O ELO PERDIDO DAS CIÊNCIAS DO ARTIFICIAL
(ou da Economia como uma das Ciências do Artificial)

Consideramos "ciências do artificial" aquelas teorias e práticas científicas que procuram realizar em máquinas, concebidas ou construídas pelos humanos, certos comportamentos ou capacidades tomadas como objecto de interesse por serem consideradas típicas dos próprios humanos (ou de animais) que encontramos na natureza. Exemplos de linhas de investigação das ciências do artificial são a Inteligência Artificial ou a Nova Robótica, de que são empreendimentos típicos o xadrez computacional e o futebol robótico, respectivamente. Um problema clássico das ciências do artificial consiste em saber se as máquinas podem ter relações intrinsecamente significativas com o mundo ou apenas relações atribuídas por humanos. Algo como perguntar: será que há mesmo computadores que jogam xadrez e robots que jogam futebol, ou eles não fazem nada disso e nós é que os vemos desse modo?

Neste texto usaremos esse problema para tentar avançar na compreensão do cimento que une plurais linhas de investigação no mesmo arquipélago das ciências do artificial, que não são uma disciplina científica, mas uma constelação de disciplinas – e uma constelação em evolução, com fronteiras imprecisas. Começamos por sugerir que é comum aos diferentes ramos das ciências do artificial o facto de nelas, sistematicamente, se laborar na ilusão de certas entidades terem relações significativas genuínas com o mundo, quando essa aparência é estritamente dependente da atribuição de significado por parte de intérpretes humanos. Propomos, depois, que o mecanismo básico dessa ilusão é a invisibilidade da interpretação aos olhos dos próprios intérpretes. Terminando defendendo que outras linhas de investigação, ao assentarem metodologicamente no recurso à invisibilidade da interpretação, se qualificam também como parte da constelação das ciências do artificial.

A ilusão constitutiva das ciências do artificial

Desde 1956 que a Inteligência Artificial (IA) desenvolve uma tentativa sistemática para construir máquinas com uma relação significativa genuína com o mundo. Experiências clássicas da IA assentavam na ideia de que era possível dar um mundo a uma máquina – programando-a. O computador, programado numa linguagem que tivesse capturado adequadamente a relação com a realidade exterior, passava a estar numa relação inteligente com o mundo. É o humano, programador, que desenha antecipadamente essa relação – de tal modo que a semântica, afinal, dispensa qualquer relação actual entre linguagem e realidade. O facto é que a história da IA mostrou dificuldades

fundamentais desta concepção. Para compreender o essencial dessas dificuldades é adequado focar um dos problemas teóricos mais persistentes das ciências do artificial, que consiste em saber se as máquinas podem ter alguma forma de intencionalidade e, sendo o caso, se essa intencionalidade pode ser própria ou apenas intencionalidade derivada (por atribuição de intérpretes humanos). Vejamos.

O paradigma central do programa de investigação da IA clássica é a "hipótese do sistema simbólico físico" (HSSF), cuja formulação canónica se deve a Allen Newell e Herbert Simon[1], para quem a HSSF constitui solução para "aquilo a que os filósofos chamam o problema da intencionalidade": "como é que os símbolos num sistema simbólico representam algo externo ao sistema simbólico". Os banais computadores digitais electrónicos constituem o exemplo mais familiar dos sistemas simbólicos físicos (SSF) de que a HSSF é uma teoria – mas a HSSF abrange algo mais.

Sendo os símbolos conjuntos de padrões físicos susceptíveis de certas relações físicas entre si (permitindo combinar espécimes em expressões), um SSF é uma máquina que, por aplicação sucessiva de processos modificativos, produz no tempo séries de estruturas simbólicas. Num exemplo do que seria um SSF, estes autores colocam uma memória, que armazena um conjunto de expressões que constituem as referências de um conjunto de símbolos; um conjunto de operadores que processam símbolos; um controlo que aplica um operador à expressão simbólica activa; uma via receptora para novas expressões que descrevem o ambiente externo; certas ligações entre operadores e órgãos motores produtores de comportamento externo[2]. Então, segundo a *HSSF, um SSF tem os meios necessários e suficientes para a acção inteligente geral*[3]. A HSSF associa-se explicitamente à ideia de que quer humanos quer computadores são instâncias de SSF e que os símbolos dos computadores e os dos humanos são os mesmos[4].

Como um SSF é uma máquina que existe num mundo de objectos mais vasto do que o conjunto das expressões simbólicas, precisamos de duas noções centrais para compreender a relação de intencionalidade entre símbolos e outros objectos: *designação* e *interpretação*. Ora, na explicação do mencionado exemplo, embora se fale de órgãos "receptores" e "motores", parecendo haver ligações de e para o mundo, essa ilusão desfaz-se quando se explicitam as noções de "designação" e "interpretação". A *designação* ("Uma entidade X designa uma entidade Y relativamente a um processo P se, quando P toma X como input, o seu comportamento depende de Y"), supostamente, dota o sistema de uma "acção à distância" (na expressão de Newell), porque o comportamento do sistema não é uma função dos símbolos propriamente ditos, mas uma função das entidades que os símbolos designam. Só que, quando vamos à descrição técnica do operador que implementa a designação ("acesso"), percebemos que as relações de acesso que podem ser criadas são apenas as que ligam símbolos dentro da máquina a outras entidades dentro da mesma máquina[5]. Quando se trata de descrever a *interpretação* – "o acto de aceitar como input uma expressão que designa um processo e então executar esse processo" – afirma-se que os símbolos que designam operadores são essenciais, porque contêm uma semântica externa, apontando para comportamentos

[1] Newell e Simon, 1976; Newell, 1980.
[2] Newell, 1980:142-147.
[3] Newell e Simon, 1976:116.
[4] Newell, 1980:135-136.
[5] Newell, 1980:156, 160.

que incorporam o sentido que as operações do sistema fazem no mundo exterior[6]. Só que não é dada qualquer explicação acerca de como isso se produz. Porque não se produz – e assim continuará nas sucessivas reelaborações desta proposta.

