A CONSTRUÇÃO DA CONFIANÇA

–

Teoria dos Jogos e Ética

André Barata

2008

www.lusosofia.net

LusoSofia: press

FICHA TÉCNICA
TÍtulo: *A Construção da Confiança. Teoria dos Jogos e Ética*
Autor: André Barata
Colecção: Artigos **LusoSofia**
Direcção da Colecção: José Rosa & Artur Morão
Design da Capa: António Rodrigues Tomé
Logótipo: Catarina Moura
Paginação: José Rosa
Universidade da Beira Interior
Covilhã, 2008

A Construção da Confiança
–
Teoria dos Jogos e Ética

André Barata
Universidade da Beira Interior

Índice

A teoria dos jogos é um capítulo da matemática aplicada consistindo num estudo formal de interacções entre dois ou mais agentes racionais que se comportam estrategicamente.

Uma maneira mais condensada de apresentar a teoria dos jogos consistirá em dizer que a teoria dos jogos tem por objecto de estudo a *decisão social*. Isto, entendendo-se por "decisão social" a decisão que envolve, além da posição do agente decisor, a consideração da posição dos outros agentes que com ele estejam em interacção.[1]

[1] Note-se que, apesar de tratar de certo campo de decisões, não é habitual considerar-se a Teoria dos Jogos como um ramo da Teoria da Decisão, mas duas disciplinas consideravelmente autónomas, ainda que contíguas.

Três conceitos nesta definição devem ser elucidados: *interacção*, *comportamento estratégico* e *racionalidade*. Por interacção entende-se as acções de cada agente terem efeito nas dos outros agentes. E por comportamento estratégico entende-se a consideração racional, por parte de cada agente, das condições de interacção com os restantes agentes. A racionalidade dos agentes, por fim, pode ser pensada ou sob a ideia de maximização do interesse próprio ou sob a ideia de maximização de objectivos.[2]

Veremos adiante que uma dificuldade central nas aproximações entre pensamento estratégico e pensamento moral passa justamente pelo que se deva entender por racionalidade.

A teoria dos jogos, tal qual a reconhecemos hoje, aparece em 1944 com *Theory of Games and Economic Behaviour* de John von Neumann e Oskar Morgenstern. Antes disso, houve naturalmente importantes teorizações que são, no entanto, entendidas como precursoras da moderna teoria dos jogos. Contam-se, entre estas teorizações percursoras, os trabalhos de Pascal (1623-1662), Fermat (1601-1665) e Huygens (1629-1695), ligados sobretudo aos jogos de azar, como o jogo do dado, e a problemas de probabilidade. Só bem mais tarde, com o matemático Emile Burel (1971-1956) surgem teorizados aspectos realmente ligados à decisão sob um pressuposto de interacção entre jogadores. Por Xemplo, foi Burel quem formulou o influente princípio *minimax* da minimização das perdas máximas.

Na moderna teoria dos jogos, Von Neumann (1903-1957), Mor-

[2] Robert Frank distingue, a propósito de uma definição de racionalidade, dois padrões de racionalidade prática – "Há duas abordagens distintas à definição de racionalidade. Uma delas considera o interesse próprio como única motivação; as pessoas racionais apenas atribuem um peso significativo aos custos e benefícios que lhes dizem directamente respeito. Esta abordagem põe explicitamente de lado motivações tais como tentar tornar os outros felizes, tentar fazer o seu dever, e por aí diante. Um conceito alternativo toma como possíveis quaisquer objectivos imediatos. O seu único requisito é o de que as pessoas actuem eficazmente para os atingir." (Frank, 1997: 18)

genstern, John Nash são referências centrais. Este último, popularizado pelo filme *A Beautiful Mind*, demonstrou a existência, desde que admitidas estratégias mistas, de pelo menos um ponto de equilíbrio para quaisquer jogos, cooperativos ou não-cooperativos, de soma zero ou variável, com dois ou mais jogadores. Antes disso, Von Neumann e Morgenstern apenas haviam conseguido generalizar o teorema minimax aos jogos cooperativos e de soma zero.[3]

A teoria dos jogos tornou-se particularmente importante em diversos domínios de aplicação, onde nos deparamos com agentes racionais em interacção. Desde as relações internacionais às relações entre agentes sociais, passando pela estratégia militar, a teoria dos jogos revelou-se uma forte ferramenta de análise e de produção de conhecimento. No campo da economia vários prémios Nobel foram atribuídos a investigadores que, de uma forma ou de outra, lidaram com esta teoria matemática no estudo do comportamento de agentes económicos. John Nash em 1994, Thomas Schelling em 2002 e David Kahneman em 2000 são três laureados com o Nobel da Economia que introduziremos na nossa discussão. Mesmo no âmbito da biologia, do estudo dos ecossistemas e da evolução natural, investigadores têm conseguido amplas aplicações da teoria dos jogos. A teoria do "gene egoísta" desenvolvida por Richard Dawkins assimila as condições concorrenciais na biologia às da economia. Resulta disto uma cooperação concorrencial-dependente[4]. Na biologia evolucionária, o *altruísmo recíproco* consiste num tipo de altruísmo em que um organismo concede um benefício a outro na expectativa de uma reciprocidade futura. Também Maynard-Smith aplicou a teoria dos jogos à biologia, designadamente ao comportamento estratégico de populações em contexto de evolução, tendo disso resultado o conceito de "estratégia evolucionária estável" *(Evolutionary stable strategy)*. Tal estratégia, sendo generalizada numa população,

[3] Entende-se por jogos de soma zero jogos em que um jogador só pode ganhar o que outro jogador perde – somando, por isso, ganhos e perdas zero – e por jogos cooperativos jogos em que são permitidos acordos entre jogadores.

