (i.e. o facto de os agentes racionais humanos, por exemplo, não serem temporalmente imparciais relativamente aos resultados desejados das suas acções: em princípio tendemos a descontar um benefício futuro no presente) seja um dos problemas mais interessantes tratados neste contexto.

vir a querer alguma coisa específica de entre as alternativas possíveis, de acordo com uma razão. Trata-se portanto de uma mudança interna que acontece no agente através de uma razão. Quanto à razão poder-se-ia propor que uma razão seria um estado mental, em parte desejo em parte crença (a crença de que fazendo x se satisfará o desejo y)[9] e em parte entendimento do agente (o agente tem razões para fazer aquilo que deseja de acordo com um dado entendimento da situação).

Se várias coisas lhe aparecerem como 'a melhor' (a escolher, a fazer) um agente racional escolherá uma qualquer, indiferentemente [10].

Aparentemente seriam coisas diferentes para um agente racional escolher a opção que trará o melhor resultado (maximizar) e seguir as suas preferências. Esta é uma ambiguidade quanto àquilo que se entende por 'o melhor' que aparece frequentemente nas teorizações da escolha racional.

Qual é a forma tradicional da escolha racional em situação de incerteza? O apostador e o benefício monetário

O tratamento formal da decisão racional é bastante anterior à teoria da decisão racional e ao desenvolvimento da economia como ciência. Matemáticos como B. Pascal (1623-1662) e D. Bernouilli (1700-1782) procuraram conceber a relação de agentes com bens ou valores em situações em que agentes pretendem obter bens ou evitar males devendo para isso considerar não apenas o bem a obter ou o mal a evitar pelo seu valor para o agente mas também a probabilidade da ocorrência ou não ocorrências de tais bens e males e portanto uma certa 'proporção' que todas estas coisas (os valores e as probabilidades de ocorrência dos resultados) têm em conjunto. O significado de utilidade esperada é muito claro quando o problema da decisão ou escolha racional é formulado para bens e males monetários: a maior ou menor utilidade consiste

---

[9] Existe uma grande literatura filosófica pós-wittgensteiniana (uma vez que os wittgensteinianos defendem que não faz sentido falar de razões como sendo causas) acerca de se as razões podem ser causas. Donald Davidson, nomeadamente, defendeu que razões podem ser causas. Cf SCHICK 1997: 13.

[10] Esta formulação é evidentemente apenas um modo se ocultar um problema sério. No entanto pode-se dizer algo: estas são as situações em que é racional escolher ao acaso, atirar uma moeda ao por exemplo.

directamente em valores monetários maiores ou menores e um agente será racional se escolher de modo a maximizar o benefício (monetário) esperado, i.e. se escolher a acção que provavelmente lhe permitirá receber o valor mais alto. Em suma, ou o apostador é racional ou perde dinheiro sistematicamente (coisa que normalmente ninguém deseja).

O benefício, o bem esperado, é, no caso do apostador, o valor monetário resultante da aposta. A teoria das probabilidades constitui desde há vários séculos um instrumento de abordagem de tal situação, ajudando a definir por exemplo quanto é razoável apostar e quanto se pode razoavelmente esperar ganhar numa determinada situação de aposta.

Como se afirmou, a caracterização da escolha racional supõe uma produção de cenários de futuro, situações possíveis que são os resultados alternativos das várias acções. Além de obrigar a uma opção por entre alternativas, a escolha racional não é sempre feita em condições de certeza. Grande parte das escolhas racionais são feitas em situação de incerteza e de risco. Problemas de escolha em incerteza e com risco são tratados, como se disse, desde há vários séculos por meio da teoria das probabilidades e têm o seu paradigma na figura do apostador ou jogador racional. A decisão por uma acção A envolve risco quando o resultado de A não é conhecido com certeza. Em tais casos, o apostador racional deve considerar não apenas a desejabilidade de um determinado resultado como também a probabilidade da ocorrência desse resultado. O benefício monetário da situação do apostador é tratado actualmente na teoria da decisão de forma neutra, como a 'utilidade' (de uma acção, de uma escolha) para o agente. A utilidade pretende assim ser uma medida neutra da valoração pelo agente e um conceito que pode ser aplicado a todos os 'bens' possíveis.

Assim, na literatura económica e filosófica fala-se da utilidade para o agente de por exemplo escrever um livro, comprar uma casa, mudar de profissão, o que for necessário. O que interessa é que a utilidade mede o quanto o agente deseja alguma coisa, a prefere ou está disposto a preferi-la. Em economia fala-se da utilidade como o grau de satisfação que um bem propicia a um agente. A utilidade é portanto um auxiliar da descrição das preferências do agente.

De que trata a teoria dos jogos?

A teoria dos jogos, elaborada no quadro da teoria da decisão racional é, uma lógica do confronto entre agentes racionais. Foi criada nos anos 40 pelo matemático J. Von Neuman e pelo economista O Morgenstern para modelizar situações em que as escolhas e decisões de agentes racionais dependem daquilo que outros agentes racionais escolhem. Ela permite conceber situações em que as escolhas não são levadas a cabo por agentes isolados, escolhendo sozinhos e sim num contexto em que existe mais do que um agente. O Dilema do Prisioneiro é uma situação emblemática da teoria dos jogos e o paradigma dos jogos racionalmente prejudiciais *(rationally hurtful games)*. Neste problema, dois prisioneiros, conjuntamente acusados de um crime, são mantidos separados. Cada um tem duas opções: confessar ou não confessar. Se nenhum deles confessar cada um terá uma pena de dois anos. Se ambos confessarem, cada um terá uma pena de dez anos. Se um deles (A) confessar e o outro (B) não confessar, A será libertado e B terá uma pena de doze anos. Se for B a confessar, não tendo A confessado, B será libertado e A terá uma pena de doze anos.

As combinações possíveis são as seguintes:

A confessa B confessa	A não confessa B não confessa
A confessa B não confessa	A não confessa B confessa

Os resultados respectivos são os seguintes:

A 10 anos B 10 anos	A 2 anos B 2 anos
A posto em liberdade B 12 anos	A 12 anos B posto em liberdade

O problema é obviamente decidir o que será mais racional para cada um fazer (de facto, deve-se considerar que cada jogador lista hierarquicamente os resultados que prefere, de entre as quatro combinações possíveis). A situação é de alto risco pois nenhum dos jogadores pode ter qualquer certeza quanto ao que o outro fará. E a coisa mais racional a fazer, no caso do Dilema

do Prisioneiro, é aparentemente confessar (esta é a opção *dominante*), embora isso faça com que ambos os jogador fiquem pior do que ficariam se nenhum confessasse, ou se apenas um confessasse (esse seria o melhor resultado para cada um, se fosse ele próprio a confessar, evidentemente). No entanto cada jogador, se for racional, pensará (ou será racional se pensar) 'Faça o outro o que fizer, eu fico melhor confessando, por isso vou confessar'. Assim, vão ambos para a cadeia 10 ano [11], o que é o pior resultado na escala de preferências de cada um.

Situações políticas e diplomáticas, situações de guerra, situações de competição entre agentes económicos e estratégias evolutivas na biologia [12] configuram dilemas do prisioneiro e uma grande quantidade de literatura tem sido produzida avançando propostas relativas à melhor estratégia para enfrentar o dilema.

A situação análoga ao dilema do prisioneiro para vários agentes chama-se 'The tragedy of the commons' [13]. *Commons* são terras comunitárias. Imagine-se que vários lavradores partilham um determinado terreno comunitário para dar de pastar ao seu gado. Tal como a situação está, a terra comum dificilmente suporta a quantidade de gado e cada cabeça de gado a mais que é para lá mandada faz com que o seu rendimento diminua. No entanto, o benefício para cada lavrador de mandar mais uma cabeça de gado supera a sua parte do custo do dano que esse animal pode causar, façam os outros lavradores o que fizerem. Assim, a decisão racional para cada um dos lavradores é enviar mais cabeças de gado para a terra comum. Se todos os lavradores forem agentes racionais todos farão o mesmo e a terra comum será destruída. Este resultado é obviamente menos desejável do que o resultado que teria sido obtido se nenhum enviasse mais gado.

Como se vê, as grandes alternativas nas situações do género Dilema do Prisioneiro são 'desertar' *(defect)*, i.e. agir por si, e co-

---

[11] Como diz R. Nozick, «Individual rationalities combine to produce a joint mess» (NOZICK 1993: 51).

[12] Cf. por exemplo MAYNARD SMITH 1982 e AXELROD 1984 e SCHICK 1997: 103-105.Vários modelos da teoria dos jogos têm sido propostos e explorados para compreender a evolução biológica de características determinadas em termos de estratégias evolutivamente estáveis.

[13] Cf. HARDIN 1968. Para um comentário cf. por exemplo SCHICK 1997: 87.

operar. Na biologia evolucionista têm sido analisadas situações de Dilema do Prisioneiro iterado, i.e. casos em que a mesma situação entre partes é colocada repetidamente para ver qual é a estratégia evolutivamente estável (um conceito usado por exemplo pelo biólogo John Maynard Smith [14]) para os agentes ao longo das gerações). Uma estratégia é uma política de comportamento de um agente (por exemplo: atacar ou fugir sempre na situação x), que não é necessariamente elaborada de forma consciente. Diz-se que uma estratégia é evolutivamente estável quando essa estratégia, uma vez adoptada pela maioria dos indivíduos de uma população não é superável por uma estratégia alternativa, i.e. a estratégia alternativa não 'vingaria' (não viria ser dominante na população) se fosse introduzida. Claro que a melhor estratégia para um agente particular depende daquilo que o resto da população fizer.

Um exemplo usualmente citado é o exemplo das estratégias de luta Pombo e Falcão. Pombo e Falcão não são esses animais mas sim nomes para duas estratégias de luta que agentes de uma população animal podem adoptar: os falcões lutam sempre, fazem-no agressivamente e só se retiram quando feridos; os pombos apenas ameaçam e nunca ferem o adversário. Se um falcão luta com um pombo, o pombo foge e não é ferido. Se um falcão luta com outro falcão, a luta só acaba com um deles ferido ou morto. Se um pombo luta com outro pombo, apenas se ameaçam longamente, durante muito tempo, finalmente retirando-se. Na medida em que os agentes não sabem à partida qual será a estratégia de luta do adversário, o problema é saber o que é que compensa (*pays off*) ser (um falcão ou um pombo). A questão não se põe relativamente ao agente individual, porque aquilo que lhe compensa ser depende do que os outros forem. O problema é saber qual é a estratégia evolutivamente estável, ser pombo ou ser falcão [15].

