GUAIRACÁ
REVISTA DE
FILOSOFIA

ALÉM DO TESTE DE TURING: EM BUSCA DE UMA DEFINIÇÃO RAZOÁVEL E TESTÁVEL DE CONSCIÊNCIA

CARLOS LUCIANO MONTAGNOLI[1]

Resumo: Neste artigo, discutimos a questão da possibilidade de se produzir uma inteligência artificial. Começamos chamando a atenção para o fato de que uma solução para tal questão deve, necessariamente, passar pela obtenção de uma definição de inteligência. Passamos, então, a considerar definições possíveis de inteligência, e, ao fazermos isso, distinguimos, como costumeiramente faz a literatura acerca desse assunto (com exceção, talvez, para o que chamamos de IA média), três tipos de inteligência artificial (IA), que chamamos, respectivamente, de IA fraca, IA média e IA forte. Observamos, na sequência, que a solução da questão da possibilidade é trivial no caso da IA fraca, e nos concentramos nos casos da IA média e da IA forte. Fazemos, então, uma análise crítica de possíveis definições de autoconsciência, com base em alguns critérios, como testabilidade e intuitividade. Encerramos propondo uma definição de autoconsciência que consideramos capaz de suportar a crítica com base nos critérios mencionados, e, empregando tal definição, concluímos que os conceitos de IA média e IA forte se confundem, e que a questão da possibilidade deve permanecer em aberto sempre, ou até que seja produzido um agente artificial capaz de satisfazer a definição de autoconsciência que propusemos.

Palavras-chave: inteligência artificial, autoconsciência, possibilidade, testabilidade, intuitividade

1. Universidade Estadual de Londrina. Email: montagnoli@uel.br

DOI 10.5935/2179-9180.20180001

BEYOND THE TURING TEST: AN ATTEMPT OF OBTAINING A REASONABLE AND TESTABLE DEFINITION OF CONSCIOUSNESS.

Abstract: In this paper we shall discuss the problem of the possibility of creating an artificial intelligence. We begin considering that a solution to such problem requires a satisfactory definition of intelligence. We then assess a bunch of possible definitions of intelligence and, while doing so, we distinguish three types of artificial intelligence (AI), viz., weak AI, medium AI, and strong AI (such distinction is quite common in the literature concerning AI, with the possible exception of medium AI). After that we observe that the solution of the possibility problem is trivial in the case of weak AI, and we so focus on the cases of medium AI and strong AI. We then make a critical analysis of some possible definitions of self-awareness, using some criteria like testability and intuitiveness. At last we propose a definition of self-awareness which we believe able to stand up to criticism based in the mentioned criteria, and we conclude, using the definition in question, that the notions of medium AI and strong AI converge, and that the possibility problem must either remain open forever, or until someone creates an artificial intelligence capable of satisfying our definition of self-awareness.

Keywords: artificial intelligence, self-awareness, possibility, testability, intuitiveness.

INTRODUÇÃO

A inteligência artificial (IA), nas últimas décadas, tem sido um tema recorrente no cinema, bem como em romances de ficção científica. Contudo, como não tem sido raro na história, o que é ficção científica em um momento às vezes se torna fato científico em um momento posterior. Como estamos vivendo uma situação desse tipo no que atine à IA, não é surpresa que ela tenha se tornado alvo de extensa discussão filosófica.

Há, de fato, um vasto leque de problemas filosóficos relacionados à IA, muitos deles explorados pelo cinema, com ênfase em problemas éticos, como o dos direitos que teria uma

ISSN 2179-9180
GUAIRACÁ REVISTA DE FILOSOFIA, GUARAPUAVA-PR, V34, N1, P. 1-14, 2018

máquina ou programa que suspeitássemos possuir consciência. Neste artigo, pretendemos discutir um único problema filosófico relacionado à IA, que num certo sentido precede todos os demais, qual seja: é mesmo possível produzir uma IA?

Agora, é evidente que, se vamos nos questionar sobre se é possível produzir inteligência artificialmente, devemos, antes de mais nada, esclarecer o que estamos entendendo por inteligência.