[4] Cf. Dawkins, 1976.

revela-se vantajosa face a qualquer outra estratégia mutante presente com uma frequência suficientemente baixa[5]. Richard Dawkins, ilustra a ideia das estratégias evolucionárias estáveis com a parábola do falcão e da pomba. Esta parábola exprime a interacção entre comportamento agressivo e comportamento pacífico num contexto de competição por um recurso alimentar. Com efeito, verifica-se não ser evidente que seja preferível, estrategicamente falando, ser falcão a ser pomba. O comportamento pacífico não é *a priori* menos razoável que o comportamento agressivo, e isto mesmo num mundo estritamente competitivo e não-cooperativo. A parábola do falcão e da pomba explicita, pois, a natureza de *preferências estratégicas*, isto é, preferências que dependem do número de agentes que a partilham – se houver demasiados falcões será preferível ser pomba; se, pelo contrário, houver demasiadas pombas, então será preferível ser falcão.

É neste amplo quadro de aplicações da teoria dos jogos que resultam pertinentes aproximações ao problema do relacionamento ético entre agentes racionais. Inspirada no hobbesianismo político, uma dessas aproximações recebeu a denominação de "contratarianismo moral" (*moral contractarianism*). Sustenta, por um lado, que os indivíduos são primariamente motivados pelo interesse próprio e, por outro, que, em vista da maximização do interesse próprio, contratam uns com os outros normas morais que proporcionam resultados cooperativos melhores. David Gauthier é a principal referência contratarianista.[6] Convém distinguir esta corrente do *contratualismo moral*. Neste, a motivação para os indivíduos fazerem concessões uns aos outros contratando normas morais não é interessada. Pelo contrário, os indivíduos são motivados a agir moralmente pelo compromisso racional com a universalizabilidade da acção. Numa palavra, o contratarianismo está para Hobbes como o contratualismo moral está para Kant ou para Rawls.

Feita esta sucinta apresentação, propomo-nos, neste estudo, fazer um itinerário que culminará numa questão central para a filosofia mo-

[5] Cf. Maynard-Smith, 1974.
[6] Cf. Gauthier, 1986.

ral. Começaremos por dar conta das possibilidades que a teoria dos jogos oferece para uma explicação da construção social da confiança. Para isso, tomaremos como matéria da nossa atenção a discussão do célebre dilema do prisioneiro e das suas versões iteradas. Evidenciaremos, em seguida, que as estratégias que a teoria dos jogos tematiza no âmbito da decisão social encontram uma razoável correspondência em certos preceitos fortemente presentes nos mais importantes sistemas morais que a cultura da humanidade dispõe. Mau grado esta nítida aproximação, exporemos uma razão, a nosso ver de peso, contra uma redução da moral a estratégias de maximização da utilidade. Tal razão dá pelo nome de altruísmo sem contrapartidas, ou genuíno. Cremos que sem uma compreensão deste fenómeno o essencial do que denominamos *dever* fica por compreender. Teremos, no entanto, ocasião de verificar que não são apenas as nossas intuições sobre moralidade e ética que contrastam com a racionalidade de um *homo economicus*. São os próprios comportamentos económicos dos humanos que, em contraste com um princípio de acção assente exclusivamente numa racionalidade maximizadora, revelam uma racionalidade prática não instrumental, que visa antes a autonomia, semelhante, pois, àquela que Kant defendia, ainda que localizada e sempre sustentada em interacções sociais. Assim, é no quadro da discussão das aplicações da teoria dos jogos e da racionalidade prática aí pressuposta que julgamos poder encontrar as bases para a compreensão de uma emergência do dever, que caracterizaremos como agencial, dever para com a acção, considerada em si mesma e por si mesma. Por fim, procuraremos articular um tal dever agencial, que assinala o requisito da liberdade como autonomia, de herança manifestamente kantiana, com o não menos essencial requisito da alteridade, isto é, de que a moralidade seja algo que não pode deixar de ter que ver com os outros, em sua radical irredutibilidade, ou seja, heteronomia. Estes que são, pelo menos de acordo com as nossas mais básicas intuições acerca do comportamento moral, os dois aspectos essenciais da moralidade têm aparentado, porém, entre si, uma contradição. Não será exagero afirmar que há um paradoxo

da moralidade: Como resolver esta aparente contradição, se é que nos podemos permitir falar em aparência, entre autonomia e heteronomia?