Voltando à questão mais geral da forma racional de enfrentar o dilema do prisioneiro, testes feitos (através de simulações em computador) mostram que a estratégia estável em situações de

[14] MAYNARD SMITH 1982.

[15] Há uma proporção estável entre pombos e falcões. Atingida essa proporção (5 pombos para 7 falcões) a lucro médio dos agentes que são pombos será igual ao lucro médio dos que são falcões. Estes lucros são calculados atribuindo valores numéricos aos resultados dos confrontos (considerando vitórias, derrotas, perda de tempo em ameaça, etc).

dilema do prisioneiro iterado é a estratégia usualmente chamada '*TIT for TAT*' (retribuição passo a passo, ou olho por olho dente por dente), i.e. a estratégia do agente que ajusta a sua própria estratégia àquilo que o adversário for fazendo [16].

O que mostra a teoria da decisão racional quanto às situações em que existem agentes em interacção?

Uma aportação curiosa da teoria da escolha racional em situação de agentes múltiplos (teoria dos jogos, teoria da escolha social) é ter mostrado que quando cada agente maximiza o seu interesse o resultado pode ser desastroso para todos, ou pelo menos muito pior do que aquilo que poderia ter sido se todos tivessem aceite uma contenção na prossecução dos interesses próprios. A situação da terra comunitária atrás referida é evidentemente um exemplo daquilo que sucede com os recursos naturais (oceanos, ar, florestas, etc). Em termos de estratégias evolutivamente estáveis de comportamentos, o que se verifica é que estratégias evolutivamente estáveis podem trazer aos agentes benefícios médios muito menores do que o que seria o caso se todos os agentes chegassem a um acordo: por exemplo no caso dos pombos e falcões se todos os agentes 'concordassem' em ser pombos teriam um lucro médio muito mais elevado.

Esse facto permitiria argumentar que em contextos de racionalidade interactiva ou de múltiplos agentes, a noção elementar de racionalidade económica, de acordo com a qual o agente racional age movido pelo interesse próprio de modo a maximizar a utilidade esperada, se arrisca a ser *incoerente*. De facto essa noção reporta-se a agentes individuais, considerados isoladamente. O problema é que um agente racional individual não pode agir de forma a maximizar a utilidade esperada assumindo que todos os outros agentes maximizarão identicamente: ele não pode simplesmente atribuir probabilidades a todas as combinações de acções e escolher conforme a hierarquização de

[16] Os resultados são frequentemente interpretados da seguinte maneira: compensa ser disponível (não ser o primeiro a 'voltar as costas'), compensa perdoar (i.e. ter propensão a retomar uma estratégia de cooperação após o adversário a ter abandonado) e compensa retaliar (imediatamente abandonar a estratégia de cooperação se o adversário a abandonou). Este *Tit-for-tat* representa a possibilidade de ligar em cooperação agentes racionais agoístas.

utilidades atribuídas. Agentes racionais em confronto vêm-se, assim, frequentemente obrigados pela sua própria racionalidade a fazer opções desastrosas ao mesmo tempo que prevêm perfeitamente a indesejabilidade das consequências [17].

Qual é a perspectiva assumida nas teorias da racionalidade (i.e. as razões para a acção são razões de quem)?

Teorias como a teoria da decisão racional, na qual se incluem a teoria dos jogos e a teoria da escolha social, são teorias da racionalidade na medida em que supõem agentes racionais, finalidades, acções, e desejabilidade de resultados mesmo que não envolvam qualquer consciência dos agentes relativamente aos *rationales* das suas acções. O imperativo que elas atribuem ao agente racional é, numa primeira caracterização grosseira: *Age sempre de forma a maximizar o teu próprio benefício* (são por isso também chamadas teorias do egoísmo racional). O egoísmo racional não tem no entanto que ser egoísmo-consciente-do-ponto-de-vista-do-agente (o agente não tem que saber o que faz para a teoria o considerar racional).

Recapitulando, a situação estrutural elementar analisada nas teorias da racionalidade pode ser caracterizada da seguinte maneira:

Se o agente A tem o conjunto de crenças C e o conjunto de desejos D, A será um agente *racional* se decidir em cada situação pela opção, e apenas por essa opção, que seja (apareça às luz das suas crenças como sendo) a mais apropriada para atingir as sua finalidades.

Ora, é óbvio que os conteúdos desta caracterização relativamente a uma dada situação podem diferir se se considerar de forma diferente o parâmetro 'conhecimento do agente' ou conjunto C das suas crenças. Existe aí uma alternativa: ou se considera uma relativização ao agente do conhecimento, e portanto uma relativização ao agente daquilo que seria bom para si (apropriado às

[17] Um resultado básico análogo da teoria da escolha social é o chamado Teorema da Impossibilidade de Arrow, que é muito importante para a abordagem teórica dos propósitos da votação democrática e da noção de bem-estar geral. De acordo com o teorema, «várias condições naturais e desejáveis, que aparentemente deveriam ser satisfeitas por qualquer procedimento para determinar o bem-estar (*welfare*) geral ou a alternativa democrática preferida acima de todas, não podem ser satisfeitas conjuntamente» (NOZICK 1993: xv).

suas finalidades) ou se fala em geral e em abstracto daquilo que seria bom para o agente, sem qualquer relativização àquilo que o agente sabe, que é sempre pouco, limitado. Esta última opção é evidentemente muito praticada pelos filósofos e economistas utilitaristas, que têm como principal objectivo teórico considerar as escolhas racionais para o todo e não para cada agente, assumindo (procurando assumir) um ponto de vista imparcial e agregado. O utilitarismo é uma teoria consequencialista [18] do bem-estar (*welfare*) [19] global feita a partir de um ponto de vista 'imparcial'.

Note-se que uma diferença como a acima mencionada (a relativização ou não relativização ao conhecimento do agente individual daquilo que constitui o melhor para si) abre um espaço para a correcção das preferências dos agentes individuais que é muito importante na filosofia política e na filosofia da economia: esse espaço permite argumentar que se o agente possuísse toda a informação relevante não teria a preferência a mas a preferência b e essa seria uma correcção racional. Não é raro ver teóricos utilitaristas, filósofos ou economistas, servirem-se desta diferença de modo a distinguirem entre preferências manifestas e verdadeiras preferências dos agentes individuais, e para considerarem que existem razões para a acção mesmo que o agente não as conheça, o que propicia espaço de manobra para a teorização das políticas públicas que são racionais mas não seriam escolhidas pelos agentes isoladamente (pense na acima referida tragédia das terras comuns).

Evidentemente, o que está aqui em jogo é a possibilidade de um ponto de vista 'impessoal' sobre a 'qualidade' da acção. Mesmo os adversários do pensamento utilitarista admitem que este é fortemente apelativo do ponto de vista epistémico: a injunção que o pensamento utilitarista representa é praticamente 'quem quer ser menos inteligente do que o parceiro do lado?' (supondo que essa inteligência conduzirá o agente individual a uma impessoalidade e imparcialidade que não são psicologicamente muito normais ou primeiras). Essa é a razão pela qual a

---

[18] I.e. que considera que o princípio que deve reger a escolha da acção correcta é a consideração dos resultados dessa acção.

[19] O utilitarismo enquanto teoria da avaliação dos estados de coisas no mundo é um *welfarism*, i.e., a pretensão de que a base correcta para essa avaliação é o bem-estar, a satisfação, o facto de as pessoas obterem aquilo que preferem.

maior parte dos utilitaristas vêm as abordagens dos problemas éticos e políticos feitas a partir de qualquer outra óptica como muito pouco racionais.

No entanto como a possibilidade de um ponto de vista impessoal e agregado sobre a qualidade da acção e da decisão racional, tipicamente desejado pelos utilitaristas, é ainda assim negada por muitos economistas e filósofos, 'o melhor' numa situação de escolha racional é normalmente capturado em termos da listagem de preferências do agente individual como aquilo que está mais alto.

PARTE DOIS

Quais é a estrutura geral das teorias filosóficas da racionalidade?

As teorias filosóficas da racionalidade tomam como mais ou menos estabelecido que as suas duas grandes áreas são a racionalidade da crença e a racionalidade da acção. É de acordo com esse pressuposto que opera uma certa divisão do trabalho na epistemologia normativa contemporânea. Quanto às linhas gerais de análise da racionalidade da acção, elas foram apresentadas na primeira parte destas notas. No que respeita à racionalidade da crença os problemas tratados são os seguintes:

1. O que faz de uma crença uma crença racional?
2. Porque queremos que as nossas crenças sejam crenças racionais?
3. O que é possível fazer para melhorar a racionalidade das crenças?

Quanto ao primeiro problema, básico e definicional, os dois ramos explorados são os seguintes: a racionalidade de uma crença depende das razões que o agente tem para a sustentar e a racionalidade de uma crença depende da fiabilidade (*reliability*) dos processos que a produziram.

Em que é a que a racionalidade da crença se distingue da racionalidade da acção?

As acções racionais de um agente tanto podem ser decisões por uma inferência válida num processos de pensamento como escolhas da melhor opção num curso de acção. A caracterização da agência (*agency*) racional não tem por que estabelecer distinções de princípio entre o que seria uma racionalidade prática (relativa a acções e decisões que afecta planos e intenções do

agente) e uma racionalidade teórica (relativa ao curso dos processos de pensamento e às inferências voluntáriamente controladas, que afecta as crenças do agente). Ou melhor, os processos de pensamento, tanto quanto são racionais e voluntários, também são acções . É certo que alguma diferenças entre racionalidade da crença e racionalidade da acção merecem ser notadas [20]. Por exemplo em situações práticas, de indiferença entre A, B C, escolhas arbitrárias quanto ao que fazer são racionais, enquanto que escolher arbitrariamente o que acreditar em situação de alternativa entre A, B e C não é racional (se eu não sei se Albert foi pelo caminho 1 ou pelo caminho 2 quando estou a segui-lo, não é racional decidir arbitrariamente acreditar que ele foi pelo caminho 2 [21]). Outra diferença entre racionalidade de âmbito prático e de âmbito teórico é a razoabilidade do *wishful thinking* (acreditar ou quase-acreditar naquilo que se gostaria que fosse o caso): se o *wishful thinking* prático é razoável (o desejo de ter uma boa nota no exame faz Jane estudar muito para que seja o caso que teve uma boa nota), o *wishful think* teórico não é razoável (Jane já fez o exame e o desejo de ter uma boa nota leva-a a concluir que teve uma boa nota) [22].

Como se relaciona o núcleo mínimo da noção de racionalidade com as teorias científicas da cognição?