Fodor foi cortante nesta questão. Comentando o robot simulado SHRDLU, para quem Winograd tinha programado um "mundo" acerca do qual o robot pudesse "falar", critica a pretensão de que as frases das linguagens de programação ganhem uma semântica genuína quando interpretadas para linguagem-máquina, por esta ligar directamente "ao mundo", isto é, aos estados físicos mais elementares do computador. Interpretar desse modo, digamos, a frase "Boise é uma cidade" é dizer que a expressão "BOISE" aponta para um endereço de memória com o rótulo "CIDADE". Isso não é mais do que pretender que "Napoleão venceu a batalha de Waterloo?" quer dizer "Verifique se a frase 'Napoleão venceu a batalha de Waterloo' ocorre no volume que tem o número XXX,XXX na numeração decimal de Dewey na Secção da Rua 42 da Biblioteca da Cidade de Nova York"[7].

Haugeland (1985) identifica o núcleo duro da IA clássica com a HSSF, que traduz na tese de que tanto os computadores como os humanos são sistemas formais automáticos interpretados – mas precisa que essa tese depende essencialmente de outra, a "divisa formalista". Tendo os espécimes de símbolos num sistema formal "duas vidas" – uma "vida sintáctica", na qual são marcas sem significado, manipuladas exclusivamente de acordo com as regras internas do jogo, e "uma vida semântica", na qual têm significados apontando para o mundo exterior – a "divisa formalista" é: "Trata da sintaxe, que a semântica trata dela própria"[8]. Quer dizer: aceites como verdadeiros os axiomas do sistema formal, se as regras de inferência preservam a verdade, então qualquer processamento pelo "sistema formal automático interpretado" de uma fórmula com sentido à entrada resultará, à saída, numa fórmula com sentido na mesma acepção. Quando uso uma calculadora, o resultado obtido carregando na tecla "=" tem sentido debaixo da mesma interpretação que empreguei para escolher o arranjo de teclas com que inseri os dados, ordenadas pela pergunta que pedia aquela resposta. O problema desta leitura é que renuncia directamente, no caso das máquinas, a qualquer forma de relação com o mundo que não seja meramente derivada das atribuições de sentido a cargo dos humanos: sou eu, humano, que escolho o arranjo de teclas para inserir os dados, tal como sou eu que leio e uso o resultado devolvido pela máquina. Tudo isso faz um determinado sentido na minha vida – mas não na "vida" da calculadora.

A IA começou a enfrentar resolutamente este problema depois da sua formulação explícita por Stevan Harnad (1990), que o baptizou como "problema da fundação dos símbolos": como é que a semântica de um sistema formal automático interpretado podia ser intrínseca e não parasita dos intérpretes humanos, se um computador digital com programa armazenado está face ao mundo como alguém tentando aprender chinês como primeira língua a partir do zero, apenas usando um dicionário chinês-chinês, ou mesmo todas as obras existentes escritas em chinês, mas nada além disso: nem outras linguagens, nem qualquer experiência acerca do mundo (Harnad 1989). A sua resposta implicava que um SSF carecia de algum subsistema capaz de captação senso-

[6] Newell, 1980:158.
[7] Fodor, 1978:204-211.
[8] Haugeland, 1985:106.

rial, pelo qual o mundo exterior impressionasse por via não simbólica o processamento simbólico. Posteriormente (Harnad 2002), ao reformular o problema para extirpar o enviesamento internalista, sublinha o interesse de considerar o conteúdo lato (a parte do mundo exterior no significado), mas alerta contra outro erro: tentar escrever um modelo do mundo e programá-lo directamente na máquina (algo como escrever o modelo do mundo "dentro da cabeça") será igualmente improdutivo.

Claramente, mesmo os críticos da IA clássica mostravam dificuldades em descolar do erro de considerar apenas o mecanismo, distraindo-se do impacte da interacção histórica na moldagem da relação entre um sistema e o seu mundo. A Nova Robótica prometeu novidades nesse campo, ao desafiar a legitimação que o funcionalismo (Putnam 1960) emprestava ao desprezo generalizado pela questão da realização física da mente das máquinas. Nessa "Nova IA" (Robótica) jogou papel destacado Brooks (1999), que tenta dispensar os símbolos, concentrar-se nos comportamentos, dar aos sistemas perceptivo e motor o trabalho da "fundação física" dessas "criaturas", tentando assim uma ligação tão directa ao mundo exterior que a própria necessidade de representações é suprimida ("o mundo é o seu melhor modelo"). Felizmente, a nova IA Robótica sobreviveu a essa estratégia radical, que se revelou incapaz de compreender uma inteligência pelo menos tão sofisticada como a humana (Steels 2003).