O dilema do prisioneiro

Por vezes, em situações de interacção a opção por uma estratégia de compromisso é vantajosa. Uma tal vantagem na opção pelo compromisso é exemplarmente exposta no dilema do raptor e da vítima, desenvolvido por Thomas Schelling. A situação é simples: por razões que não importa, uma pessoa arrepende-se de ter raptado outra, procurando, por isso, garantir condições pelas quais possa libertar a sua vítima sem que, naturalmente, com isso lhe sucedam consequências desagradáveis, como ser denunciado pela vítima, perseguido pelas autoridades, etc. Porém, libertar sem mais a vítima pode ser bastante arriscado. Seria negar a garantia da salvaguarda do interesse próprio do raptor. A simples aplicação do modelo racional de decisão, com a maximização do interesse próprio da vítima, pode, assim, e algo ironicamente, condená-la à morte, como única forma de o raptor garantir o não prejuízo do seu interesse próprio. A saída para o dilema envolve algum tipo de compromisso perene entre a vítima e o raptor, de tal maneira que aquela possa ser libertada sem que isso signifique uma fragilização do interesse próprio do raptor.[7] Tal compromisso, contra o interesse próprio da vítima, acaba por a beneficiar.

Note-se, porém, que neste dilema o compromisso não corresponde ainda a uma cooperação baseada na confiança. Pode-se, por certo, falar de cooperação – vítima e raptor interagem de forma a se beneficiarem mutuamente, mas justamente por não haver entre elas nenhuma espécie

[7] "Se a vítima cometer um acto cuja revelação poderia levar à chantagem, poderá o raptor garantir o seu silêncio; se não, pode cometer um na presença do seu raptor, para criar uma situação que assegurará o seu silêncio." (Schelling, 1960: 43-44. Citado por Frank, 1997: 225)

de confiança. A vítima não denuncia simplesmente porque *não pode* e não porque o deseje, menos ainda porque confie no raptor. Se o compromisso tem uma função é a de colmatar uma ausência de confiança.

A situação que melhor pode dar conta de uma cooperação baseada na construção da confiança é o dilema do prisioneiro, dilema imaginado por Tucker e que influenciou profundamente quer aspectos ligados à Teoria dos jogos, quer suas aplicações em múltiplos domínios. Boa parte do nosso estudo será dedicada a este dilema. Começaremos por o expor na sua versão mais simples, não iterada e com apenas dois jogadores; em seguida, considerá-lo-emos enquanto jogo comunitário e, por fim, discutiremos a sua versão iterada.

Sejam, então, dois prisioneiros, sem possibilidade de se contactarem. Imaginem-se Bonnie e Clyde por exemplo, finalmente detidos após as suas perigosas aventuras. Havendo apenas prova suficiente para condenar ambos a uma pena leve, digamos 1 ano, um procurador sujeita-os, em interrogatório, e cada um por si, ao seguinte dilema: ou um delatar o outro, podendo por isso ser premiado com a liberdade, sucedendo ao outro arcar com uma pena pesada, digamos 3 anos, ou, caso a delação seja recíproca, calhando a ambos, por cumplicidade, uma pena intermédia, 2 anos; ou não delatar o outro, caso em que, se a decisão for recíproca, ambos sofrerão a pena mais leve, digamos 1 ano, mas se não houver reciprocidade, arcará com a pena mais pesada, ficando, além disso, o outro livre.

Dispondo os dados, obtém-se uma matriz

	Bonnie fala	Bonnie cala
Clyde fala	2/2	3/0
Clyde cala	0/3	1/1

Ora, racionalmente preferir-se-á delatar a não o fazer, pois delatando as opções serão ou 0 ou 2 anos de pena, ao passo que não traindo as opções serão ou 1 ou 3 anos de pena. Seria pois, pouco razoável, se está em causa maximizar o interesse próprio, e sem a introdução de factores especiais, alguém não delatar. E isto é pensar raconalmente.

Importa, neste momento, introduzir dois conceitos de análise da teoria dos jogos:

Estratégia dominante: Sejam A e B dois jogadores, A terá uma estratégia dominante quando, entre as suas estratégias, existe uma que responde melhor do que todas as outras às estratégias de B.

Equilíbrio de Nash: Conjunto de estratégias, um para cada um de dois ou mais jogadores, em que nenhum jogador pode incrementar o seu ganho sem, com isso, prejudicar o ganho dos restantes jogadores.

A propósito destes pontos de equilíbrio, é frequentes vezes acentuado o contraste entre John Nash e Adam Smith, ainda que, também frequentemente, de maneira equívoca. De acordo com a teorização de Nash, baseada na teoria dos jogos, o bem-estar social é maximizado quando cada indivíduo persegue o seu bem-estar, sob a consideração do bem-estar dos demais agentes que consigo interajam. Já de acordo com Adam Smith, habitualmente reconhecido como o pai da Economia, o máximo nível de bem-estar social emerge quando cada indivíduo persegue egoisticamente o seu bem-estar individual[8]. O contraste é óbvio: onde Smith considera apenas o interesse individual, Nash pensa também, e como condição para aquele, o interesse dos outros. Mas, sem contradição, é também Adam Smith quem logo reconhece que o egoísmo – que não é o mesmo que o comportamento agressivo (tal como o pacifismo comportamental não equivale ao altruísmo) – deve ser sujeito a uma condição, pois, embora apenas pela promoção, por parte de todos, dos interesses próprios individuais se possa alcançar o melhor para todos, Adam Smith reconhece que nem sempre assim sucede, pelo que o valor do egoísmo tenha de ser condicionado pela efectiva obtenção de ganhos para a sociedade como um todo. Não condi-

[8] "Não é da benevolência do açougueiro, do cervejeiro ou do padeiro que esperamos nosso jantar, mas da consideração que eles têm pelo seu próprio interesse", ("It is not from the benevolence of the butcher, the brewer, or the baker, that we expect our dinner, but from their regard to their own interest. We address ourselves, not to their humanity but to their self-love, and never talk to them of our own necessities but of their advantages.) " (Smith, 1776: I.ii.2)

cionado, o egoísmo será moralmente censurável, precisamente porque prejudicial ao interesse da sociedade como um todo. Onde o contraste por vezes erra é em pensar que Smith sustentaria um egoísmo incondicional.