Da caracterização elementar, que pode ser extraída dos modelos formais de racionalidade até agora referidos, caracterização essa que liga representações de conhecimento, estruturas de finalidades e acções ou decisões obtém-se que existe um núcleo mínimo comum quando se fala de racionalidade. Esse núcleo mínimo comum é a racionalidade meios/fins, ou racionalidade instrumental [23], i.e. o uso estratégico de meios por um agente para alcançar os fins pretendidos. Seja qual for a teoria da racionalidade que se apresenta ela deve cobrir esse aspecto. É importante sublinhar que a caracterização se aplica quer à racionalidade da acção quer à racionalidade da crença. O fim prosseguido, no que

[20] CF HARMAN 1995: 179-181

[21] O exemplo é de G. Harman (HARMAN 1995: 180).

[22] Os exemplos são de Harman, em HARMAN 1995: 180-181.

[23] Cf. NOZICK 1993: 136: «the nature of rationality would be, let's suppose, wholly instrumental, but its value would not be».

respeita à racionalidade das crenças, é a verdade das crenças, o evitamento do erro, a obtenção de poder explicativo. No entanto, se esta óptica põe em relevo uma base instrumental do interesse pela verdade, isso não significa que as pessoas ou outros agentes por hipótese consciente desejem acreditar verdades por verem isso como instrumentalmente útil. O que isso significa é que foi devido à utilidade de acreditar verdades que houve 'selecção para' (*selection for*) [24] uma preocupação pela verdade na crenças. É natural pensar na racionalidade das crenças e das acções como um processo dirigido por fins (*goal-directed behavior*), como uma questão relativa a alcançar eficazmente fins, escolher os meios mais apropriados para alcançá-los. Em qualquer dos casos a crença ou a acção têm que ser sensíveis (*responsive*) a razões por ou contra. Assim, como nota R. Nozick, a racionalidade não é qualquer tipo de instrumentalidade, mas instrumentalidade de razões e raciocínio [25]: «se apanhar uma pancada na cabeça ou ingerir mescalina fosse uma maneira de chegar a crenças verdadeiras sobre um determinado tópico (...) então a crença ela própria não seria uma crença racional. Para alguém que soubesse isso, no entanto, poderia ser racional escolher apanhar essa pancada para adquirir a tal crença verdadeira» [26].

O desenho do agente racional resulta em princípio de selecção natural – mas exactamente o que é seleccionado se a racionalidade de agentes resulta de selecção natural?

Antes de mais, quando se fala da evolução da racionalidade fala-se da evolução de um sistema de controlo de um sistema físico – que não é em si e por si nem racional nem inteligente – para produzir resultados com êxito (relativamente a circunstâncias independentes, exteriores).

Nessas condições, em que supostamente poder guiar-se por regras abstractas é vantajoso para o sistema, é razoável pensar que o que é seleccionado não é uma «capacidade de reconhecer conexões racionais válidas independentemente existentes» (o que é razoável pensar é que) existe uma conexão factual e existiu

[24] A distinção entre selecção de e selecção para é uma discussão da filosofia da biologia na qual não se entrará aqui.

[25] NOZICK 1993: 71

[26] NOZICK 1993: 71

selecção entre os organismos para a situação de tal conexão parecer válida, para notar esse tipo de conexão e para que isso conduzisse a certas crenças adicionais, inferência, etc. Há selecção para reconhecer como válidos alguns tipos de conexões que são factuais, i.e. para eles virem a parecer-nos como mais do que apenas factuais» [27]. Mas, como conclui Nozick, «*that something seems self-evidently true to us, does not garantee that it ever was strictly true*» [28].

Já um filósofo como C. S. Peirce abordava a questão da racionalidade perguntando que hábitos (regras ou leis de comportamento) relativos a inferências e transições seria melhor para a mente ter. Evidentemente, a pergunta que se segue é: mas porque se há-se confiar em métodos de pensamento instalados por selecção natural? A única maneira de obter uma resposta é verificar em que proporção de casos eles nos conduzem a verdades. De acordo com esta abordagem, a lógica, note-se, teria o estatuto de auto-controlo de tais processos.

Será então a racionalidade pura e simplesmente identificável com a racionalidade instrumental?

Enquanto característica de sistemas cognitivos físicos, muitos deles resultantes de evolução por selecção natural, a racionalidade é fundamentalmente racionalidade instrumental, a qual cumpre uma função adaptativa relativamente ao ambiente. Saber se a racionalidade é ou pode ser mais do que racionalidade instrumental é um problema em aberto. Como D. Hume afirmou de forma célebre, «não é contrário à razão eu preferir a destruição

---

[27] NOZICK 1993: 109

[28] NOZICK 1993: 109. Cf. também NOZICK 1993: 111 «the strength and depth of our intuitions about certain statements cannot be used as powerful evidence for their necessity if these atatements are of a kind that were they contingent fact, would have led to selection favoring strong intuitions of their self-evidence». Digamos que o evolucionismo aplicado à racionalidade representa uma alternativa à visão kantiana: Kant propôs que os factos empíricos não são a variável dependente, que a sua dependência relativamente à razão explica a correspondência entre eles (entre razão e factos). A visão evolucionista acerca da racionalidade considera que é a razão que é a variável dependente moldada pelos factos, e que é isso que explica a correlação e a correspondência.: «reason tells us about reality because reality shapes reason, selecting for what seems 'evident'» (NOZICK 1993: 112). Esse é para Nozick o significado da ideia segundo a qual «the nature of rationality includes the Nature in it» (NOZICK 1993: 181). A especial aportação de uma perspectiva evolucionista a este problema é trazida pelo facto de a selecção operar sobre a coisa racional a fazer e não sobre a coisa racional a acreditar (NOZICK 1992: 113).

do mundo inteiro a um arranhão no meu dedo» [29]: nada fica dito acerca dos fins com a definição operatória de racionalidade, ela diz respeito exclusivamente ao uso estratégico dos meios para prosseguir fins, sejam estes quais forem. A racionalidade instrumental nada diz acerca de para onde ir, apenas diz alguma coisa acerca de como ir para onde se deseja ir. Evidentemente, grande parte dos problemas filosóficos interessantes acerca da racionalidade relacionam-se com a inseparabilidade em última análise da racionalidade dos meios e da racionalidade dos fins. Não se trata tanto de tipos diferentes de racionalidade como de áreas teóricas em que a clareza e desenvolvimento conceptuais são mais ou menos conseguidos. Os desenvolvimentos técnicos a que se fez referência no início destas notas dizem respeito à racionalidade dos meios (é certo que não se vê bem como se poderia ter uma racionalidade dos fins a não ser por meio de uma racionalidade dos meios...)

O núcleo onde se juntam a racionalidade dos fins e a racionalidade dos meios são os sistemas de preferências dos agentes. É certo que é possível fazer, como se faz na teoria da decisão, i.e. definir constrangimentos formais de coerência, transitividade, etc, sobre estes sistemas de preferências que deverão ser o caso tanto quanto o agente for um agente racional. O problema com que os teóricos se deparam é que as concepções substanciais que os agentes têm do seu bem acabam por intervir na moldagem da lista de preferências [30].

A racionalidade é uma questão de correcção?

É muito importante notar que nas teorias da racionalidade não se trata apenas de imperativos de correcção acerca das práticas (práticas teóricas e práticas *práticas*...) de agentes. É certo que para um conjunto de crenças desejos e acções de um agente constituir uma vida mental racional, esse conjunto (ou os processos sobre ele definidos) tem que ser adequado ou correcto de várias maneiras, no 'interior' (do agente) e relativamente ao exterior. No entanto o que é importante notar é que se não for esse o caso,

[29] HUME 1985, Book II, Section 3, III

[30] Este é o problema organizador da recolha de artigos organizada por DUPUY&LIVET 1997.

pelo menos tendencialmente, não existe, no limite, um agente. A racionalidade é assim constitutiva do agente (seja este um agente consciente como um humano ou um agente artificial inconsciente). Se no início destas notas se afirmou que as teorias da racionalidade propriamente ditas são teorias de agentes, constata-se agora que sem racionalidade não existem agentes. A racionalidade constitutiva do agente [31] forma um pano de fundo sem o qual são inconcebíveis particulares racionalidades, irracionalidades ou erros [32] de pensamentos e acções [33], i.e. sem o qual é impossível a avaliação correctiva das escolhas e decisões do agente.

Aliás, erros não são necessariamente sintoma de irracionalidade do agente: um exemplo seria um erro cometido por qualquer um de nós ao somar 25 parcelas com 5 dígitos. Aconteceu um erro, a soma fica mal feita. No entanto, para ser irracional, um acto ou um pensamento tem que ser 'cometido' deliberadamente pelo agente (dado o conhecimento do agente, o passo dado não é o melhor passo para o conduzir ao objectivo prosseguido, ou dado o resto do seu conhecimento o agente não deveria ter adoptado a crença c, que o conduz a cometer o acto ou pensamento em causa) enquanto que um erro cometido involuntariamente como no exemplo dado não é irracional (não é o caso de naquele momento o agente ter adoptado como crenças suas regras aritméticas inconsistentes com o resto do seu conhecimento acerca de números). Esta diferença será retomada no fim do presente artigo na medida em que se relaciona com uma aparentemente inevitável circularidade na justificação da racionalidade.

O que se deve pensar acerca do estatuto idealizante de teorias da racionalidade como a teoria da decisão racional?

Há sem dúvida problemas (filosóficos?) inerentes ao estatuto idealizante das teorias da racionalidade e nomeadamente ao facto de aparentemente estas requererem agentes racionais ideais, capazes de inferências perfeitas, cujas crenças seriam caracterizadas pelo fechamento dedutivo («as crenças de um agente ideal são dedutivamente fechadas ou fechadas debaixo de implicação

[31] Cf. CHERNIAK 1994

[32] Para a distinção entre irracionalidade e erro, cf. HARMAN 1995: 176-177.

[33] DENNETT 1987

lógica se e só se qualquer proposição logicamente implicada por alguma dessa crenças for também acreditada»[34]) e pela coerência perfeita. Como G. Harman comenta[35], uma tal entidade seria um génio lógico ou um ser dotado de ominisciência divina e não um agente racional real. No entanto, e apesar das idealizações problemáticas que as teorias da racionalidade envolvem, elas estão de boa saúde (ou melhor, colocando a questão como frequentemente o fazem vários filósofos que são teóricos da racionalidade, não existe alternativa para elas quando se trata de conceber o comportamento de agente, e nomeadamente o comportamento – inclusive o pensamento – de agentes humanos, sendo por isso mais razoável escolher as mais desenvolvidas e sofisticadas). É de resto precisamente por essa razão que a filosofia não pode razoavelmente pretende falar sozinha acerca da racionalidade (embora frequentemente o faça como quando com grandiloquência se refere à racionalidade em massa, como em 'a racionalidade ocidental' ou 'a racionalidade científica' – a qual seria uma coisa má ou incompleta).