Interessante é que a compreensão biológica das funções podia ter evitado alguns grandes problemas à IA: é que, como explicou Millikan (1984), as funções, não sendo causas, mas efeitos – efeitos de uma história evolutiva – não podem ser instaladas "à mão" por projectistas humanos, instantaneamente, em máquinas que assim ficam com "a história errada" – e, assim, com mecanismos incapazes de intencionalidade genuína. Não obstante, a maioria dos robots continuam a padecer da mesma doença: saem directamente da mão do artesão, ou da linha de montagem, e não têm qualquer história evolutiva. Pode esperar-se uma futura viragem na robótica, por via da Robótica Evolucionista (Nolfi e Floreano 2000) e das suas ferramentas inspiradas na evolução natural (como o algoritmo genético). Mas, enquanto essas promessas não se cumprem, teremos de insistir em que a evidência disponível sugere que as "máquinas inteligentes" continuam a ser apenas partes do nosso mundo, objectos das nossas atribuições de sentido, sem traços de qualquer relação intrinsecamente significativa com o mundo, algo que provavelmente só poderiam adquirir se pertencessem a uma espécie com todas as contingências de uma história de (co-)evolução.

Esta viagem a passos largos pela história das ciências do artificial mostra que começa a haver a compreensão de que esse programa de investigação é atravessado por uma ilusão constitutiva. A pretensão de que computadores ou robots têm uma relação significativa genuína com o mundo (fala-se de "computadores que jogam xadrez" e de "robots que jogam futebol"), apesar de tudo o que nessas entidades aparenta uma relação própria com o mundo ser tão-somente o resultado de actos de atribuição de significado por intérpretes humanos – é uma ilusão. Mas é uma ilusão constitutiva das próprias ciências do artificial, que ajuda a delimitar as suas fronteiras. Aqueles cientistas que trabalham com os mesmos tipos de entidades (computadores e robots), mas os assumem como ferramentas e não falam deles como agentes intencionais, estão, apenas por isso, fora das ciências do artificial (mesmo que estejam nos mesmos departamentos de engenharia ou de ciências da computação). Vale a pena tentar perceber como funciona essa ilusão constitutiva – o que faremos voltando à ideia de que as "máquinas inteligentes" são sistemas formais automáticos interpretados.

Da invisibilidade da interpretação

A mencionada tese de Haugeland (1985) acerca da IA clássica implica renunciar à ambição de que as máquinas programadas tenham uma relação intencional genuína com o mundo: a calculadora não faz contas, quem faz contas somos nós. O mesmo se a máquina for um computador sofisticado ou um robot. Quer dizer: os sistemas da IA clássica só são o que parecem ser debaixo das nossas interpretações. Seria útil compreender como pode ser tão subtil a intervenção das nossas atribuições de sentido – ao ponto de elas possibilitarem a ilusão de que falámos antes. Tentaremos agora clarificar esse ponto centrando-nos na questão da interpretação de sistemas formais. Se "tanto os computadores como os humanos são sistemas formais automáticos interpretados", donde vem a interpretação?

Dada uma linguagem formal, cuja sintaxe tenha importado o sentido habitual nessa linguagem para as suas expressões lógicas (como conectivas e quantificadores), a possibilidade de estabelecer uma relação entre essa linguagem e alguma realidade exterior passa principalmente por fornecer uma interpretação às expressões não lógicas (como letras para variáveis, letras para nomes, letras para frases, letras para predicados). Isso implica dar um significado a essas expressões, por meio de várias operações, tais como: especificar um domínio D, atribuir uma referência em D às letras para nomes e às letras para predicados, … Uma interpretação dá significado às expressões da linguagem na medida em que especifica como é que as expressões elementares dessa linguagem contribuem para a determinação do valor de verdade das fórmulas que contêm essas expressões. Por isto, uma mesma fórmula de uma linguagem pode ser verdadeira debaixo de uma interpretação e falsa debaixo de outra interpretação. Podemos "dar um mundo" a uma teoria formal (axiomas e regras de inferência) construída numa linguagem formal (com os seus elementos primitivos e as suas regras de sintaxe) – dando-lhe uma interpretação.

É conveniente, para tentar fazer luz sobre esta questão, considerar como Alfred Tarski tratou de captar o essencial da concepção clássica de verdade como correspondência numa definição formalmente correcta[9]. De acordo com a proposta de Tarski, lidamos com a verdade de uma frase numa determinada linguagem particular por meio de frases-V. Uma frase-V é uma definição, na forma de uma frase bicondicional, tendo do lado esquerdo o *definiendum* e do lado direito o *definiens*, como no exemplo:

A frase "Sócrates é mortal" é verdadeira se e somente se Sócrates é mortal.

Cada frase-V diz como tem de ser o mundo para que uma certa frase seja verdadeira. O esquema-V será o esquema de todas as frases-V:

X é verdadeira se e somente se p.

No esquema-V, a letra "p" é para ser substituída por qualquer frase declarativa e "X" é para ser substituído por um nome dessa frase. A forma mais comum de nomear uma frase é citá-la. Nessas circunstâncias, no *definiendum* de uma frase-V ocorre a menção da frase que se quer dizer em que condições é verdadeira – sendo a mesma frase usada no *definiens*. Se não atendêssemos a esta distinção entre menção e uso de uma frase, esta parcial definição de verdade poderia parece circular. Essa sensação de

[9] Particularmente relevantes para compreender este projecto são os textos "O Conceito de Verdade nas Linguagens Formalizadas" (1933) e "A Concepção Semântica de Verdade" (1944). Para as breves noções que aqui serão apresentadas, seguiremos a exposição contida em (Santos, 2003).

circularidade pode ser esclarecida se tornarmos mais fácil de ver que mencionamos uma frase de uma língua diferente daquela em que se está a fazer a definição, como em

"Snow is white" é verdadeira se e somente se a neve é branca.