Sobretudo por esta razão, mais do que opor Nash a Smith, alguns autores entendem a posição de Nash como uma resposta que resolve o problema que Smith deixara em aberto, a saber, como discriminar o egoísmo condicional, que promove o bem-estar do todo, do egoísmo individualista que não quer saber do bem-estar do todo em interacção. Neste sentido, Nash não se opõe a Smith, antes vem completá-lo. Dizer isto não obscurece, porém, o facto de que, com Nash, se efectiva realmente uma revolução face à teoria de Smith – tratando-se em ambos de promover o interesse individual, Smith pensa tal promoção como uma preocupação *exclusiva* pelo interesse próprio de cada um, de que emerge, pelo efeito da "mão invisível", o bem-estar comum; já Nash pensa a mesma promoção do interesse individual como uma preocupação *inclusiva* pelo interesse dos outros. A revolução pode ser enunciada da seguinte maneira: se o autor de *A Riqueza das Nações* (1776) diria que a ambição individual gera bem-estar comum, Nash contraporia que, pelo contrário, a ambição relativa ao bem-estar comum gera bem-estar individual.

Posto isto, interessa mostrar que, com os conceitos de estratégia dominante e de equilíbrio de Nash, é possível identificar qual é o conjunto de estratégias que deixam o jogo numa situação de equilíbrio. Tanto Clyde como Bonnie possuem uma estratégia dominante – delatar, para ambos. A vantagem de se possuir uma estratégia dominante reside em um jogador poder identificar qual é a melhor estratégia para si, isto é, qual é a que maximiza o seu interesse próprio, independentemente da estratégia que o outro jogador venha a adoptar.

Quer isto dizer que tanto Bonnie como Clyde podem racionalmente definir a melhor estratégia para si sem que haja, entre eles, qualquer dependência.

O equilíbrio de Nash neste jogo deixa-se identificar através das estratégias dominantes de ambos os jogadores, Bonnie e Clyde.[9]

	Bonnie fala	Bonnie cala
Clyde fala	2/2	3/0
Clyde cala	0/3	1/1

É fácil verificar, porém, que a maximização do *interesse próprio* de Bonnie e de Clyde não corresponde à maximização do *interesse comum* de ambos, que seria, claramente, ambos permanecerem calados. O resultado obtido, através da identificação do equilíbrio de Nash fica, assim, longe da solução óptima. Apenas se obtém uma solução subóptima (ou ainda, Pareto-ineficiente); na verdade, evita-se apenas o pior.

Este facto suscita a percepção de que há qualquer coisa de paradoxal ou, ao menos, de ineficaz, na racionalidade do dilema do prisioneiro. Além disso – ou, sobretudo, apesar disso –, o mesmo facto tem suscitado a expectativa de que uma tal ineficácia, não se verificando empiricamente na generalidade das interacções sociais que têm a forma de dilemas do prisioneiro, possa encontrar uma resposta capaz na regulação moral dessas interacções. Por outras palavras: a moral pode ser pensada como tendo por função a optimização.[10]

[9] Nem sempre, contudo, ambos os jogadores dispõem de uma estratégia dominante. Pode suceder que só um disponha de uma estratégia dominante ou mesmo que nenhum disponha de uma estratégia dominante. No primeiro caso, o jogador que não dispõe de uma estratégia dominante deverá escolher como sua estratégia a que melhor responde à estratégia dominante do outro jogador. No segundo caso, e havendo estratégias dominadas (isto é, estratégias que respondem sempre pior que uma outra das estratégias disponíveis), procede-se eliminando estas até surgir uma estratégia dominante. Não sendo possível apurar uma estratégia dominante, será sempre possível encontrar o equilíbrio de Nash recorrendo a estratégias mistas, isto é, estratégias que combinam probabilisticamente as estratégias originais ou puras.

[10] Escreve, a propósito, Bruno Verbeek: "Morality commits agents to avoid Pareto-inefficient or suboptimal outcomes. (...) On this view, the function of morality is to prevent the failures of rationality." (Cf. Verbeek, 2004)

Pensada nestes termos, funcionalmente pois, a moral justificar-se-ia como *compromisso estratégico* em vista de uma optimização que não seria alcançável se pensada apenas como resultado de interacções conduzidas aos pontos de equilíbrio de Nash. Veremos, em seguida, que tal compromisso estratégico pode efectivamente ser deduzido a partir da teoria dos jogos através de uma versão iterada do dilema do prisioneiro. Mais adiante, não deixaremos de discutir a expectativa de que este compromisso estratégico possa justificar uma moral, ou ainda, de que a moral possa ser pensada funcionalmente.

Versão iterada do dilema do prisioneiro

Consideremos, pois, uma versão do dilema do prisioneiro na qual o dilema seja iterado dez vezes, isto é, em que o mesmo dilema suceda, repetidamente, numa série de dez lances, importando, então, avaliar os resultados gerais da série completa.