IA FRACA

O que é inteligência? Para os nossos propósitos, poderíamos muito bem definir inteligência simplesmente como habilidade de resolver problemas. Nesse caso, pode ser interessante adotarmos uma definição de inteligência relativizada a determinados conjuntos de problemas. Isso porque, como é óbvio, um mesmo agente (pessoa, animal, máquina ou programa) pode ter uma grande habilidade para resolver determinados problemas, e pouca ou nenhuma habilidade para resolver outros. Por exemplo, eu posso ser muito bom para jogar xadrez, e não possuir habilidade nenhuma para consertar um carro com problemas mecânicos. Nesse caso, podemos dizer[2] que sou bastante inteligente para jogar xadrez, mas nem um pouco para consertar um carro enguiçado.

Um agente artificial (máquina ou programa) capaz de resolver problemas específicos é o que se costuma chamar de IA fraca. Muito bem, se nos perguntarmos se é possível criar uma IA fraca, a resposta é obviamente positiva, porque já foram criados diversos agentes artificiais capazes de resolver problemas específicos com um nível de performance igual ou superior àquele humano. Um dos exemplos mais célebres é o do programa enxadrista *Deep Blue*, que, em 1996, venceu em um torneio de xadrez o então mestre mundial Garry Kasparov. Entretanto, há uma série de outros exemplos, entre os quais se destacam os assim chamados sistemas especialistas, que são programas dotados, normalmente, de bancos de dados com o conhecimento de uma determinada área, como a medicina, de raciocínio lógico, e de habilidades de processamento de linguagem natural. Um exemplo é o programa *Watson*, da IBM, que, em alguns casos, mostrou capacidade de diagnóstico de enfermidades ao menos equivalente à de médicos experientes.

Então a IA já é uma realidade? Bem, a IA fraca certamente sim. No entanto, parece que não estamos dispostos a considerar *Deep Blue* ou *Watson* como entidades inteligentes. Isso significa que nossa definição de inteligência como habilidade de resolver problemas específicos não captura corretamente nossos usos da palavra 'inteligência'. De fato, parece que costumamos dizer que há tipos de inteligência, mas não que alguém seja muito inteligente para jogar xadrez e nada inteligente para consertar carros. Parece que essa nossa definição de inteligência tem mais a ver

2. De um modo que não faz muita justiça ao nosso uso comum da palavra 'inteligência'. Vamos falar mais sobre isso mais adiante.

ISSN 2179-9180
GUAIRACÁ REVISTA DE FILOSOFIA, GUARAPUAVA-PR, V34, N1, P. 1-14, 2018

com o conceito de conhecimento que com aquele de inteligência, pois, no caso de nosso exemplo, poderíamos dizer que alguém é inteligente ou não, embora possa ter conhecimentos em determinadas áreas, como jogar xadrez, mas não em outras, como consertar carros.

IA MÉDIA

Nesse sentido, poderíamos substituir nossa definição de inteligência como habilidade para resolver problemas específicos, por uma outra que considere uma habilidade geral, por assim dizer, para resolução de problemas. Para tornar mais claro do que estamos falando, vamos considerar a discussão que Descartes faz, no seu 'Discurso do método', sobre a possibilidade de se construir uma máquina que simule de forma convincente um ser humano[3].

De acordo com Descartes, não é impossível construir uma máquina que simule convincentemente o comportamento, por exemplo, de um macaco. Se uma tal máquina for construída de modo a, inclusive, apresentar a aparência externa de um macaco, Descartes considera que é perfeitamente plausível que, postos diante da máquina em questão, pensássemos se tratar de fato de um macaco. Mas seria possível fazer o mesmo com um ser humano? Descartes acredita que não. Eis as razões.

Descartes considera que alguém poderia sem problemas criar uma máquina com a aparência externa de um ser humano, e dispor suas peças de forma tal que a máquina pudesse responder a determinados estímulos da forma que um ser humano poderia normalmente responder. Por exemplo, a máquina, se tocada em determinada parte, poderia perguntar o que se quer com ela[4].