Em rigor, no sistema de Tarski a frase citada é sempre de uma linguagem diferente da linguagem em que se está a fazer a definição. É a distinção entre linguagem-objecto e metalinguagem, necessária ao rigor que evitará o caos das linguagens naturais.

Uma forma de caracterizar o que é comum nas frases-V é dizer que uma frase é verdadeira se há concordância entre aquilo que ela afirma e aquilo que existe na realidade. Dar uma definição de verdade para uma determinada linguagem-objecto é dizer o que é comum nas frases-V sobre essa linguagem. A convenção-V, que é o critério para julgar a adequação material de uma definição formal de verdade numa determinada linguagem, estipula que uma tal definição do predicado "é (uma frase) verdadeira" deve ter todas as frases-V de uma dada linguagem como suas consequências. Assim, a convenção-V é uma forma de tentar dar, em termos mais precisos, a "concepção clássica da verdade". O que mostra que esta concepção de verdade é uma concepção semântica.

Do ponto de vista do contexto histórico, é interessante notar que a teoria da verdade de Tarski encontra uma forma subtil de contornar o cepticismo semântico do positivismo lógico. Se, por um lado, entra numa zona (relação entre linguagem e realidade) interdita à filosofia concebida como análise lógica da linguagem, por outro lado evita uma tarefa teórica espinhosa: dar uma noção rigorosa de significado. De acordo com (Santos, 2003), a estratégia de Tarski para contornar o problema do significado passa por deixar completamente ao utente da linguagem a responsabilidade pelo conteúdo factual, libertando as suas definições de verdade de qualquer conteúdo empírico. Afinal, a abordagem formalista à questão da relação entre linguagem e mundo cria uma ilusão: o que parece dar-nos, afinal limita-se a pedir-nos que façamos nós, enquanto utilizadores da linguagem, de fora do próprio formalismo. Explicitemos.

A operação fundamental de Tarski para fugir à dificuldade de dar uma noção rigorosa de "significado" consiste em impor uma condição meta-teórica à construção de frases-V. Sendo o esquema-V, como vimos,

X é verdadeira se e somente se p

porque é que a frase

"A neve é branca" é verdadeira se e somente se 1+1 = 2

não se qualifica como instância do esquema-V? Porque Tarski impõe que a frase do lado direito seja a tradução na metalinguagem da frase do lado esquerdo da bicondicional, entendendo-se que só há tradução se ambas têm o mesmo significado. Ricardo Santos clarifica assim a operação[10]: "Se olharmos apenas para a definição, nós *vemos* que quem a formulou esteve de facto a traduzir as expressões da linguagem-objecto para a metalinguagem. Mas isso não é *dito* em momento algum da definição. Deste modo, Tarski conseguiu evitar a aparente necessidade de uma teoria do significado."

Podemos extrair desta análise que os mecanismos da teoria só funcionam quando nós, utentes da linguagem, fornecemos conhecimento do significado das frases envolvidas. Seja L um fragmento do inglês e seja a frase

"Snow is white" é verdadeira em L se e somente se a neve é branca

[10] Santos, 2003:231-235, 234-235 para a citação. Os realces são do original.

Como é que eu posso reconhecer estar perante uma definição parcial de verdade se não conhecer o significado da frase citada do lado esquerdo da bicondicional? Bastaria, para isso, não saber ler inglês. Só que o caso é mais geral: mesmo que aquela frase estivesse em português, tudo continuaria a depender de eu ser capaz de fornecer o respectivo reconhecimento do significado.

É que as definições tarskianas de verdade não possuem conteúdo empírico. Elas são (apenas) verdades lógico-matemáticas. Só porque estamos anteriormente munidos desse conhecimento é que sabemos que a definição dá, para cada frase, as respectivas condições de verdade. Aliás, só por isso é que sabemos que se trata de uma (parcial) definição de verdade[11]. Então, temos que concluir que, se a teoria tarskiana evita a necessidade de uma teoria do significado é, apenas, porque cada utente fornece o seu próprio conhecimento do significado. A teoria formal só dispensa o que nós lá colocamos de forma independente da teoria – e sem isso a teoria formal não faria o seu trabalho.

O que aqui está em causa é o horizonte escamoteado dos sistemas formais. É que a linguagem formalizada tem um horizonte (não formal) de onde retira o seu sentido e a sua justificação (mesmo que esta se pretenda "meramente" pragmática). A teoria formal, numa linguagem formal, não surge do nada: "a selecção dos axiomas é orientada pelo desejo de que eles, quando interpretados, se transformem em frases verdadeiras; e a escolha das regras de inferência é igualmente guiada pela intenção de que elas sejam preservadoras da verdade, quer dizer, que elas não permitam derivar frases falsas a partir de frases verdadeiras; além disso, queremos ainda que o conjunto de axiomas e de regras seja suficiente para derivar todas as verdades da teoria"[12].

Ricardo Santos descreve assim o que constitui uma "segunda leitura" do processo formal – mas a "segunda" leitura vem antes. Trata-se de uma "leitura interpretativa" que começa antes da própria construção formal – e que, como vimos, a guia. É que sem isso, como lembrou o próprio Tarski, o exercício formal é em si mesmo irrelevante. A relevância está ligada ao facto de que a verdade ou falsidade de uma frase depende (também) daquilo que ela significa – e significa algo concreto. Uma frase não é apenas um conjunto de marcas de tinta no papel, razão pela qual a questão da verdade não se pode colocar a respeito de uma linguagem formal não interpretada. Uma linguagem formal sem interpretação não "quer dizer" nada. E ninguém pode dizer nada com ela.