A análise da versão iterada do dilema do prisioneiro mostrará como se pode preferir, ainda no quadro de uma racionalidade que vise a maximização da utilidade, estratégias de cooperação a estratégias baseadas no equilíbrio de Nash.

Quadro 1 – cooperação depois de convite

Bonnie (perdoa 1x)	Clyde (coopera p/ reacção)
1)3	1)
2)4	2)1
3)5	3)2
4)6	4)3
5)7	5)4
6)8	6)5
7)9	7)6
8)10	8)7
9)11	9)8
10)12	10)9

Quadro 2 – recusa de convite à cooperação

Bonnie (perdoa 1x)	Clyde (não coopera)
1)3	1)
2)6	2)
3)8	3)2
4)10	4)4
5)12	5)6
6)14	6)8
7)16	7)10
8)18	8)12
9)20	9)14
10)22	10)16

A comparação entre o Quadro 1 e o Quadro 2 mostra que o não cooperante ao 2^o lance não age de forma racional ao destruir a possibilidade de cooperação com Bonnie. Da sua estratégia, resultarão mais

7 anos de pena para ele. Portanto, é de se excluir racionalmente tal estratégia.

Uma segunda estratégia, menos benovolente, não equaciona o perdão, retaliando logo ao primeiro indício de não cooperação. Neste caso, a sequência de dez iterações obtém o seguinte cúmulo de penas a cumprir:

Bonnie (não perdoa)	Clyde (não coopera)
1)3	1)
2)5	2)2
3)7	3)4
4)9	4)6
5)11	5)8
6)13	6)10
7)15	7)12
8)17	8)14
9)19	9)16
10)21	10)18

As duas estratégias por que Bonnie poderia ter optado são estratégias puras. Tais estratégias receberam, na terminologia da teoria dos jogos, as seguintes designações:

– Tit for Tat ("toma lá, dá cá"): estratégia que retalia à primeira não-cooperação, e que podemos fazer corresponder à máxima do Antigo Testamento "olho por olho, dente por dente";

– Tit for two Tats:estratégia que perdoa uma vez, retaliando apenas à segunda não cooperação, e que corresponderá, naturalmente, ao preceito neotestamentário do "dá a outra face".

É ainda possível conjugar estas duas estratégias puras com procedimentos aleatórios de perdão – por exemplo, o *Tit for Tat generoso* – de maneira a deter efeitos de retaliação em cadeia no caso de se optar pela estratégia menos tolerante. Este tipo de conjugação origina estratégias mistas, em contraste com as puras.

Posto isto, em que falha exactamente o teorema de Nash quando se trata do dilema do prisioneiro iterado? É que nessa versão do dilema sucede a estratégia adoptada não ser a *estratégia dominante* (alcançando-se o ponto de equilíbrio); bem pelo contrário, ocorre uma *estratégia cooperante* (dependente da estratégia do outro participante), e o resultado, embora instável – sem ponto de equilíbrio pois – é melhor do que o subóptimo que se obtinha na versão não iterada do dilema do prisioneiro.

O que está em causa, note-se, é apenas a construção de uma cooperação vantajosa, cooperação que depende essencialmente da estabilização de um padrão de confiança entre os jogadores. Com efeito, as estratégias de acção visam estabelecer quanto antes um padrão de confiança e a confiança construída visa compensar a ausência de equilíbrio, permitindo a estabilidade de um resultado óptimo mas dependente de algo que é, por natureza, opaco, justamente a confiança.

Este enfoque na confiança, como finalidade relativamente à qual as *estratégias de acção*, suas promotoras, se definem como meios, torna-se especialmente evidente quando procedemos a uma segunda complexificação do jogo, de maneira a nele haver a possibilidade de recusar parceiros. Com efeito, nesta situação verifica-se que o mais importante não residirá na escolha entre estratégias de acção, mas na escolha entre *estratégias de selecção* de parceiros. A rejeição de parceiros não cooperantes é bem sucedida independentemente da estratégia de acção empregue.

Note-se que um *rationale* moral como a regra de ouro "Não faças aos outros o que não queres que te façam a ti", que se encontra tanto na *Bíblia* (*Êxodo*, IV, 16; *Lucas* 6,31), como em *Mahabarata* (XIII, 113), ou em Confúcio e Hillel, no Zoroastrimo e no Taoísmo, resulta bem como interpretação moral de uma estratégia geral que compreende alternativas de estratégias de acção, como, por exemplo, as mais bem sucedidas "dente por dente, olho por olho"veterotestamentária (*Êxodo* 21,22), ou "dá a outra face"neotestamentária (*Mateus* 5,38). Esta presença da *regra de ouro* em vastíssimas religiões candidata-a à posição

de universal de interacção comunitária. Mas, e talvez como explicação para essa possível universalidade, a teoria dos jogos permite deduzi-la de uma *racionalidade prática em acto* enquanto formulação da estratégia mais vantajosa.

Construção da confiança e ética

A moralidade, tal qual a reconhecemos, pode coincidir ou seguir-se desta construção racional da confiança? – Esta é a questão que nos importará.