Será a racionalidade de facto maximização como a teoria da decisão pretende? H. Simon e a racionalidade limitada

Na teoria da decisão a idealização central é a maximização da utilidade esperada. A lógica da maximização da utilidade, por vezes chamada racionalidade bayesiana, aplica-se sem problemas a situações de apostas com resultados monetários. É claro no entanto que a maioria das situações reais de escolha não são assim claras, é claro que é polémico assumir que utilidade e probabilidades em situação de escolha têm uma formulação numérica precisa. No entanto, para além desses problemas, o que é mas importante para muitos críticos[36] é que um agente racional não é de facto um maximizador racional.

Supostamente a racionalidade de sistemas cognitivos físicos como humanos e outros animais resulta de evolução por selecção natural. Todas as inteligências fisicamente realizadas têm recursos (de conhecimento, de tempo) limitados. Uma célebre e influente teoria da racionalidade limitada (*bounded rationality*) deve-se ao

[34] HARMAN 1995: 187.

[35] HARMAN 1995

[36] Cf. por diferentes razões Herbert Simon e Amartya Sen

cientista de computadores Herbert Simon, um dos fundadores da Inteligência Artificial. De acordo com Simon [37], o que num número significativo de casos os agentes racionais procuram fazer é fazer o melhor possível (*satisfice*) e não maximizar ou optimizar. Dadas as limitações de partida (nomeadamente limitações de tempo e de conhecimento), o uso de estratégias heurísticas não maximizadoras por agentes racionais limitados não constitui uma irracionalidade sendo antes o resultado de uma troca (*trade-off*) da correcção por rapidez e utilizabilidade. Ser capaz de funcionar suficientemente bem (i.e. de tomar decisões, de escolher) em condições normais (nomeadamente em tempo real) é mais racional para um agente racional do que funcionar perfeitamente bem.... mas nunca.

Por exemplo na sua obra *The Sciences of the Artificial* [38] Simon analisa os processos de *satisficing* na resolução de problemas. Os processos de *satisficing* são processos de decisão que encontram soluções relativamente boas para problemas em situações muito complexas, situações em que uma decisão tem que ser tomada no desconhecimento e na incerteza – mas não encontram soluções óptimas (estas são inalcançáveis na medida em que não é possível – ao contrário do que a maximização exigiria – medir todas as alternativas através daquilo a que Simon chama uma função de utilidade comum [39]). Na medida em que em larga medida as crenças e escolhas humanas resultam de processos de *satisficing*, elas não são, respectivamente, consistentes e transitivas, como deveria ser o caso se os humanos fossem agentes idealmente racionais. O processo de *satisficing* caracteriza a resolução de problemas por humanos em inúmeras circunstâncias e também, aliás, de acordo com Simon a resolução de problemas pela natureza, na evolução por selecção natural. Note-se que as suposições de optimização em situações de escolha ou de alternativa são comuns não apenas como já se referiu na economia como também na biologia precisamente na teoria da evolução, com as considerações acerca de adaptação. Como diz Simon, a racionalidade está para a psicologia como a adaptação está para

[37] SIMON 1969
[38] SIMON 1969
[39] SIMON 1969: 29

a biologia evolucionista. Simon questiona em ambos os casos as suposições de optimização. O problema dos processos usados pela evolução na criação do *design* do agente racional é muito importante pois põe em jogo a natureza impura da razão real, i.e. da arquitectura dos agentes racionais naturais. De acordo com Simon, o agente real é um *satisficer*, que aceita muito frequentemente «alternativas suficientemente boas não porque prefira o menos ao mais mas porque não tem escolha» [40]. São as suposições (idealizantes) de optimização realtivamente a um agente que é, por natureza, um *satisficer* que são irrealistas. Se no mundo real os agentes são *satisficers*, a racionalidade real é racionalidade limitada.

Se a racionalidade real dos agentes é limitada, como se relaciona essa condição com as teorias idealizantes das transições mentais ou formais para as quais existem razões, como a teoria da decisão e a lógica, i.e. o que é que agentes reais como pessoas têm a ver com a teoria da decisão e com a lógica?

Se a racionalidade real é uma racionalidade limitada, *standards* ideais de racionalidade, por exemplo *standards* lógicos de validade de argumentos ou axiomas da teoria da decisão, são qualquer coisa outra que não uma caracterização, uma descrição, da racionalidade real de agentes. Aliás, como se afirmou não é sequer muito razoável considerar a lógica como uma teoria da racionalidade: a lógica é mais propriamente uma teoria da dedução ou um estudo das propriedades de sistemas formais. Mas a consistir em alguma coisa do ponto de vista da lógica a racionalidade consistiria na validade de derivações, na validade de argumentos. O problema é que a caracterização da validade não é uma descrição daquilo que se passa em agentes reais nas inferências e decisões.

A caracterização que Simon faz dos processos de *satisficing* é uma crítica das idealizações comuns nas teorias da racionalidade comuns. De acordo com a caracterização de Simon, ser racional exige apenas um processo de *satisficing* (fazer o melhor possível em direcção a) a utilidade esperada e não a maximização da utilidade esperada.

---

[40] SIMON 1969: 29

Se a teoria da decisão é utilizada como um modelo do comportamento de agentes na economia, que forma teria uma crítica a esse modelo?

Como se afirmou, o modelo do agente racional está intimamente ligado com a economia. Embora sendo economista, H. Simon, na sua teoria da racionalidade limitada, fala em geral de agentes físicos reais, sejam eles naturais ou artificiais. No entanto o economista e filósofo Amartya Sen, Prémio Nobel da Economia em 1998, fala sobretudo (caricaturando-a) da suposição frequente na economia segundo a qual uma pessoa, um agente económico, é um maximizador racional. No seu conhecido artigo *Rational Fools – A critique of the behavioral foundations of economic theory*, Sen defende que o *homo oeconomicus*, o agente idealmente racional, escolhendo em função de uma listagem única de preferências coerentes, se existisse seria um tolo (*fool*). Evidentemente, o problema é que grande parte da teoria económica assume que esse agente existe (é vulgar encontrar distinções entre economia e sociologia, por exemplo, nas quais se faz apelo ao comportamento racional de agentes razoavelmente bem informados movidos pelo interesse próprio – esses serão o objecto do economista – e outra coisa qualquer que será o objecto da sociologia). A suposição do comportamento racional tem um papel fulcral na ciência económica, por mais que as verificações empíricas dessa suposição sejam raras ou inexistentes.

As razões que Sen apresenta para negar a existência do *homo oeconomicus* são várias. A sua conclusão geral é que a ideia comum segundo a qual o egoísmo é a componente essencial da descrição da motivação dos agentes, sendo o egoísmo do agente pura e simplesmente identificável com a racionalidade, deve ser rejeitada. Ao contrário do que é frequentemente assumido pelos economistas, o egoísmo universal (não agregado mas distributivamente concedido), a ideia segundo a qual todo o agente racional deve ser concebido como movido pelo interesse próprio, não é o melhor critério de racionalidade. Não porque os agente sejam movidos pela adesão a sistemas morais ou porque seja falso que o interesse pessoal tenha um papel fulcral num grande número de decisões de agentes, mas porque a teoria da escolha racional oferece uma caracterização ilusoriamente simples e circularmente fundamentada daquilo em que consiste a motivação

de um agente. De acordo com Sen, é sempre possível fazer parecer com que um agente maximiza a sua utilidade, seja ele egoísta, altruísta ou um militante com consciência de classe.

Antes de mais, a caracterização do agente como um egoísta racional não é particularmente realista (aplica-se sobretudo a duas actividades-tipo, a guerra e o contrato, muito importantes certamente entre as actividades hmanas mas que não esgotam o assunto). Em seguida, porque a teoria da decisão racional parte do pressuposto de que a única maneira de compreender as preferências de um agente é observar as suas escolhas reais e ao mesmo tempo do pressuposto de que as escolhas de um agente são racionais se e só se essas escolhas puderem ser explicadas em função das preferências reveladas. O comportamento racional do agente é assim explicado em função das suas preferências sendo estas definidas em função do comportamento.

Mas, sobretudo, para Sen, a ideia de uma única listagem, multifuncional, de preferências por agente é simplista e desadequada. Ela oculta por exemplo que ao que a pessoa prefere do ponto de vista pessoal se deve acrescentar o que a pessoa prefere do ponto de vista social e isso pode não representar o que lhe traz maior bem-estar a si. De acordo com Sen, é necessária uma técnica de meta-classificação de preferências de modo a exprimir juízos morais (que também existem e também podem ser racionais) para alcançar um maior realismo na descrição das preferências das pessoas. Uma das intenções de Sen é capturar o 'empenhamento' como componente do comportamento dos agentes, o qual supõe uma separação entre escolha e bem-estar (*welfare*) pessoal. Uma pessoa pode preferir racionalmente uma opção que não se traduz na maximização do seu próprio bem-estar e esse facto é dificilmente enquadrável nos termos do egoísmo racional.

A existência de outro tipo de caracterização das motivações do agente não faz do agente um agente irracional, simplesmente questiona a concepção egoísta e consequencialista de racionalidade que é assumida tacitamente em grande parte do trabalho teórico dos economistas.

A importância de análises como a de Sen reside no facto de inegavelmente a economia, como de resto qualquer ciência intencional, supôr a racionalidade. Aliás, de todas as chamadas ciên-

cias sociais e humanas a economia é mesmo aquela que selecciona a racionalidade de modo a demarcar o seu objecto. É pelo facto de essa suposição, que ancora a teorização das escolhas dos agentes não ser empiricamente confirmável, que algumas posições na filosofia da economia a consideram como um ramo da matemática aplicada e não como uma ciência empírica produzindo generalizações perfectíveis.

Que áreas da filosofia contemporânea levam mais em conta as teorias formais da racionalidade?