Reconhecemos agora como funciona a ilusão nas ciências do artificial: interpretamos, a partir do nosso mundo, o que pusemos um computador ou um robot a fazer, e atribuímos a esse sistema o significado que só está no nosso olhar. Olhamos para uma frase-V e vemos lá uma tradução entre linguagem e realidade – mas não vemos que somos nós que autorizamos a tradução. Olhamos para um computador "a jogar xadrez" e não vemos que não há xadrez nenhum para o computador – apenas para nós, que montámos o cenário e o interpretamos a partir do nosso horizonte de sentido. É tão natural (provavelmente devido à nossa história evolutiva) que interpretemos sistematicamente, que atribuamos significado a tudo o que encontramos no nosso mundo, que chegamos a não identificar o nosso papel nessa atribuição. A invisibilidade da interpretação – aos olhos dos próprios intérpretes – é, afinal, o mecanismo básico da ilusão constitutiva das ciências do artificial.

[11] Santos, 2003:248-250.

[12] Santos, 2003:156.

Ora, compreender a invisibilidade da interpretação como o mecanismo básico da ilusão constitutiva das ciências do artificial, pode levar-nos a encarar de outro modo a questão das fronteiras do arquipélago das ciências do artificial. Haverá outras disciplinas científicas, ou pelo menos certas linhas de investigação no interior de certos campos disciplinares, cujos métodos assentem na invisibilidade da interpretação – qualificando-se, assim, como parte das ciências do artificial? O que defendemos de seguida é uma resposta positiva a esta pergunta, o que faremos considerando o uso da Teoria dos Jogos em Economia.

Da economia como uma das ciências do artificial

Nas últimas décadas, a Teoria dos Jogos (TJ) tornou-se uma das principais ferramentas teóricas da Economia[13]. A TJ é uma tentativa altamente formalizada de representar e analisar situações em que jogadores (agentes que têm de tomar uma decisão) racionais interagem entre si, tendo em conta a racionalidade dos seus oponentes. Usaremos aqui o "dilema do prisioneiro" (DP), um dos modelos mais estudados em TJ, como material de base para a exposição.

Uma apresentação clássica do "dilema do prisioneiro" é como segue[14]. Dois homens suspeitos de cometerem um crime grave em conjunto são presos e colocados incomunicáveis em celas separadas. Cada um deles pode confessar ou negar o crime. Se nenhum confessar, não haverá forma de provar o crime e os homens só serão condenados por um crime muito menos grave (um ano de prisão para cada um). A confissão confere o direito a um tratamento mais favorável, por constituir colaboração com a justiça (se ambos confessarem, cinco anos de prisão para cada um). Se apenas um deles confessar, o crime será considerado provado: o que confessa é libertado, o outro é condenado a 20 anos de prisão.

A matriz que se segue traduz a situação. As possíveis estratégias do Suspeito 1 são identificadas em linha, as do Suspeito 2 em coluna. Em cada célula da matriz representa-se um perfil de estratégias: o resultado da conjugação de duas estratégias, com o pagamento obtido pelo Suspeito 1 mencionado na primeira posição e o obtido pelo Suspeito 2 na segunda posição.

		Suspeito 2	
		Negar	*Confessar*
Suspeito 1	*Negar*	(1 ano, 1 ano)	(20 anos, liberdade)
	Confessar	(liberdade , 20 anos)	(5 anos , 5 anos)

Para cada suspeito (jogador), negar é cooperar com o outro, confessar é trair. Pode parecer que o melhor resultado para ambos resultaria da cooperação (ambos negam, 1 ano para cada um). Mas, cada um analisando a sua situação concluirá que, qualquer que seja a estratégia do outro, o melhor para si próprio é não cooperar (confessar). Vejamos o raciocínio do Suspeito 1: no caso do outro negar, se eu negar apanho 1 ano, se eu confessar vou em liberdade; no caso do outro confessar, se eu negar apanho 20 anos,

[13] Rubinstein, 1990:11

[14] Em inúmeras obras de exposição básica da teoria dos jogos aparece esta apresentação ou alguma equivalente. Uma sólida introdução à Teoria dos Jogos é Osborne e Rubinstein, (1994).

se eu confessar apanho 5 anos. O raciocínio do Suspeito 2 dará o mesmo resultado.

O DP pode ser posto numa forma mais geral, como exemplificado na seguinte matriz:

		Jogador 2	
		Cooperar	*Desertar*
Jogador 1	*Cooperar*	(R , R)	(S , T)
	Desertar	(T , S)	(P , P)

O DP tem sido usado para caracterizar situações económicas, quer trocas económicas elementares entre indivíduos, quer a competição entre grandes empresas. Essencial à caracterização do DP é a sua estrutura de pagamentos (resultados obtidos pelos jogadores em função das suas acções conjugadas), dada pela desigualdade **T > R > P > S**, onde R = "Recompensa" por cooperação mútua; P = "Punição" por deserção mútua; S = "Sonso", aquele que coopera quando o outro é um desertor; T = "Tentação" a que cede aquele que deserta quando o outro coopera. O que a desigualdade significa é que "Tentação" é melhor que "Recompensa", que é melhor que "Punição", que é melhor que "Sonso". (No caso dos prisioneiros, melhor é ter menos anos de cadeia)[15].

Isto quer dizer várias coisas: o pior resultado possível para um jogador é o pagamento a uma vítima de deserção que não sabe agir em conformidade (S); o melhor resultado possível para um dos jogadores é o pagamento a um desertor individual (T); a estratégia dominante num único encontro é desertar, qualquer que seja a escolha do oponente; a melhor escolha individual para cada um dos jogadores (desertar) parece incapaz de alcançar o melhor resultado colectivo. É a desigualdade nos pagamentos que gera a tensão entre o interesse colectivo e o interesse individual. Até porque um dos pressupostos básicos da TJ é que cada jogador age apenas para maximizar o seu "pagamento".