Não obstante, Descartes acredita que é impossível construir uma máquina que simule o comportamento humano de forma convincente em *qualquer* circunstância. Atendo-se apenas ao comportamento verbal humano, Descartes considera que para cada resposta que se quisesse que a máquina desse a um determinado estímulo, seria necessário dispor as peças da máquina de um modo específico. E o fato, segundo Descartes, é que simplesmente não seria possível dispor as peças da máquina de tão variadas maneiras quantas seriam necessárias para que a mesma pudesse dar uma resposta adequada a qualquer estímulo que recebesse. Obviamente, se considerarmos que os estímulos sejam as sentenças de um interlocutor, então, como há infinitas sentenças que podem ser ditas usando-se uma língua natural qualquer, as peças da máquina deveriam ser dispostas de infinitas maneiras diferentes, de

3. Cf. DESCARTES, 1989, pp. 75-76.
4. Cf. DESCARTES, 1989, p. 75.

ISSN 2179-9180
GUAIRACÁ REVISTA DE FILOSOFIA, GUARAPUAVA-PR, V34, N1, P. 1-14, 2018

modo que a mesma estivesse programada para dar uma resposta adequada a uma sentença *arbitrária* de seu interlocutor.

É claro que, nesse argumento de Descartes, estão envolvidas duas pressuposições. A primeira de que só se pode programar o comportamento de uma máquina por meio da disposição de suas peças de formas específicas. E a segunda, de que nós humanos somos capazes de responder de forma adequada, infinitamente variada, a uma quantidade infinita de estímulos diversos. É bem possível que nosso conhecimento pessoal nos coloque em condição de responder satisfatoriamente a um número finito, embora talvez muito grande, de sentenças de um interlocutor, e que, para qualquer sentença que não pertença a esse conjunto finito de sentenças, devolvamos uma resposta *default*, ou uma resposta pertencente a um certo conjunto de respostas *default*, como, por exemplo, 'não sei do que se trata', ou 'não entendi o que você disse'. Poder-se-ia dizer que uma tal resposta seria adequada para Descartes, porque ele estava pensando em simulações convincentes do comportamento humano, de forma que qualquer resposta tipicamente humana para um dado estímulo seria por definição adequada. Mas, se isso é assim, usando um computador moderno e uma linguagem de programação como C, por exemplo, não há nenhum problema, de fato, em se desenvolver um programa que responda de formas específicas a um grande número de sentenças de um interlocutor, e que responda com respostas *default* a sentenças do interlocutor que não pertençam a tal conjunto de sentenças.

Então essa nova definição de inteligência é adequada? Podemos considerar como inteligente um agente que apresente uma habilidade geral para resolver problemas, ou que ao menos se comporte de forma tipicamente humana diante de circunstâncias arbitrárias? Há quem defenda que sim, e há quem defenda que não. Se, ao falar em inteligência, estivermos pensando que isso inclui a autoconsciência, então podemos considerar dois casos célebres: o de Alan Turing, defendendo uma resposta positiva à nossa questão acima, e o de John Searle, defendendo uma resposta negativa à mesma questão.

IA FORTE

O ponto aqui é que, de um ponto de vista intuitivo, inteligência e consciência são coisas diferentes. Intuitivamente, podemos considerar uma entidade que fosse bastante inteligente, ou seja, bastante hábil para resolver problemas, mesmo genericamente, sem, no entanto, ser autoconsciente. Mas, nesse caso, o que, então, é a autoconsciência? Ainda de um ponto de vista intuitivo, podemos definir uma entidade autoconsciente como sendo um agente que sabe de sua própria existência,

ISSN 2179-9180
GUAIRACÁ REVISTA DE FILOSOFIA, GUARAPUAVA-PR, V34, N1, P. 1-14, 2018

e que, quando está realizando uma dada ação, sabe que está ali realizando aquela ação. Essa é uma definição certamente intuitiva de autoconsciência, mas ela padece de um problema grave: não é testável.

De fato, suponha que houvesse um prêmio para o primeiro indivíduo ou grupo de indivíduos que construísse um agente artificial autoconsciente, e que você estivesse no comitê encarregado de verificar se requisitantes do prêmio de fato fazem jus a ele, ou seja, se de fato construíram o tal agente autoconsciente. Como você faria para decidir se uma máquina ou programa apresentado por um requisitante do prêmio é ou não autoconsciente, dada a definição de autoconsciência que propusemos acima? Obviamente, não bastaria perguntar à suposta IA se ela sabe que existe, e se sabe que está sendo avaliada por você, porque é bastante simples programar um *chatbot* para responder afirmativamente a essas questões. Então uma resposta sim a tais questões não é uma condição suficiente para a autoconsciência. Aliás, não é sequer uma condição necessária, já que parece perfeitamente possível que uma entidade possa ser autoconsciente no sentido especificado, sem apresentar as habilidades linguísticas necessárias para responder afirmativamente às questões sob consideração. Por exemplo, não negaríamos que uma pessoa com afasia estivesse consciente porque não consegue responder a tais questões.