Nova constatação. Embora muitas das actuais teorias da racionalidade sejam formalizadas e técnicas e totalmente independentes da filosofia há áreas da filosofia contemporânea em que se choca de frente com a questão da racionalidade. Muitos dos filósofos que trabalham nessas áreas partem da teorização que é feita noutras disciplinas a partir de teorias formais da racionalidade. É bastante claro a partir daquilo que já se afirmou que a ética e a filosofia política são duas dessas áreas, já que em ambas se trata inevitavelmente de agentes, das suas crenças e das suas estruturas de preferências. No caso da ética a questão da racionalidade coloca-se devido às dificuldades com que se debatem as éticas da maximização racional por excelência, i.e. a deontologia (as éticas de raíz kantiana) e o utilitarismo. No caso da filosofia política fundamental, não é difícil notar que a grande revolução da filosofia política contemporânea, que aconteceu com a *Teoria da Justiça* de J. Rawls [41] se deveu precisamente à decisão de tratar a questão da justiça como uma questão de racionalidade. Precisamente, Rawls faz depender de uma escolha racional a instituição social da justiça: «o conceito de racionalidade aqui evocado é (....) aquele que é comum nas ciências sociais. Assim, conforme é usual, um sujeito racional é visto como tendo um conjunto coerente de preferências, estabelecidas de entre as opções que se lhe oferecem. Hierarquiza essas opções de acordo com a sua aptidão para prosseguir os seus objectivos; prefere um plano que satisfaça um maior número dos seus desejos a um outro que o faça em menor quantidade, preferindo também aquele que

[41] RAWLS 1993

tenha maiores probablidades de ser executado com êxito» [42]. É assim racionalmente que «os sujeitos colocados na posição original tentam identificar os princípios que favoreçam tanto quanto possível os seus sistemas de objectivos» [43]. Como afirma Rawls, «A questão da justificação é resolvida pela solução a dar a um problema de deliberação: temos de identificar quais os princípios que seria racional adoptar na situação contratual dada. Por aqui se liga a teoria da justiça à teoria da escolha racional» [44].

Uma terceira área em que as questões da racionalidade são incontornáveis é a filosofia da mente.

PARTE TRÊS

Sob que forma surgem as questões da racionalidade na filosofia da mente?

Várias teorias da mente de raiz quiniana apresentam-se como teorias da interpretação de sistemas físicos supondo a racionalidade (é o caso das teorias da mente de Donald Davidson e de Daniel Dennett, por exemplo). Nessa teorias (a interpretação radical de Davidson, a estratégia intencional de Dennett, além da tradução radical do próprio Quine) começa-se por pressupôr que as pessoas pensam e agem racionalmente e depois vê-se o que se pode fazer com essa suposição (o que é que ela pode explicar, como é que ela pode ser justificada, se ela pode realmente ser justificada, o que é que se perde se ela não se sustentar). Ora, como a filosofia da mente é ou deve ser filosofia da ciência e nomeadamente filosofia da psicologia ela vê-se obrigada a considerar o estatuto desse apelo à racionalidade para falar de acontecimentos em sistemas físicos. Ao contrário do que se passa com teorias declaradamente abstractas e formais como a teoria da decisão que podem sempre remeter para depois, para a psicologia, seja lá como for que esta se passa, a implementação de processos como crenças, desejos e escolhas a que fazem apelo, a incumbência da filosofia da mente é oferecer uma explicação daquilo que são

[42] RAWLS 1993:125
[43] RAWLS 1993: 126
[44] RAWLS 1993: 37

crenças, desejos, escolhas, no seio de uma metafísica materialista.

Diga-se de passagem que a circularidade encontrada por Amartya Sen na descrição da decisão racional em termos de preferências e escolhas é o pão nosso de cada dia dos filósofos da mente que se resignam por isso a falar do mental em termos de crenças e desejos não procurando explicar as crenças pelos desejos nem os desejos pelas crenças, mas tomando a racionalidade do agente de forma 'holista' (é o caso da tradução radical, da interpretação radical e da estratégia intencional).

Que forma apresentam os estudos empíricos da irracionalidade e que problema representam esses estudos para as disciplinas que supõem a racionalidade dos agentes? O que se entende por irracionalidade?

Curiosamente, aquele que é talvez o maior problema para os filósofos da mente que fazem apelo à racionalidade é o (aparentemente) comprovado irrealismo desse apelo. Existem estudos empíricos, realizados por psicólogos – provavelmente os mais conhecidos serão os estudos acerca de heurísticas, tendências prévias (*bias*) e enquadramento de escolhas (*frame of choices*) ligados ao nome do psicólogo Amos Tversky da Universidade de Stanford-California [45] – que podem ser interpretados como provando que as pessoas são basicamente, sistematicamente e maioritariamente irracionais nas suas decisões e preferências. No mínimo tais conclusões representam um problema para as teorias de acordo com as quais a teoria do comportamento de agentes está dependente de uma suposição de racionalidade, pois esvaziam de conteúdo empírico tal suposição.

As polémicas em torno da interpretação dos resultados psicológicos relativos a irracionalidade têm sido acesas. Não há nenhuma certeza quanto ao que está envolvido na suposta possibilidade de provar empiricamente que a maior parte das pessoas agem e pensam irracionalmente na maior parte das vezes. O que está em causa em tais polémicas é em grande medida o estatuto empírico da teoria da decisão, que é ao mesmo tempo a visão

[45] Cf. por exemplo KAHNEMAN, SLOVIC & TVERSKY 1982.

dominante nos estudos acerca da racionalidade e uma idealização aparentemente vazia de conteúdo empírico ou mesmo infirmada pelos factos acerca das escolhas reais de agentes

Avançando desde já com a posição dos autores que não consideram que os resultados psicológicos infirmem as hipóteses ou idealizações da teoria da escolha racional acerca dos agentes racionais, é possível defender – é essa a posição de D. Davidson [46] – que sem o contexto teórico representado pela teoria da decisão racional, sem os constrangimentos que a teoria da decisão racional explicita acerca de crenças e preferências, não sabemos sequer dizer o que é racionalidade. Assim, não faz sentido pensar que se prova que as pessoas são maioritariamente irracionais. Chamar a alguém ou a algum comportamento irracional supõe saber ou presumir saber o que é racionalidade e a melhor teoria desenvolvida nesse sentido é a teoria da decisão racional, que permite conceber em que consiste o facto de um agente escolher por entre acções alternativas para as quais existem razões, em função de preferências e de crenças.

De acordo com a teoria, um agente racional tem um padrão coerente de preferências, que obedecem a certos requisitos, como a transitividade, e a escolha racional consiste na selecção da alternativa que de acordo com o entendimento do agente produzirá os resultados mais valorizados. O agente racional agirá sempre de modo a que a alternativa escolhida seja aquela que não é superada por outra em termos de utilidade esperada para o agente. D. Dennett, outro teórico quiniano da interpretação [47], embora não evoque a teoria da decisão como explicitação da racionalidade suposta na estratégia intencional, considera que sem constrangimentos idealizantes quanto ao que constitui a racionalidade não

[46] DAVIDSON 1980. Cf comentário em ZILHÃO 1998/1999. D. Davidson passou alguns anos a fazer experiências desenhadas no sentido de confirmar ou infirmar a hipótese segundo a qual as pessoas agem racionalmente no sentido da teoria da decisão. Como diz Davidson (DAVIDSON 1980: 270): «As time went on, I became more and more skeptical about what the experiments showed. Of course they showed something» (...) (DAVIDSON 1980: 272) «they can be taken, if we want, as testing whether decision theory is true. But it is at least as plausible to take them as testing how good one or another criterion of preference is, on the assumption that decision theory is true».

[47] DENNETT 1987

temos sequer o instrumento que permite a interpretação de sistemas físicos como intencionais e mentais.

Quanto ao próprio Quine, a tradução radical simplesmente não é possível a não ser mantendo certas suposições de comunidade entre intérprete e interpretado, nomeadamente quanto aos conectivos lógicos. A ideia de tradução afasta assim a possibilidade de racionalidades ou esquemas conceptuais radicalmente diferentes. No artigo *On the very idea of a conceptual scheme*[48] Davidson afirma algo de semelhante: (sendo aquilo que somos) não poderíamos conceber esquemas conceptuais radicalmente diferentes do nosso nem podemos fazer sentido de um falhanço total de tradução (entre línguas, entre esquemas). Que a tradução não fosse possível seria uma condição necessária para a existência de esquemas conceptuais radicalmente diferentes (a que é costume chamar incomensuráveis) e é a existência de tais esquemas conceptuais que as teorias da mente que são teorias da interpretação se vêem conduzidas a negar.

O que está aqui em causa é uma inevitabilidade da racionalidade, o facto de, como diria Davidson[49], a caridade não ser uma opção. Dada esta 'inevitabilidade' da racionalidade, esta inevitabilidade da caridade[50], explicite-se o que se entende por irracionalidade de um agente, com uma certa independência relativamente à teoria dominante i.e, a teoria da decisão racional. A irracionalidade será a característica de (1) processos pelos quais se chega a conclusões que não podem ser justificadas pelo conhecimento do agente, (2) processos que conduzem a uma conclusão ou uma decisão que não é a melhor que poderia ter sido alcançada à luz da evidência disponível e com os recursos temporais disponíveis e também a característica dos resultados de tais processos. Irracionalidades típicas dos animais racionais humanos são por exemplo acreditar em contradições, não acreditar nas consequências daquilo em que se acredita, fazer escolhas e professar crenças baseadas em más estimativas de probabilidades,

---

[48] DAVIDSON 1984

[49] DAVIDSON 1984: 197

[50] No sentido davidsoniano de interpretar o outro de forma a optimizar a concordância (*agreement*), e portanto tomar o outro como o mais racional possível, tendo crenças na sua maioria verdadeiras, etc. Os princípios da tradução radical e da estratégia intencional são igualmente caridosos.

fraqueza da vontade (*akrasia*) (não se fazer fazer aquilo que ao mesmo tempo se acredita que se deve fazer), auto-engano (sinceramente não acreditar naquilo que ao mesmo tempo se sabe que se deve acreditar), etc. Em todos estes casos se trata obviamente de uma irracionalidade dos meios e não dos fins, ou pelo menos de uma irracionalidade na gestão dos meios: quando se afirma que existe irracionalidade nestes casos não se considera qualquer pretensão quanto a finalidades e valorações. O que está em causa é a análise do instrumento que nos permite decidir inclusive quanto a finalidades e valorações, a racionalidade atrás chamada instrumental.

A forma dos estudos empíricos da irracionalidade

Os estudos empíricos realizados por psicólogos [51] acerca de irracionalidades confirmariam que a hipótese ou suposição de acordo com a qual as pessoas são racionais é falsa. Dá-se em seguida alguns exemplos muito resumidos da forma que tais estudos assumem.