Vimos que a construção da confiança possibilita e promove o comportamento cooperativo; além disso, retalia as traições à cooperação esperada. Pode, assim, ser pensada como construção de uma moralidade, entendida esta como conjunto de prescrições e proscrições práticas, e no sentido em que se deixam *interpretar* como estratégias de acção/selecção. Mas, já por outro lado se, ao optarem pela cooperação, os dois participantes no dilema do prisioneiro confiam um no outro, fazem-no, no entanto, apenas no estrito interesse próprio de cada um. Mal deixe de interessar, a cooperação *deveria* racionalmente cessar. Reconhecendo-se que a situação nunca foi outra, i.e., que a cooperação foi sempre *instrumental* então haverá que reconhecer que a cooperação não gera um genuíno e desinteressado altruísmo. Talvez mais importante: nem é de esperar que o devesse fazer.

Deduzir, ou reinterpretar, a moralidade a partir da teoria dos jogos, sob o suposto racional da maximização do interesse próprio dos agentes, significa assumir apenas uma moral interessada.

Isto contraria duas perspectivas éticas – em primeiro lugar, que a moral tenha de ser desinteressada (à maneira de Kant) e, em segundo lugar, que a moral simplesmente possa ser desinteressada. Ora, sobretudo a segunda contrariedade vai ao arrepio das nossas intuições morais

de que são possíveis actos de genuíno altruísmo ou de sacrifício sem contrapartidas.

Tomemos em consideração um exemplo que ilustra bem a dificuldade destra construção racional da confiança. No dilema do prisioneiro iterado, uma pessoa, digamos, o João, inicia a cooperação com duas outras, em separado, e que o faz, em ambos os casos, por razões, suponhamos, exclusivamente ligadas ao seu interesse próprio. Suponhamos, em seguida, que os dois cursos de iterações seguem padrões bem diferenciados, de tal maneira que, num deles, a cooperação com o parceiro corre na perfeição, ao passo que no outro curso a cooperação foi frequentes vezes traída pelo parceiro. Suponhamos, finalmente, que a iteração tem um fim e que, chegado o último lance, apenas João está ciente disso. Naturalmente, nesse último lance qualquer estratégia de João deixa de estar condicionada por uma futura retaliação por parte dos seus parceiros. A pergunta que se coloca então é a de saber se João terá um comportamento diferenciado para cada um dos seus parceiros, ou se decidirá da mesma maneira para qualquer dos casos.

As nossas "intuições morais", valham o que valerem, dizem-nos que João diferenciará o seu comportamento, ainda que nada tenha a ganhar com isso. Não se sentirá, ou sentir-se-á menos, inibido de não cooperar, no derradeiro lance, com um parceiro pouco cooperante; mas resistirá, ou resistirá mais, a não cooperar com um parceiro que se revelou dominantemente cooperante. Uma questão de consciência, pensará; ou de consciência moral.

Mas se admitirmos esta resistência desinteressada – chamemos-lhe *resistência ética* –, que fundamento, se não a maximização do interesse próprio, lhe podemos dar?

Por outras palavras, um indivíduo sentirá *dever* a cooperação a quem sempre cooperou consigo e que, além disso, tem uma expectativa bem fundada de que continue a cooperar. Já quanto a quem não revelou espírito de cooperação, o mesmo *sentido de dever* não terá razão de ser – nenhuma confiança será traída pelo simples facto de nunca ter havido construção da confiança.

Como pode, porém, a construção da confiança gerar este *sentido de dever*?

Segundo o modelo da decisão racional, a construção da confiança não ultrapassa nunca a posição de meio para um fim que não ela mesma: a maximização do interesse próprio. Isto poderia querer dizer que não há realmente lugar na moralidade racional para altruísmo ou sacrifício genuínos. Tais condutas seriam irracionais face a agentes que vêem a sua racionalidade como maximização do interesse próprio. Mas, a mesma instrumentalidade da cooperação e da confiança poderia querer dizer que, havendo de facto altruísmo e sacrifício genuínos, tais condutas seriam genuinamente morais justamente por não serem racionais.

Retomando, então, a situação do João, teríamos um dilema – ou João não agiu moralmente porque agiu irracionalmente, ou João agiu moralmente porque agiu irracionalmente. Portanto, a sermos consequentes, teríamos de concluir, das duas, uma: o altruísmo genuíno ou seria irracional mas não moral ou seria moral mas não racional. Contudo, as nossas intuições sobre racionalidade e moralidade dizem-nos precisamente o contrário, a saber, que João, ao agir por genuíno altruísmo, pode ter agido moral e racionalmente.

Será possível dar algum tipo de fundamento a estas intuições? Julgamos obter algum ganho se desdobrarmos esta pergunta em duas questões de natureza distinta: uma questão *de facto* sobre se haverá algum tipo de evidência comportamental que sustente tais intuições, e uma questão *de jure* sobre se haverá racionalidade que permita as dar por assentes em algum fundamento.