Um dos aspectos centrais para compreender o tipo de raciocínio da TJ é a proeminência teórica das soluções de equilíbrio: boas soluções são situações que tenderão a manter-se estáveis por opção racional dos jogadores, que são optimizadores da utilidade individual. É particularmente relevante o conceito de "equilíbrio de Nash" de um jogo estratégico, que é uma situação com a seguinte propriedade: para todas as acções ao dispor dos jogadores, nenhum jogador tem ao seu dispor uma linha de acção que lhe permita obter um pagamento superior ao que obtém actualmente, na condição de que todos os outros jogadores mantenham as respectivas linhas de acção.

A TJ interessa à economia, porque os "jogos" (o DP, o "jogo do ultimato", o "jogo do ditador", …) são modelos. A Economia usa-os como usa outros modelos matemáticos: constrói-os, manipula-os, analisa-os, transfere os resultados para o mundo real. Os modelos da TJ têm duas componentes: as estruturas e as narrativas. As narrativas são as histórias como a dos dois prisioneiros, as estruturas são as matrizes[16], tendendo

[15] Considera-se por vezes que uma segunda condição deve ser respeitada para termos um verdadeiro DP, dada pela desigualdade R > (S+T)/2 , traduzindo a ideia de que a cooperação deve dar melhores resultados do que a alternância de mútuo acordo entre cooperação e deserção.

[16] As matrizes são as estruturas para os jogos apresentados na forma estratégica. Para os jogos apresentados na forma extensiva, que, sem perda de generalidade, não consideramos neste texto, as estruturas são as árvores.

estas a ser vistas pelos economistas como o verdadeiro objecto de estudo da TJ. A seguinte citação de um manual de TJ exemplifica a (dominante) visão descontextualizada dos "jogos"[17]: «É útil pensar nas estratégias dos jogadores como correspondendo a várias "botões" num teclado de computador. Os jogadores são considerados como estando em salas separadas e sendo solicitados a escolher um botão sem comunicar uns com os outros.»

Contudo, de acordo com Grüne-Yanoff e Schweinzer, (2008), esta visão está errada: a estrutura de um jogo só tem propriedades formais, pelo que, só por si, não possui qualquer ligação a qualquer situação económica no mundo. As estruturas dos jogos, para significarem alguma coisa, carecem de uma interpretação – e essa interpretação é fornecida pelas narrativas. As narrativas são essenciais aos modelos na medida em que, fornecendo as interpretações, são elas que permitem ligar o modelo a certas situações económicas no mundo.

Já anteriormente tinha sido sublinhado o papel das narrativas como a sede da interpretação nos modelos da TJ. Para Morgan (2007:168-178), compreender o uso da TJ em Economia passa por dar o devido valor ao papel desempenhado pelas narrativas, que contêm, implícita e explicitamente, elementos sem os quais a TJ não pode sequer ser aplicada a situações económicas. Implicitamente, as narrativas contêm certos pressupostos tradicionais da Economia, seja acerca da racionalidade individual, seja acerca do que é uma boa explicação em teoria económica. Por exemplo, a concepção embutida de racionalidade exclui que possam ser pesadas questões morais, uma vez que elas prejudicariam o raciocínio de optimização da utilidade individual de cada jogador. Ainda por exemplo, o requisito de que a solução do jogo seja um equilíbrio exclui, no caso do DP, que a solução seja a cooperação, já que nesse caso qualquer dos jogadores tem um incentivo para trair. Explicitamente, as narrativas introduzem certas regras do jogo, por exemplo estipulando que não há concertação possível entre os jogadores ou que as jogadas (as escolhas) dos diferentes jogadores são feitas simultaneamente. Acresce que são as narrativas que permitem distinguir as situações a que se aplica ou não um determinado jogo ou solução (por exemplo, no DP original não há repetição, é um encontro único entre dois jogadores).

O que está aqui em causa, como esteve antes com a análise da teoria da verdade de Tarski, é a tentativa de esconder as fontes de significado. O que aqui muitos teóricos da Economia tratam de escamotear, as narrativas, é mais uma vez o que fornece a interpretação à construção formal. Mas isso não é exactamente uma novidade. Quando, em 1950, Melvin Dresher e Merrill Flood realizaram na Rand Corporation a primeira experiência da situação que viria a ser depois conhecida como o dilema do prisioneiro, John Nash, futuro prémio Nobel, criticou a experiência por haver nela demasiada interacção. O defeito da experiência consistia, segundo Nash, no facto de os jogadores, de facto, estarem num jogo de múltiplas jogadas e não numa sequência de jogos de uma única jogada cada – o resultado de estarem sempre a jogar contra os mesmos era que se criava uma reputação. A alternativa proposta por Nash era que os jogadores fossem sempre rodando e que nunca fosse possível a um jogador saber como tinha jogado nas rondas anteriores o seu actual oponente. Esse seria apenas um momento do esforço de depuração formalista da economia experimental envolvendo modelos da teoria dos jogos[18].

[17] Fudenberg e Tirole, 1991:4.

[18] Roth, 1995:8-13.