I. Erros de disponibilidade (*availability errors*)

a. Em geral, as pessoas não levam em conta os factos e as estimativas de probabilidades: embora tenha sido calculado que o risco de um nadador ser agarrado por um tubarão é muito menor do que o risco de morrer num acidente rodoviário no caminho para a costa, depois do filme *Tubarão* ter sido exibido, o número de nadadores na costa da Califórnia diminuiu extraordinariamente [52].

b. Numa experiência psicológica, os sujeitos lêem descrições de pessoas. As pessoas são descritas através de seis adjectivos [53]. Pede-se aos sujeitos que avaliem a pessoa que foi descrita. As avaliações resultantes das seguintes duas séries seguintes de adjectivos são completamente diferentes:

A - Inteligente, trabalhador, impulsivo, crítico, teimoso, invejoso

B - Invejoso, teimoso, crítico, impulsivo, trabalhador e inteligente

(Não será necessário dizer que a avaliação resultante da primeira série é favorável e da segunda desfavorável numa grande

[51] Cf. por exemplo TVERSKY & KAHNEMAN 1981, TVERSKY & KAHNEMAN 1993, SHAFIR & TVERSKY 1995, SUTHERLAND 1992.

[52] SUTHERLAND 1992: 15

[53] SUTHERLAND 1992: 25

maioria dos casos, embora no segundo caso os adjectivos sejam exactamente os mesmos, apresentados por ordem inversa)

c. É sabido que a maioria das pessoas sobrestima muito a probabilidade de morrer por exemplo num acidente aéreo: de cada vez que acontece um acidente aéreo tudo o que lhe diz respeito está muito disponível, i.e. as pessoas recebem uma grande quantidade de informação sobre o caso.
Do mesmo modo, depois de cada tremor de terra na Califórnia o número de seguros contra tremores de terra realizados aumenta bruscamente, descendo gradualmente de novo até ao próximo tremor de terra.

d. Os médicos que viram recentemente um grande número de casos de doença x têm uma maior tendência para diagnosticar essa doença.

e. Os corretores (pelo menos alguns) têm tendência para aconselhar os clientes a comprar acções quando o valor destas sobe e a vender quando o valor desce, quando a estratégia mais racional seria a inversa: vender nos picos e comprar nas descidas [54].

A casos deste género os psicólogos chamam 'erros de disponibilidade' (*availability errors*): os agentes julgam apoiados no conhecimento mais prontamente disponível. Tversky e Kahneman [55] relacionam este tipo de 'inclinação' (*bias*) com uma natural, muito difundida e má teoria da amostragem devida precisamente a esta heurística da disponibilidade. A disponibilidade substitui-se à consideração de probabilidades prévias, do tamanho da amostra, etc.

II. *Sunk costs*

Há custos a que os economistas chamam *sunk costs*, custos 'perdidos ou afundados'. Na economia é defendida uma teoria acerca da decisão racional de acordo com a qual apenas o presente e as consequências futuras devem ser considerados numa decisão. Os custos passados (nomeadamente os investimentos e despesas feitas num outro curso de acção) são passados. A regra racional para a maximização dos lucros monetários é portanto que *sunk costs* são coisa do passado: o que importa são benefícios futuros.

---

[54] SUTHERLAND 1992: 23-24

[55] TVERSKY& KAHNEMAN 1993.

No entanto a maior parte das pessoas tem um grande tendência para incorrer nas chamas falácias de custo perdido, continuando por exemplo a prolongar um dado curso de acção, quando assim certamente perderão mais do que simplesmente desistindo e escolhendo racionalmente um curso alternativo. Por exemplo, as pessoas geralmente vêm uma peça de teatro que as aborrece mortalmente até ao fim porque pagaram o bilhete, generais de um exército atacante continuam a deixar morrer os seus homens numa batalha mesmo quando o ataque provoca muito mais baixas no atacante do que no inimigo, e os poderes políticos têm uma enorme relutância em abandonar um projecto público que comprovadamente será mau quando já foram investidos muitos fundos e preferem desperdiçar ainda muito mais dinheiro continuando o projecto.

É evidente que a doutrina dos *sunk costs* é uma doutrina do âmbito da maximização do lucro monetário, logo não é necessriamente um bom princípio de decisão em geral. Como nota R. Nozick não é assim que agentes morais racionais tratam compromissos passados com outras pessoas ou com os seus próprios projectos [56].

III. Intransitividade de preferências, inversão das preferências e enquadramento das decisões (*frame of choice*)

---

[56] É certo que 'falácias dos custos perdidos' podem ser interpretadas como salvaguarda de princípios. Aliás um dos interesses de R. Nozick em *The Nature of Rationality* (NOZICK 1993: 3, How to Do Things With Principles) é apontar as vantagens da adopção de princípios por agentes, vantagens consideradas de um ponto de vista ele próprio 'instrumental' (mas que é constitutivo do agente que se rege por princípios de acção). Algumas dessas vantagens serão a definição da identidade do agente, a integração da sua vida ao longo do tempo, impedir que seus futuros abandonem projectos iniciados pelo eu actual, poupar esforço de decisão e tempo de cálculo em criaturas de racionalidade limitada, definir limites (*draw the line*) entre o que se faz e o que não se faz sem ter que se pensar nisso de cada vez. De acordo com Nozick, os princípios categorizam as acções em grupos fazendo com que a 'utilidade simbólica' associada a uma opção de acção não se restrinja ao caso particular. A 'utilidade simbólica ' é a proposta de alargamento que Nozick faz relativamente aos princípios da teoria decisão racional que envolvem utilidades. A regra geral da racionalidade será então maximizar a utilidade simbólica. Voltando aos *sunk costs*, a hipótese geral de Nozick é que a tendência a honrar os *sunk costs* que em geral se encontra nas pessoas lhes permite por exemplo acções regidas por princípios, os quais têm, como se viu, várias funções pessoais e interpessoais. Assim, Nozick sugere que a disposição poderia ter sido seleccionada evolutivamente (NOZICK 1993: 21-26).

«Problema (A1): Imagine que decidiu ir ver uma peça e que o preço do bilhete é $10. Quando está a entrar no teatro descobre que perdeu uma nota de $10. Ainda assim pagaria $10 pelo bilhete» [57] (88% das pessoas responde que sim, 12% que não)

«Problema (A2): Imagine que decidiu ir ver uma peça e que pagou $10 pelo bilhete. Quando está a entrar no teatro descobre que perdeu o bilhete. Pagaria $10 por outro bilhete?» [58] (46% das pessoas dizem que sim, 54% dizem que não)

Problema B1. «Imagine que os Estados Unidos se estão a preparar para a irrupção de uma doença asiática rara que se espera que mate 600 pessoas. Dois programas alternativos foram propostos para combater a doença. Assuma que as estimativas científicas exactas das consequências dos programas são as seguintes:

Se o programa A for adoptado, 200 pessoas serão salvas

Se o programa B for adoptado, há 1/3 de probabilidades de que 600 pessoas sejam salvas e 2/3 de probabilidades de que nenhuma pessoa seja salva» [59]

(72% dos sujeitos da experiência escolhem o programa A e 26% escolhem o programa B)

Problema B2 (a situação é a mesma)

«Se o programa C for adoptado 400 pessoas morrerão

Se o programa D for adoptado, há 1/3 de probabilidades de que ninguém morra e 2/3 de probabilidades de que 600 pessoas morram» [60]

(78% dos sujeitos escolhem o programa D e 22% escolhem o programa C)

As respostas dos (mesmos) sujeitos às duas formulações do problema A e do problema B são inconsistentes.

Caso real:

Nos Estados Unidos a certa altura passou a ser permitido o uso de cartões de crédito nas estações de serviço. No entanto, o uso do cartão de crédito envolvia o pagamento de uma taxa sobre

---

[57] TVERSKY&KAHNEMAN 1981: 457

[58] TVERSKY&KAHNEMAN 1981: 457

[59] TVERSKY &KAHNEMAN 1981: 453

[60] TVERSKY&KAHNEMAN 1981: 453

a transacção. As pessoas boicotaram as estações de serviço. A situação foi reformulada: as estações de serviço passaram a anunciar a (mesma) diferença de preço do combustível como um desconto associado ao pagamento em dinheiro. As pessoas deixaram de fazer boicote e passaram a utilizar sem problemas o cartão de crédito [61].

IV. Não consideração da dimensão de amostras, más concepções do acaso, más estimativas de probabilidades [62]

Dá-se apenas um exemplo de má estimativa de probabilidades, a chamada falácia da conjunção. De acordo com a teoria das probabilidades a probabilidade de A & B é menor do que a probabilidade de cada um dos acontecimentos (A e B) isoladamente. No entanto, mesmo que saibam isso não é de acordo com esse princípio que os sujeitos julgam em situações práticas.

O exemplo que se segue é um dos mais discutidos na literatura. Aos sujeitos é dada a descrição de uma pessoa ficcional, por exemplo Linda, 31 anos, solteira, interventiva (*outspoken*), muito inteligente. A área principal da sua licenciatura (*major*) foi a filosofia. Como estudante preocupava-se com questões de discriminação, justiça social e participava em manifestações antinuclear. Pede-se aos sujeitos a quem é dada a conhecer a descrição que listem por ordem de probabilidade as afirmações A1, A2, A3.... An. As afirmações incluem 'Linda trabalha num banco' e 'Linda trabalha num banco e é activista do movimento feminista'. Mais de 80% dos sujeitos listam a segunda afirmação como mais provável do que a primeira [63].

Porque é que os resultados de experiências como aquelas que se referiu provariam a irracionalidade das pessoas?

Retomando por exemplo os estudos do enquadramento das decisões (*framing of choice*) os resultados obtidos são curiosos na

[61] Exemplo em TVERSKY & KAHNEMAN 1981: 456, discutido por exemplo em SCHICK 1997: 50.

[62] Cf. TVERSKY & KAHNEMAN 1993 para a discussão dos estudos, incluindo este exemplo e outros análogos.

[63] De facto os estudos nos quais este exemplo grosseiramente descrito se enquadra são sobre conjunções representativas de traços ou características e a sua base teórica é a ideia de que a informação é armazenada e processada através de protótipos. O facto de a informação ser processada por meio de protótipos (princípio psicológico de representatividade) ergue-se contra a lógica (extensional) das probabilidades.

medida em que são aparentemente incompatíveis com a caracterização do agente racional no âmbito da teoria da decisão.

Basicamente, aquilo que Tversky e Kahneman fazem nas suas experiências com enquadramento das decisões, das quais foram acima dados dois exemplos, é formular o mesmo problema de maneiras diferentes e submetê-lo as versões a teste, se possível com os mesmos sujeitos. É a partir de resultados como os acima referidos que Tversky e Kahneman concluem que «os princípios psicológicos que governam a percepção de problemas de decisão e a avaliação de probabilidades e resultados produzem desvios previsíveis de preferências quando o mesmo problema é enquadrado de diferentes maneiras» [64]. Em suma, aquilo que as pessoas querem e escolhem pode ser revertido pela diferença de descrição dos mesmos factos.