Quanto à questão *de facto,* mencionamos duas importantes referências que põem em causa a eficácia teórica do modelo de humanidade que se tornou habitual designar por *homo economicus*, ou seja, o ser humano modelado como agente racional e auto-interessado, cuja racionalidade é definida em termos de comportamento que maximiza a utilidade.[11] Por um lado, Philip Johnson-Laird e o seu estudo do raci-

[11] Esse é o pressuposto subjacente aos modelos da economia neo-clássica – *uma meta-teoria económica baseada na oferta e procura dependendo de agentes económi-*

ocínio silogístico. Por outro, Daniel Kahnemann e Amos Tversky com o seu modelo comportamental da decisão. Relativamente ao primeiro, basta-nos dar conta do facto de que, com o seu estudo, de natureza empírica, evidenciou que os seres humanos, homens e mulheres concretos, não raciocinam como seria de esperar dado o modelo de racionalidade clássico, mas que nem por isso deixam de ser racionais. Howard Gardner, em *A Nova Ciência da Mente*, tece, a propósito das investigações de Philip Johnson-Laird, a seguinte consideração:

Demonstrou serem insustentáveis as formas como se concebia que o ser humano abordava problemas de raciocínio; o ser humano não raciocina como sugeria a lógica clássica. No entanto, os seres humanos também não são irracionais.[12]

Já Kahnemann e Tversky demonstraram, com resultados experimentais, diversas dissonâncias entre o comportamento previsto pelo modelo de decisão racional – em que os agentes são "homens económicos" – e o comportamento real de seres humanos reais. Não obstante essas dissonâncias, o comportamento humano, entenda-se o de seres humanos reais, continua a exibir regularidades que justificam o desenvolvimento de *modelos comportamentais da decisão*

Concretizando um pouco mais, uma dessas dissonâncias comportamentais face ao modelo de decisão racional consiste num padrão de aversão às perdas que não se deixa traduzir por uma simples função de utilidade em que se contabilizem ganhos e perdas. A verdade é que, seguindo explicação de Robert Frank, "as pessoas não avaliam as alternativas com uma função de utilidade convencional, mas em vez disso, com uma *função de valor*, que se define nas *alterações* de riqueza"[13] e que, em particular, "uma propriedade importante desta função de valor é que é muito mais inclinada nas perdas do que nos ganhos".[14] Quer

cos operando racionalmente, cada um procurando maximizar a sua utilidade através de escolhas baseadas na avaliação de informação.

[12] Gardner, 2001: 489.

[13] Frank, 2001: 249.

[14] Frank, 2001: *ibidem*.

isto dizer que as pessoas não avaliam da mesma maneira os seus ganhos e as suas perdas, agravando estas face àqueles.

Um outro exemplo de dissonância que nos importa sobremaneira, veremos em seguida porquê, e que é perfeitamente reconhecível na experiência quotidiana de qualquer pessoa consiste em *não ignorarmos custos afundados*. Se paguei antecipadamente por algo, ir ao cinema por exemplo, então, apesar de isso por que paguei ter deixado, não importa por que motivo, de ser a minha preferência (já não maximizando o meu interesse pois), eu tendo a ter em conta na minha decisão sobre o que fazer a seguir os custos entretanto afundados, podendo, por isso, decidir contrariamente à minha preferência actual. Acabo, enfim, por ir ao cinema contrariado, simplesmente para justificar uma despesa que, de uma forma ou de outra, já não era recuperável. Quem ainda não se surpreendeu a pensar "se paguei por tudo, então hei-de querer tudo, mesmo que já não seja o caso de que prefira querer tudo"? Importa notar que esta tendência, além de intuitiva, está experimentalmente verificada. Portanto, o que merece a pena considerar é o seu sentido.

Obviamente, o modelo racional da decisão prescreve que deveríamos, racionalmente, seguir a tendência contrária, ou seja, *ignorar os custos afundados* nas nossas decisões. Se paguei por uma preferência – por exemplo, um fim-de-semana paradisíaco –, e se esse é um custo, em todo o caso, irrecuperável, que perco em não sair de casa se essa é, agora, a minha preferência? Que sentido faz, além do custo já afundado, teimar em fazer o que não prefiro? Ao fim e ao cabo, não foi uma preferência o que justificou o custo? Então, por que não dar sequência a uma preferência que não traz nenhum custo suplementar?

De certo modo, estas questões insinuam uma certa irracionalidade na nossa tendência a *não dar por perdidos custos afundados.* Mas será realmente assim?

Comecemos por notar que há uma *dívida potencial* que instala o *dever* de não dar por perdidos custos, independentemente do facto de se terem convertido, entretanto, em custos afundados. Por isso, os custos passam a ter não apenas um valor instrumental, relativo ao interesse

próprio do agente ou, ao menos, ao seu objectivo, mas um valor em si mesmos. *Mutatis mutandis*, se um indivíduo investiu na cooperação para obter um certo resultado, então tende a suceder que *resista* a não perseverar na cooperação, apesar de saber que alcançaria um resultado melhor caso não cooperasse. Tal qual como para os custos em geral, a cooperação tende a tornar-se, uma vez assumida, fim em si mesma, independente, pois, do facto de já não servir o interesse próprio. Há, pois, um *compromisso não instrumental* com os custos, mesmo que afundados, de uma acção.

Isto que se verifica relativamente a custos, pode igualmente ser dito de acções. Com efeito, se admitirmos que a acção humana integra não só o resultado, mas também os custos nela envolvidos, aliás todo o esforço nela investido (até mesmo o de a conceber), então o compromisso não instrumental com os custos é, na verdade, um compromisso não instrumental com a acção. Admitindo pôr as coisas nestes termos, então não ignorar custos afundados exibe um *compromisso agencial* dos agentes.