Diante desse problema, Turing idealizou seu famoso teste. O teste de Turing serve precisamente para se determinar se um dado agente artificial deve ou não ser considerado autoconsciente. A versão original do teste possui algumas peculiaridades, que não vamos considerar aqui. A grosso modo, o teste de Turing pode ser descrito da seguinte maneira. Colocamos, em seis salas distintas, cinco pessoas e uma máquina, de modo que em cada sala ficará uma pessoa ou a máquina. Se em vez de uma máquina tivermos um programa, não interessa onde ficará o hardware que vai rodá-lo. Esses seis agentes vão conversar, em um ambiente virtual, sobre qualquer assunto, durante um tempo pré-determinado, digamos, durante uma hora. Cada uma das cinco pessoas saberá desde o início que um dos seus cinco interlocutores será um agente artificial, mas não saberá qual deles é o tal agente. Após a conversa, cada uma das cinco pessoas deverá apontar qual dos seus cinco interlocutores ele acredita ser a IA. Caso mais da metade das pessoas, três delas, no caso, erre ao apontar um dos interlocutores como sendo a IA, então o agente artificial passa no teste, e deve ser considerado autoconsciente. Do contrário, o agente não passa no teste e se deve considerar que não possui autoconsciência.

Ora, parece claro que um agente pode passar no teste de Turing sem ter as propriedades que usamos há pouco para definir autoconsciência. O que significa, evidentemente, que o teste de Turing nos dá, de fato, uma nova definição de autoconsciência, essa evidentemente testável, que é a seguinte: uma entidade autoconsciente é qualquer entidade capaz de passar pelo teste de Turing. Essa

ISSN 2179-9180
GUAIRACÁ REVISTA DE FILOSOFIA, GUARAPUAVA-PR, V34, N1, P. 1-14, 2018

definição é razoável? Intuitivamente parece que não é. Há, nos dias atuais, programas capazes de passar no teste de Turing, que não são considerados autoconscientes nem por seus criadores.

Analisando com mais cuidado o problema de se definir autoconsciência, podemos chegar a algumas conclusões interessantes. A definição com a qual começamos era intuitiva, mas não testável. Aliás, ela não era testável não só para agentes artificiais, mas tampouco para pessoas que não nós mesmos. De fato, como não temos acesso direto a nenhuma outra consciência que não a nossa própria, não podemos saber se qualquer outro agente, artificial ou biológico, sabe de sua própria existência, e se, quando está realizando uma ação, sabe que está ali realizando aquela ação. Este é, é claro, o antigo problema filosófico do solipsismo. O que Turing tenta fazer é então substituir essa definição de consciência que, embora seja intuitiva não é testável, por uma outra que o seja. E ele faz isso. Mas ao fazer isso o que ele obtém é uma definição que é testável, porém não intuitiva.

QUESTÕES SEMÂNTICAS

Aqui nós estamos, sem dúvida, diante de uma questão semântica. Se definirmos autoconsciência de modo que a definição seja testável, então, desde que aceitemos tal definição, a questão de se saber se uma entidade é ou não autoconsciente reduz-se a testar se ela satisfaz ou não a definição. O problema do comitê que deve avaliar se os requerentes do prêmio que mencionamos antes fazem jus a ele estará resolvido. Mas como decidimos se vamos aceitar ou não uma definição? Nós podemos julgar a qualidade de uma definição, para além do fato de ela ser testável ou não? Nós certamente podemos fazer isso. É possível, por exemplo, considerar que uma definição é ruim por ser restritiva demais, ou seja, por determinar que não caiam sob o conceito definido objetos que deveriam cair sob tal conceito. Ou podemos considerar que ela é ruim por ser concessiva demais, ou seja, por permitir que caiam sob o conceito definido objetos que não deveriam cair sob aquele conceito. Mas como determinar, de antemão, se um objeto deve ou não cair sob um dado conceito? Parece haver aqui uma certa arbitrariedade inevitável. Se vamos definir o que é uma teoria científica, podemos considerar que é restritiva demais uma definição com base na qual a mecânica quântica não deva ser considerada científica, e leniente demais uma definição segundo a qual a astrologia deva ser considerada científica. Isso porque consideramos, de antemão, que a mecânica quântica deve cair sob o conceito de teoria científica, mas a astrologia não. Mas podemos especificar critérios não-arbitrários com base nos quais fazemos tais considerações? Parece que tudo o que podemos fazer para regular de algum modo a definição de um conceito é