Importa sublinhar que a operação de separar a construção formal das fontes de significado (narrativas) pode ter efeitos importantes. Um exemplo[19] é o contributo da TJ para a forma de pensar as relações internacionais que foi típica da Guerra Fria. Jogos como o DP deram uma aparência racional à corrida ao armamento nuclear, como se a guerra não fosse mais do que um elemento de um exercício intelectual, altamente matematizado, de estratégia. O DP, entendido como modelo do conflito entre o bloco comunista e as democracias ocidentais, representava o dilema entre a opção de cooperar e evitar a guerra ou a opção de lançar primeiro um ataque nuclear e arriscar a destruição mútua total. O ponto é que essa representação da situação pelo DP restringia o campo das possibilidades, desqualificando como irracional a possibilidade de cooperação, anulando as tentativas para imaginar alternativas na política internacional do mundo real. (Lembrar que, no DP, a solução de cooperação mútua não é um equilíbrio de Nash, uma vez que contém incentivos aos jogadores para trair e assim melhorar o seu pagamento.) Afinal, existiam as alternativas e elas acabaram mesmo com a Guerra Fria. Ora, o que estava subjacente àquela desqualificação da possibilidade de cooperar era o uso da autoridade da matemática (da TJ) na atribuição de um tipo específico de racionalidade aos agentes.

Esta análise é completada por Grüne-Yanoff e Schweinzer (2008), sublinhando que as narrativas desempenham em exclusivo certas funções essenciais na TJ. Primeiro, fornecem uma interpretação aos termos da estrutura, dando conteúdo a essa forma (quem são os jogadores, em que consistem as suas escolhas, como avaliam as consequências das decisões, em que consistem os pagamentos). Segundo, integram os termos interpretados no contexto de uma situação estratégica (qual é a situação inicial, o que aconteceu antes e está acontecer, o que está em causa). Só sobre essa estrutura interpretada é que faz sentido identificar a solução e avaliá-la: verificar se é relevante. Estes autores vão, contudo, mais longe: as narrativas, além de indispensáveis à interpretação das estruturas, são também encaradas como necessárias à aplicação da teoria à estrutura. A TJ contém uma "teoria propriamente dita", enquanto disciplina matemática, a que cabe, designadamente, o desenvolvimento de conceitos de equilíbrio, a demonstração de teoremas, a construção de provas de existência (por exemplo, provar que em qualquer jogo estratégico simétrico, para dois jogadores, no qual cada jogador tem duas estratégias puras e os pagamentos para os quatro perfis estratégicos são diferentes, há uma estratégia mista que é uma estratégia evolucionariamente estável). Ora, a ligação da teoria matemática ao modelo, e por via desta a situações económicas reais ou imaginárias, depende também das narrativas. O ponto é que, para um dado tipo de jogo, a teoria fornece um menu de conceitos para soluções, mas é necessário recorrer a razões extra-teóricas para seleccionar a solução cuja aplicação é viável numa determinada situação – e essas razões extra-teóricas virão das narrativas. Exemplifiquemos.

O conceito-solução mais simples é "eliminar todas as estratégias dominadas" (EED). Do ponto de vista de um jogador, uma estratégia é estritamente dominada quando há sempre, qualquer que seja a jogada do oponente, uma outra estratégia que garante um pagamento melhor. Então, o conceito EED pode oferecer uma solução, mas só será uma solução efectiva se forem preenchidas certas condições. Designadamente, essa solução

[19] Morgan, 2007:157-160.

está indisponível se algum dos jogadores não conhecer todas as estratégias que pode seguir, ou não conhecer as consequências de cada uma dessas estratégias, ou não for capaz de determinar as suas preferências relativamente a essas consequências. Para saber se é este o caso, tem de recorrer-se à narrativa.

Para a maior parte dos jogos este tipo de solução não está disponível. Um conceito-solução mais sofisticado é o equilíbrio de Nash, que corresponde a um perfil de estratégias que, para cada jogador, representa a estratégia óptima, de tal modo que nenhum jogador tem vantagem em modificar a sua opção. Mas também há situações em que o encaminhamento do jogo para um equilíbrio de Nash depende de elementos de contexto que só podem ser fornecidos pela narrativa. Um caso interessante é quando um jogo tem dois equilíbrios de Nash, como o que é representado na seguinte matriz.

	C1	C2
L1	1 , 1	0 , 0
L2	0 , 0	1 , 1

Não há nada na matriz que indique aos jogadores se devem tentar encontrar-se na solução (L1, C1) ou na solução (L2, C2), apesar de elas serem perfeitamente equivalentes e representarem ambas equilíbrios. A existir alguma pista para uma solução, ela teria, mais uma vez, de estar na narrativa. O caso é que há muito tempo que foi assinalado por certos autores que existem certas situações em que, apesar de a estrutura formal do jogo não dar pistas para a sua solução, os jogadores encontram maneiras de se coordenarem com base em elementos que extravasam o formalismo.

Thomas Schelling realizou em 1957 experiências que mostravam isso mesmo. Numa delas, três indivíduos entram num jogo para tentar ganhar uma certa quantia em dinheiro. Os jogadores são designados pelas letras A, B, C. A jogada, a realizar separadamente e sem comunicação, consiste em apresentar as letras que designam os jogadores numa sequência qualquer. Se todos propuserem a mesma sequência, um prémio de montante x será distribuído por todos da seguinte maneira: $\frac{1}{2}x$ para o jogador cuja letra apareça na primeira posição, $\frac{1}{3}x$ para o jogador cuja letra apareça na segunda posição, $\frac{1}{6}x$ para o jogador cuja letra apareça na terceira posição. Se nem todos propuserem a mesma sequência, ninguém recebe nada. Obviamente, cada um ganharia o máximo possível se a lista resultante tivesse à cabeça o seu próprio "nome". Contudo, dos 40 indivíduos que realizaram a experiência, 33 propuseram a sequência ABC. Dos 40, só 12 tinham a letra A. O que está aqui em causa, para Schelling, é que os jogadores encontram, fora da estrutura formal do problema, a partir da sua cultura partilhada, uma maneira de se coordenarem para alcançar um certo objectivo, enquanto a teoria dos jogos, pelo seu formalismo, frequentemente ignora esses factores[20].