Apurando um pouco mais as consequências que nos é permitido extrair daqui, é razoável afirmar a validade deste compromisso agencial não só para *acções com termo definido*, como ir ao cinema, onde o "tempo de agência" começa, por exemplo, com a decisão de ir ao cinema e termina com a ida ao cinema, mas também para *acções para toda a vida*, como deixar de fumar, ter um filho ou casar, cujo tempo de agência coincida com o tempo de vida. Nestas últimas, observe-se, o sentido de dever para com a acção será permanente e permanentemente reforçável.

Com isto, fica lançada uma resposta positiva à nossa questão *de facto* sobre o comportamento do João em respeitar a cooperação mau grado a soberana perda de oportunidade de obter um ganhos acrescido. De facto, tal comportamento, por mais irracional que possa parecer, inscreve-se, como um exemplo entre tantos outros, no padrão de comportamento que nos leva a não ignorar custos afundados, ou seja, a nos

comprometer com as nossas próprias acções, independentemente das preferências que puderam estar na sua génese.

Supondo agora que os nossos comportamentos altruísticos tenham por base o que chamámos compromisso agencial, há que enfrentar a questão *de jure* sobre a racionalidade, ou não, de tais comportamentos e de tal compromisso. Será uma moral baseada no dever de não ignorar custos afundados, centrada na acção em todo o tempo de agência e não no interesse, uma moral irracional?

Naturalmente, esta questão (não sobre a realidade de uma moral desinteressada, mas sobre a sua justificação) só pode receber uma resposta negativa se for outro o conceito de racionalidade que tivermos em mente. E é neste ponto que a racionalidade prática kantiana ressurge com um rosto novo, mais dado à prova empírica, e como alternativa a uma racionalidade meramente instrumental, subordinada a preferências.

Já pudemos avaliar a ideia de que só há genuíno agir se for livre; por isso, o agir está sempre em questão na sua liberdade. Mas, assim veríamos apenas metade do que há a ver. A própria liberdade está também sempre em questão na acção. Se fazemos prova de que agimos genuinamente na liberdade com que agimos, também é só no nosso agir, e em como agimos, que fazemos prova da nossa liberdade. Por isso, *a liberdade é agencial*. Estando ela em causa na acção, então mais irrazoável do que ignorar uma preferência seria não ter em conta uma liberdade sempre em questão nas nossas decisões acerca do curso que imprimimos às nossas acções. Ora, é justamente Kant, com a sua filosofia moral, quem coloca a liberdade do agente na posição de fundamento racional do comportamento moralmente desinteressado. Se Kant recusa o dilema entre uma *moralidade racional mas instrumental* e uma *moralidade genuína mas irracional*, fá-lo através da afirmação de uma outra racionalidade prática que não a instrumental e que se opõe de algum modo a esta. Trata-se, é sabido, da contraposição aos imperativos hipotéticos de um imperativo categórico, ou ainda, da

contraposição a um agir heterónomo de uma liberdade pensada como autonomia.

Dar por perdidos custos afundados é dar por perdida, anulada, a agencialidade desses custos; é, pois, anular a liberdade agencial. E isto por não se resistir à promoção do interesse próprio. Ou seja, em termos mais kantianos, não ignorar os custos afundados é afirmar a autonomia da vontade – "dar-se a si mesmo a lei" – contra a heteronomia do interesse próprio. Recorde-se que a ética kantiana assenta na possibilidade, para uma vontade, de universalizar as suas máximas, isto é, de um agir determinado já não pelo interesse e pela inclinação mas pela própria razão, enquanto faculdade de legislar a regra para a acção. Em suma, a liberdade agencial exprime a liberdade como autonomia kantiana. Não ignorar os custos afundados, não os dar por perdidos, consistirá numa *resistência* à anulação da acção, resistência bem fundada sob a consciência de a acção dever ser livre e de a sua anulação ter por fundamento o interesse próprio, justamente aquilo de que nos podemos libertar fazendo assim prova da nossa condição livre.

Assim, a racionalidade para um *modelo ético da decisão* já não pode estar centrada no interesse próprio, mas na acção propriamente dita, entendida, pois, já não como meio, *apenas* como meio, mas *também* como fim a perseguir.

Se com isto se reencontra Kant e a sua racionalidade prática como resposta à nossa questão *de jure* sobre a justificação do comportamento humano altruístico, e sobre o sentido do sacrifício de preferências face a um compromisso agencial, a verdade é que também é um kantismo novo que assim encontra razão de ser, assente na interacção social concreta entre pessoas humanas reais, interacção movida sempre por interesses e preferências particulares. Jean-Pierre Dupuy, bem a propósito, fala de um kantismo de rosto "humano":

Um kantismo que conserva da tradição empirista inaugurada por Hume uma atenção prestada aos interesses das partes, interesses que os põem em conflito, mas também os incitam à cooperação; mas, igual-

mente, kantismo de rosto "humano", porque se dirige a seres interesseiros e não a anjos.[15]

É claro que o contraste com a inumanidade do Kant original está na indiferença deste para com a base relacional, de interacção entre agentes racionais, que *constrói* localmente a moralidade. Por isso, a moral de Kant, longe de dever ser oposta à teoria dos jogos, deve encontrar nesta as condições para a emergência de uma moralidade relacional concreta.

[15] Dupuy, 1999: 374.