ISSN 2179-9180
GUAIRACÁ REVISTA DE FILOSOFIA, GUARAPUAVA-PR, V34, N1, P. 1-14, 2018

especificar condições de adequação para boas definições para tal conceito, tal como Tarski fez para o conceito de verdade[5]. Mas, é claro, ainda permaneceria uma certa arbitrariedade inevitável na especificação de tais condições de adequação.

De todo modo, o fato é que houve quem considerasse a definição de consciência que resulta do teste de Turing insatisfatória, por ser concessiva demais. Uma dessas pessoas é J. Searle, que desenvolveu o famoso argumento da sala chinesa na tentativa de evidenciar essa inadequação. Vamos considerar um pouco esse argumento.

Imagine que uma pessoa p1, que não tem nenhum conhecimento da língua chinesa, é trancada dentro de uma sala. Nessa sala, se encontra um número infinito de caixas numeradas. Nessas caixas, por sua vez, se encontram bilhetes, também numerados, contendo sentenças em chinês. Qualquer sentença da língua chinesa pode ser encontrada em um bilhete, que por seu turno está em uma dessas caixas. Assim, podemos nos referir a uma sentença arbitrária da língua chinesa como sendo a sentença que está escrita no bilhete de número x, da caixa de número y, dado algum par x, y de números naturais. Além dessas caixas, encontra-se na sala também um manual, com um número infinito de páginas, contendo todas as sentenças da língua chinesa e, ao lado de cada uma delas, dois números naturais, o primeiro indicando uma das caixas, e o segundo o número de um bilhete na caixa em questão. Todas as entradas do manual têm, portanto, a forma $s - x - y$, onde s é uma sentença em chinês, e x e y são números naturais. Cada entrada deve ser entendida do modo seguinte: a sentença s é tal que a sentença que está escrita no bilhete de número y da caixa de número x é uma resposta, ou um comentário[6], que um falante competente do chinês poderia dar ou fazer para a tal sentença s.

Pois bem, do lado de fora da sala, um falante competente p2 do chinês conduz uma conversa, em chinês, com p1, da seguinte maneira: p2 escreve um bilhete em chinês, e passa para p1 por baixo da porta da sala; p1 pega o bilhete de p2, e vai ao manual. Usando mera inspeção visual dos caracteres chineses do bilhete de p2, p1 procura no manual até encontrar aquela mesma sequência de caracteres. Isso vai ocorrer eventualmente, já que o bilhete de p2 contém uma sentença em chinês, e todas as sentenças em chinês estão listadas no manual. Ao encontrar a sentença de p2 no manual, p1 memoriza os números x e y que se encontram ao lado de tal sentença. Em seguida, p1 vai até a caixa de número x, e retira dali o bilhete de número y. Então p1 passa esse bilhete por baixo da porta para p2, que tem a impressão de que p1 entendeu o que ele havia escrito no bilhete inicial, e então escreve outro bilhete, e continua a conversa passando-o por debaixo da porta a p1, que repete o processo.

5. Cf. TARSKI, 1983.
6. Ou, enfim, uma continuação adequada de uma conversa.

ISSN 2179-9180
GUAIRACÁ REVISTA DE FILOSOFIA, GUARAPUAVA-PR, V34, N1, P. 1-14, 2018

Obviamente, a conclusão que Searle deseja tirar desse experimento mental é que, embora p1 simule de modo convincente para p2 o comportamento de um falante competente do chinês, p1 não têm, como sabemos, nenhum conhecimento da língua chinesa. Da mesma forma, pensa Searle, um agente artificial poderia simular o comportamento verbal humano de forma extremamente convincente, sem de fato entender nada daquilo que ele diz, e tampouco daquilo que lhe é dito. Dizendo a mesma coisa de outra forma, o agente artificial poderia estar simplesmente manipulando símbolos sem dar nenhum significado a esses símbolos.