Ora, o que os procedimentos metodológicos aqui expostos mostram é que certas linhas de investigação em Economia, fazendo um uso intensivo da TJ, trabalham com o mesmo mecanismo que suporta a ilusão constitutiva das ciências do artificial: produzir a invisibilidade da interpretação. Este uso é complementar do que tínhamos anteriormente detectado: a IA atribui significado genuíno à relação de certas máquinas com o mundo, quando esse significado é um artefacto do observador humano; a Economia

[20] Roth, 1995:8-13.

formalista, por obra da TJ, ao escamotear as interpretações que dão significado às suas estruturas formais, e as raízes dessas interpretações, oculta o horizonte de sentido (humano, social) que deveria estar bem à vista numa ciência das sociedades humanas como se pensava ser a Economia.

A invisibilidade da interpretação, como mecanismo fundador de uma ilusão acerca do lugar do humano na configuração do seu mundo de sentido, junta certas abordagens da Economia ao arquipélago das ciências do artificial. Provavelmente com os mesmos custos de incompreensão do mundo que essa ilusão trouxe às ciências do artificial no último meio século. Porque pode ser tão contraproducente confundir as máquinas com os humanos como confundir os humanos com as máquinas.

Referências

(Brooks 1999) BROOKS, R., ***Cambrian Intelligence: the Early History of the New AI***, Cambridge: MA, The MIT Press

(Osborne e Rubinstein, 1994) OSBORNE; Martin J., e RUBINSTEIN, Ariel, A ***Course in Game Theory***, Cambridge, Massachusetts, The MIT Press

(Fodor 1978) FODOR, Jerry A., "Tom Swift and His Procedural Grandmother", *in* FODOR, Jerry A., *Representations*, Brighton:Sussex, The Harvester Press, 1981, pp. 204-224

(Fudenberg e Tirole 1991) FUDENBERG, Drew, e TIROLE, Jean, *Game Theory*, Cambridge: Massachusetts, The MIT Press

(Grüne-Yanoff e Schweinzer, 2008) GRÜNE-YANOFF, Till, e SCHWEINZER, Paul, "The role of stories in applying game theory", in *Journal of Economic Methodology*, 15 (2), pp. 131-146

(Harnad 1989) HARNAD, Stevan, "Minds, Machines and Searle", in *Journal of Theoretical and Experimental Artificial Intelligence*, 1, pp. 5-25

(Harnad 1990) HARNAD, Stevan, "The Symbol Grounding Problem", in *Physica* D, 42, pp. 335-346

(Harnad 2002) HARNAD, Stevan, "Symbol Grounding and the Origin of Language", *in* Matthias SCHEUTZ (ed.), *Computationalism: New Directions*, Cambridge: Massachusetts, The MIT Press, pp. 143-158

(Haugeland 1985) HAUGELAND, John, *Artificial Intelligence: The Very Idea*, Cambridge: Massachusetts, The MIT Press

(Millikan 1984) MILLIKAN, Ruth G., *Language, Thought, and Other Biological Categories*, Cambridge: MA, MIT Press

(Morgan 2007) MORGAN, Mary S., "The Curious Case of the Prisoner's Dilemma: Model Situation? Exemplary Narrative?", in A. CREAGER, E. LUNBECK, e M. NORTON WISE (eds.), Science Without Laws, Durham, Duke University Press, pp. 157-185

(Newell 1980) NEWELL, Allen, "Physical Symbol Systems", in *Cognitive Science*, 4, pp. 135-183

(Newell e Simon 1976) NEWELL, Allen, e SIMON, Herbert A., "Computer Science as Empirical Inquiry: Symbols and Search", in *Communications of the Association for Computing Machinery*, 19 (3), pp. 113-126

(Nolfi e Floreano 2000) NOLFI, Stefano, e FLOREANO, Dario, *Evolutionary Robotics*, Cambridge: MA, MIT Press

(Putnam 1960) PUTNAM, Hilary, "Minds and Machines", (republicação in PUTNAM, *Mind, Language and Reality*, Cambridge, CUP, 1975, pp. 362-385)

(Roth 1995) ROTH, Alvin E., "Introduction to Experimental Economics", in KAGEL, John H.,e ROTH, Alvin E. (eds.), *Handbook of Experimental Economics*, Princeton, Princeton University Press, 1995, pp. 3-109

(Rubinstein 1990) RUBINSTEIN, A., "New Directions in Economic Theory – Bounded Rationality", in *Revista Española de Economia*, 7, 3-15

(Steels 2003) STEELS, Luc, "Intelligence with representation", in *Philosophical Transactions of the Royal Society (Mathematical, Physical and Engineering Sciences)*, 361 (1811), pp. 2381-2395

(Santos 2003) SANTOS, Ricardo, *A Verdade de um Ponto de Vista Lógico-Semântico*, Lisboa, Fundação Calouste Gulbenkian e Fundação para a Ciência e a Tecnologia, 2003

Porfírio Silva[1]
IST, Lisboa

[1] Instituto de Sistemas e Robótica (Pólo do Instituto Superior Técnico, Lisboa). Bolseiro de Pós-Doutoramento da Fundação para a Ciência e a Tecnologia.