É obviamente possível defender que as opções colocadas às pessoas são diferentes, que as situações não são de factos as mesmas, que as utilidades são atribuídas de forma diferente aos resultados das escolhas. É apenas porque Tversky e Kahneman consideram que uma pessoa será racional se e só se as suas escolhas não forem sensíveis às descrições das situações (eles chamam a isto o princípio da invariância ou a extensionalidade [65]) que concluem que a maioria das pessoas não é racional.

Admitindo que a definição de racionalidade é muito debatida é ainda assim difícil negar que escolhas racionais devam satisfazer requisitos de consistência e coerência. No entanto, o que os estudos psicológicos atestam é que as pessoas sistematicamente violam esses requisitos.

O que significa afirmar que estudos deste género não provam que as pessoas são irracionais? Será razoável defender uma tal posição?

Como foi dito, Davidson considera que experiências empíricas acerca de racionalidade não podem infirmar as caracterizações que a teoria da decisão faz da estrutura do agente racional. Davidson considera que não sabemos imaginar o que é para um agente ser racional (ou irracional) sem o quadro teórico da teoria da

---

[64] TVERSKY e KAHNEMAN 453

[65] Tversky e Kahneman consideram que há uma mesmidade objectiva nas situações descritas nas suas experiências e esse é obviamente um pressuposto contestável..

decisão. Mas será que Davidson tem razão, i.e. será que a caracterização do agente feita pela teoria da decisão é mesmo constitutiva daquilo que é ser racional e agir racionalmente? Em que sentido o seria? Que posição se obteria não aceitando que as propostas de Davidson são as melhores perante os resultados obtidos por Tversky e pelos seus colaboradores?

Há uma espécie de circularidade na posição de Davidson: sendo a racionalidade constitutiva do agente, ela é encontrada porque é 'lá posta'. Na terminologia de A.Zilhão [66] e forçando um pouco os termos do próprio Davidson no artigo *'Hempel on Explaining Action'* [67], isto significa que os axiomas da teoria da decisão são verdades sintéticas a priori acerca de seres racionais quaisquer. Essa é a razão pela qual Davidson pensa que a teoria da decisão racional tem que ser preservada (por exemplo as considerações acerca da transitividade das preferências do agente) face a resultados como os de Tversky e Kahneman [68] que mostram como é comum a inversão de preferências. A teoria tem que ser preservada porque é tudo o que temos, caracteriza tudo o que somos tanto quanto chegamos a ser racionais.

Será legítimo pensar que se confirma ou infirma empiricamente a racionalidade?

Davidson, que conhece os trabalhos de Tversky e que colaborou ele próprio em investigações experimentais acerca de racionalidade, pensa que não é legítimo pensar que se pode confirmar ou infirmar empiricamente a racionalidade e procura chamar a atenção para o seguinte: o que deve ser levado em conta quando se enfrenta o problema da adequação empírica de caracterizações idealizantes do agente racional é que nenhuma interpretação de factos experimentalmente obtidos pode ser feita sem utilizar as próprias propostas da teoria da decisão. Para além disso, nós próprios não podemos deixar de nos considerar como agentes racionais já que toda a *folk psychology* [69] se baseia nesse pressuposto. Pura e simplesmente não sabemos nem podemos

[66] ZILHÃO 1998/1999

[67] DAVIDSON 1980

[68] TVERSKY & KAHNEMAN 1981

[69] A *folk psychology* ou psicologia de senso comum é a natural atribuição de crenças e desejos que as pessoas fazem relativamente aos outros e a si próprias de modo a fazer sentido do comportamento.

pensar de outra maneira, não podemos pensar que nós próprios não pensamos, ou que qualquer outra pessoa não pensa, geralmente de forma racional (a não ser deixando de pensar que pensamos, por exemplo atribuindo erros a irrupções de distúrbios ao nível do *hardware* cognitivo). Para Davidson, a teoria da decisão é uma teoria tão poderosa e tão simples que as descobertas têm que se lhe adaptar: não se poderia fazer sentido da situação inversa (em que se procuraria elaborar uma teoria da racionalidade a partir de dados como os recolhidos pelas experiências sobre a racionalidade de agentes). Por essa razão, Davidson compara a teoria da decisão com a teoria tarskiana da verdade: «A teoria é em cada caso tão poderosa e simples e de tal modo constitutiva de conceitos assumidos por posteriores teorizações (...) que devemos tentar adaptar os nossos resultados, ou as nossas interpretações, de modo a preservar a teoria» [70].

Não há como negar que se está perante uma circularidade na tentativa de justificar ou fundamentar a racionalidade de agentes. Esta é de resto a situação mais frequente nas teorias da racionalidade de qualquer género.

Quais são as alternativas quando se considera a relação entre teorias abstractas da racionalidade como a teoria da decisão e experiências psicológicas acerca da racionalidade de agentes?

A. Zilhão [71] enumera as alternativas:

1. Existem critérios independentes daquilo em que consiste a racionalidade e a teoria da decisão é confirmada.

2. Existem critérios independentes e a teoria da decisão é infirmada (i. e. os seus axiomas não são aplicáveis ao comportamento dos agentes racionais reais).

3. Não existem critérios independentes daquilo em que consiste a racionalidade a não ser os próprios axiomas da teoria da decisão (esta é obviamente a posição de Davidson).

A. Zilhão subscreve a segunda alternativa e conclui por uma alternativa: ou os humanos não são racionais, uma grande parte das suas acções é simplesmente irracional ou os humanos são racionais mas a sua racionalidade não é a racionalidade caracteri-

[70] DAVIDSON 1980:273

[71] ZILHÃO 1998/1999

zada pela teoria da decisão. Tudo o que é empiricamente verificável e consensual é que as pessoas usam procedimentos de decisão sub-óptimos nas situações particulares com que são confrontadas. De qualquer modo Zilhão rejeita uma pressuposição geral de Davidson, a ideia segundo a qual a psicologia de senso comum é uma aproximação a uma realidade cuja verdadeira natureza seria capturada apenas pela teoria da decisão.

Seja como for, supostamente a racionalidade é a chave da psicologia, no sentido em que é mediante uma suposição de racionalidade que atribuímos crenças e desejos a outros seres e a nós próprios. Dever-se-á deixar de pensar que a racionalidade é a chave da psicologia? Dever-se-á defender que a racionalidade não tem verdadeira realidade nenhuma, uma vez que a melhor hipótese de captura da verdadeira natureza da racionalidade era a teoria da decisão? Mas se a racionalidade não é a chave da psicologia (estando a racionalidade para o domínio psicológico como a adaptação está para o domínio biológico), o que é a psicologia[72]*?*

Então a racionalidade não é a chave da psicologia (humana ou outra)... Pode ser que a psicologia não seja nada de especial, pode ser que a psicologia não tenha pretensões legítimas a ser uma ciência (diferente da neurofisiologia ou da teoria de qualquer outro *hardware* que seja suporte de mentalidade)[73]. Mas se a racionalidade não existe, não tem verdadeira natureza nenhuma, a psicologia humana não existe na realidade, de um ponto de vista de Deus, digamos, ou metafisicamente, fundamentalmente (a psicologia humana ou outra, porque o domínio do psicológico, das crenças e desejos cuja interacção guia o comportamento de agentes só se deixa abrir com a chave da racionalidade).

Como é que esta conclusão se aplica à estratégia intencional?

Que a psicologia só se abre com a chave da racionalidade é também a suposição subjacente à Teoria dos Sistemas Inten-

[72] Não se fala da psicologia como disciplina mas de uma realidade que seria psicológica, constituída por pensamentos, crenças, desejos, etc (interioridade mental intencional).

[73] Esta posição não é muito estranha nem inédita. Qualquer materialista eliminativo tende a defendê-la. Por exemplo R. Rorty, nas ocasiões em que se assume como filósofo da mente, fá-lo.

cionais (TSI) de Dennett. Como a filosofia de Davidson, também a teoria dos sistemas intencionais dá uma resposta circular à questão acerca da natureza e da justificação da racionalidade: encontra-se aquilo que se põe ou não se encontrará coisa nenhuma (o intérprete, que supõe a racionalidade de sistemas físicos é ele próprio supostamente racional). No entanto, ao contrário de Davidson, que procura um fundamento para a racionalidade 'em cima', i.e. procura estabelecer uma relação entre a suposição de racionalidade e a teoria formal da decisão, a TSI procura um fundamento 'em baixo', no *design* dos sistemas cognitivos: a suposição de racionalidade não é absolutamente infundada nem apenas instrumental, caso em que suporia provisoriamente algo que de facto não existe, na medida em que a racionalidade é um resultado da evolução de *design* por selecção natural. O nível intencional dos sistemas está portanto, garantido pelo *design*. Agentes como os humanos estão *desenhados para* acreditar maioritariamente verdades, procurar a satisfação de desejos e deliberar racionalmente de modo a alcançar esses dois objectivos. Se o *design* para a racionalidade de um agente resulta de evolução por selecção natural e se toda a evolução é um processo de *satisficing* esse desenho será imperfeito. Assim, não haverá nada de especialmente 'tocante' (nomeadamente tocante na 'natureza íntima' da realidade) nos adquiridos de tal racionalidade. Qualquer princípio de racionalidade deve ser relativizado a uma particular arquitectura mental suficientemente boa mas não certamente perfeita, não perfeitamente penetrante na estrutura da realidade.

É certo que mesmo procurando um fundamento 'em baixo' a TSI encontra um outro problema de circularidade. A adaptação é aquilo que ao nível do *design*, das formas funcionais, corresponde à racionalidade ao nível psicológico a garante. A racionalidade poderá mesmo ser considerada como *design* cognitivo óptimo. No entanto, por trivial que pareça a observação, a racionalidade não é uma característica do funcionamento neurofisiológico dos neurónios ou outro *hardware* qualquer: o seu estatuto é o artificial de H. Simon, o interface, a adaptação, e esse artificial supõe outro artificial, um ponto de vista psicológico, um intérprete. Discernir a adaptação supõe a racionalidade no sentido de um pensamento, um ponto de vista, que possibilita uma abordagem do mundo físico atribuindo finalidades a sistemas.

(De novo) De um ponto de vista evolucionista o que se deve pensar acerca da racionalidade e do seu estatuto?

Não é apenas a TSI que reporta a racionalidade à evolução por selecção natural de sistemas cognitivos físicos que agem de forma adaptada ao seu ambiente. Essa é uma suposição comum nas teorias psicológicas e filosóficas da natureza da racionalidade. A racionalidade é uma adaptação evolutiva, com um propósito determinado, e o *design* da racionalidade, resultante de evolução por selecção natural não é óptimo mas apenas suficientemente bom. Ora, como Nozick sublinha [74] se deixar nas mãos de regras a produção de inferências cumpre uma função biológica interessante para os seres em quem tal funcionamento está instalado, não convem esquecer que as regras são seleccionadas pelo êxito obtido na acção pelos agentes que as seguem e que tal êxito não depende da perfeição das regras e sim do facto de elas conseguirem o máximo possível na negocia-ção (*trade-off*) entre as capacidades cognitivas do agentes (limitadas), o tempo de resposta e a quantidade de informação obtida.

Consideradas todas as (supostas) 'infirmações empíricas' da racionalidade dos agentes como se poderá abordar o estatuto da racionalidade sem subscrever a posição de Davidson? Em que situação ficamos se aceitarmos que a caracterização do agente racional feita pela teoria da decisão é mesmo psicologicamente inadequada e não há substituto? Ficamos sem psicologia? (Talvez não, porque psicologia é aquilo que somos)

F. Broncano [75] sugere que a fiabilidade do desenho desenvolvido nas condições evolutivas referidas, o manteria e o garantiria no presente: «mentes mal desenhadas (...) mas com plasticidade suficiente para se auto-corrigirem podem gerar pontos de equilibrio no controlo da informação, que identificamos precisamente com a racionalidade» [76]. Com esta ideia Broncano pretende retomar uma ideia por vezes utilizada para a justificação da racionalidade: o equilibrio reflexívo rawlsiano.

Algo como um equilíbrio reflexivo à maneira de Rawls seria a única justificação possível da racionalidade. A ideia da aplicação

---

[74] NOZICK 1993, Capítulo 4, Evolutionary Reasons.

[75] BRONCANO 1995

[76] BRONCANO 1995:327

do equilíbrio reflexivo rawlsiano à racionalidade produz o seguinte: aceita-se uma dada norma de inferência porque ela produz inferências que consideramos intuitivamente válidas e por outro lado considera-se válidas as inferências que sejam produto das regras que tiverem sido aceites. Um equilíbrio não definitivo entre intuição e regras é a única alternativa de justificação da racionalidade numa situação em que existem mentes 'mal desenhadas' e não existem fundamentos reais.

Evidentemente nada garante que as caracterizações que vão resultando do equilíbrio entre intuição e regras sejam universalizáveis. S. Stich [77], outro filósofo da racionalidade, poderia aqui evocar o problema de Nisbett (R. Nisbett é mais um psicólogo que estuda as irracionalidades do raciocínio). Quando se apresentam resultados de investigações psicológica da irracionalidade a uma audiência, ouvir-se-á questões como: 'como é que sabe a sua maneira de atribuir probabilidades é correcta (para poder dizer que a dos sujeitos é incorreta)'? 'O quê exactamente faz de uma inferência uma inferência válida?' 'Que direito temos de afirmar que os indivíduos estão enganados, estão a cometer irracionalidades?'.

Este tipo de problemas coloca em dúvida o equilíbrio reflexivo, pois põe em causa que os pontos de equilíbrio sejam aceites em geral. Stich [78] aliás especifica ainda melhor o problema que o equilíbrio reflexivo rawlsiano aplicado ao problema da racionalidade coloca: a intuição poderá ser um mecanismo universal de avaliação somente se for independente da informação que considera e valoriza. Se pelo contrário e como parece acontecer a intuição estiver conformada por exemplares paradigmáticos de inferência e se aplicar tomando estes casos como amostra a partir da qual generaliza de modo a que as próprias regras têm como referência tais protótipos, a intuição deixa de ser um mecanismo universal sendo antes dependente da história cognitiva dos sujeitos. Apoiando a suspeita de Stich quanto à intuição a que um equilíbrio reflexivo de género rawlsiano faria inevitavelmente apelo, se há coisa que os estudos psicológicos mostram mesmo é que os sujeitos raciocinam mediante protótipos e heurísticas (é esse o

[77] STICH 1990
[78] STICH 1990

sentido de procedimentos sub-óptimos). A conclusão de Stich em *The Fragmentation of Reason*[79] é então que não existe qualquer possibilidade de formular constrangimentos a priori para todos os agentes racionais possíveis: não existe A Racionalidade, o Agente Racional.

O que se conclui de uma situação em que há teorias que supõem a racionalidade (todas as disciplinas intencionais, i.e. as chamadas ciências sociais e humanas) e outras que analisam empiricamente os funcionamentos da racionalidade e que chegam a conclusões 'tristes'?

Os estudos empíricos da racionalidade mostram que, ao contrário do que é assumido em teorias por exemplo económicas e filosóficas da racionalidade, longe de serem perfeitamente racionais (ou tendencialmente racionais) as pessoas não são sequer consistentemente racionais nas suas crenças e preferências e portanto na maior parte dos seus pensamentos e acções. Mas os estudos não podem mostrar também que as pessoas são irracionais na grande maioria dos casos ou sempre, pois sem racionalidade nenhumas entidades são concebíveis como irracionais ou como agindo ou sequer como pessoas, como seres mentais. Há uma racionalidade (mesmo que talvez mínima) definidora do próprio conceito de agência, definidora daquilo que se entende por agente, por intencionalidade, por psicologia. Não há como negar que essa racionalidade é uma idea-lização, uma idealização de alguma forma reportada aos funcionamentos imperfeitos que são os funcionamentos reais. Se o núcleo mínimo de racionalidade é a racionalidade instrumental esta dá-se tão bem com o conceito de *satisficing* como com o conceito de optimização. Evidentemente o que acontece é que nenhuma qualidade intrínseca dos produtos e processos de uma racionalidade tal fica garantida.

No seu livro *The Nature of Rationality* R. Nozick faz uma pergunta curiosa[80]: se Quine acha que a lógica (tomada aqui como o expoente da análise da racionalidade) é contínua com, ou faz parte da, investigação científica em ciências empíricas, por que

[79] STICH 1990

[80] NOZICK 1993: 110

razão não se continua para uma avaliação mais profunda dos princípios da lógica, até se alcançar uma explicação da razão por que os princípios da lógica se sustentam, exactamente como se continua com a investigação em física? Já se teria chegado às verdades fundamentais definitivas na lógica ao contrário do que acontece na ciência empírica? É bastante implausível pretendê-lo. A razão será talvez outra: nós não somos, enquanto pensantes, outra coisa que não isso, racionalidade. Essa não é, como se pretendeu fazer ver, grande garantia de coisa nenhuma se a racionalidade na melhor das hipóteses consiste em pontos de equilíbrio em mentes que são, tanto quanto sabemos, mal (ou apenas suficientemente bem) desenhadas.

Sofia Miguens

BIBLIOGRAFIA

ARROW, Kenneth, 1963, *Social Choice and Individual Values*, New York, John Wiley and Sons

AXELROD, R., 1984, *The Evolution of Cooperation*, New York, Basic Books

BRONCANO, Fernando, 1995, El control racional de la conducta, in F. Broncano (ed), 1995, *Enciclopedia Iberoamericana de Filosofia* – La Mente Humana, Madrid, Trotta

CHERNIAK, Christopher, 1994, «Rationality», in Guttenplan 1994 (ed), *A Companion to the Philosophy of Mind*, Oxford, Blackwell

DAVIDSON, Donald, 1980, Hempel on Explaining Action, in Davidson, D.,1980, *Essays on Actions and Events*, Oxford, Oxford University Press

DAVIDSON, Donald, 1984, On the very idea of a conceptual scheme, in Davidson, D.,1980, *Essays on Truth and Interpretation*, Oxford, Oxford University Press

DENNETT, Daniel, 1987, *The Intentional Stance*, Cambridge MA, MIT Press

DUPUY, Jean-Pierre & LIVET, Pierre, 1997, *Les limites de la rationalité – Rationalité, éthique et cognition*, Paris, La Découverte

HARDIN, Garrett, 1968, *The tragedy of the commons*, Science 162

HARMAN Gilbert, 1995, Rationality, in SMITH &OSHERSON (eds) 1995

HUME, David 1985 [1739], *A Treatise of Human Nature*, London, Penguin

JEFFREY, Richard, 1983, *The Logic of decision*, Chicago, Chicago University Press

JEFFREY, Richard, 1995, «Choice Theory», in R. Audi (ed), Cambridge Dictionary of Philosophy, Cambridge, CUP, 1995

KAHNEMAN, D., SLOVIC, P. & TVERSKY, A. (eds), 1982, *Judgment under uncertainty: Heuristics and Biases*, Cambridge, Cambridge University Press

MAYNARD SMITH, J., 1982, *Evolution and the Theory of Games*, Cambridge, Cambridge University Press

NOZICK, Robert, 1993, *The Nature of Rationality*, Princeton, Princeton University Press

RAMSEY, F., 1926, Truth and probability, in H. Mellor (ed), 1990, *Philosophical Papers*, Cambridge, Cambridge University Press

RAWLS, J., 1993 [1971] *Uma Teoria da Justiça*, Lisboa, Presença

SCHICK, Frederic, 1997, *Making Choices – A Recasting of Decision Theory*, Cambridge, Cambridge University Press

SEN, Amartya, 1977, Rational Fools, A Critique of the Behavioral Foundations of Economic Theory, *Philosophy and Public Affairs*, 6, 4

SEN, Amartya, 1991, *Éthique et Économie*, Paris, PUF

SEN, Amartya & WILLIAMS, Bernard, 1982, Introduction: *Utilitarianism and Beyond*, in SEN & WILLIAMS, 1982, Utilitarianism and Beyond, Cambridge, Cambridge University Press

SHAFIR, E. & TVERSKY, A., 1995, Decision Making in SMITH & OSHERSON (eds) 1995

SIMON, Herbert, 1998 [1969] *The Sciences of the Artificial*, Cambridge MA, MIT Press

SMITH & OSHERSON (eds), 1995, *Thinking - An Invitation to Cognitive Science*, Vol. II, Cambridge MA, MIT Press

SUTHERLAND, Stuart, 1992, *Irrationality*, New Brunswick, New Jersey, Rutgers University Press

TVERSKY, A. & KAHNEMAN, D., 1981, The Framing of decisions and the psychology of choice, *Science* 211, 453-458

TVERSKY, A. & KAHNEMAN, D., 1993, Probabilistic Reasoning, in A. Goldman (ed), *Readings in Philosophy and Cognitive Science*, Cambridge MA, MIT Press

SHAFIR &TVERSKY 1995, Decision Making in SMITH &OSHERSON (eds), 1995

Von NEUMANN, J. & MORGENSTERN, O. 1944, *Theory of Games and Economic Behavior*, New York, Wiley

FRANK YATES, J. & ESTIN, P, 1998, Decision-making, in *A Companion to Cognitive Science*, Oxford, Blackwell

ZILHÃO, António, 1998/1999, Folk-psychology, Rationality and Human Action, *Grazer Philosophische Studien*, 56