Considerando o argumento da sala chinesa, somos tentados a concluir que uma definição de autoconsciência baseada no teste de Turing seja mesmo demasiado concessiva: ela permite que sejam considerados como autoconscientes agentes artificiais que não passam de *simulações* do comportamento de entidades autoconscientes (os humanos, no caso), mas que não possuem, de fato, consciência de si. Mas se a definição de autoconsciência baseada no teste de Turing é concessiva demais, que outra definição, que seja testável, poderíamos propor em seu lugar? Searle não nos dá uma tal definição.

Convém aqui insistir no fato, que mencionamos há pouco, de que não há como formular algo como *a* definição correta de autoconsciência. Como dissemos, qualquer definição que formularmos para autoconsciência, e, aliás, para qualquer outro conceito, terá uma certa arbitrariedade inevitável. O que podemos fazer é *convencionar* algumas condições de adequação que julguemos imprescindíveis para uma boa definição de autoconsciência, e então procurar obter uma definição que satisfaça tais condições. Uma delas parece ser ponto pacífico. Por uma questão de ordem evidentemente pragmática, a definição deve ser testável, ou seja, para um objeto x qualquer, deve ser possível decidir, por meio de algum teste, se x cai ou não sob o conceito que foi definido. Outra condição que poderíamos adotar é que a definição fosse formulada de modo a capturar, tanto quanto possível, as intuições que temos relacionadas ao conceito que está sendo definido. Ou seja, deveríamos formular a definição de autoconsciência, por exemplo, de modo tal que, dada tal definição, devessem ser considerados autoconscientes, tanto quanto possível, aqueles agentes que intuitivamente consideraríamos como tais. Esse critério teria que ser necessariamente frouxo, até porque podemos ter intuições conflitantes sobre se um dado objeto cai ou não sob um dado conceito, isto é, podemos ter intuições que nos levam a crer que isso seja o caso, e outras que nos levam à direção oposta[7].

7. Para ilustrar, vamos utilizar um exemplo oriundo da filosofia da matemática: quando nos perguntamos qual conjunto tem mais elementos, se o dos números naturais ou o dos números reais, a intuição de que dois conjuntos infinitos devem ter a mesma quantidade de elementos, a saber, uma infinidade deles, nos leva a crer que ambos tenham a mesma quantidade de elementos. Mas a intuição de que o todo é sempre maior que a parte nos leva a crer que o conjunto dos reais deva ter mais elementos, já que ele contém todos os naturais e mais os negativos, as frações e os irracionais.

ISSN 2179-9180
GUAIRACÁ REVISTA DE FILOSOFIA, GUARAPUAVA-PR, V34, N1, P. 1-14, 2018

ALÉM DO TESTE DE TURING

Estivemos aqui falando de IA em pelo menos três sentidos diferentes: a) IA como um agente artificial com uma capacidade para resolver problemas de um conjunto específico de problemas; b) IA como um agente artificial com uma capacidade genérica para resolver problemas, do tipo que, supostamente, nós humanos temos; para evitarmos falar, de um modo um tanto impreciso, sobre essa capacidade genérica de resolução de problemas, poderíamos falar em um agente artificial capaz de simular de modo convincente o comportamento humano; c) IA como um agente artificial autoconsciente. Como chamamos a IA do tipo a) de IA fraca, podemos chamar a IA do tipo b) de IA média, e a IA do tipo c) de IA forte (como é costume da literatura da área, nesse último caso). Obviamente, quem defende uma definição de autoconsciência baseada no teste de Turing vai identificar IA média e IA forte.

Quando tentamos pensar intuitivamente sobre se devemos ou não identificar IA forte e IA média, parece que nos deparamos com um caso de intuições conflitantes. Se considerarmos o argumento da sala chinesa, somos levados a crer que os dois conceitos em questão não devem ser identificados. No entanto, se considerarmos a questão do solipsismo, e o motivo que nos leva a crer que outros humanos são autoconscientes, percebemos que o fazemos porque eles se comportam do modo que consideramos típico de entidades autoconscientes. Eu tenho acesso direto à minha própria consciência apenas, e sei que sou autoconsciente naquele sentido de que sei que existo, e sei, quando estou realizando uma ação, que estou ali realizando tal ação. Mas não tenho esse acesso à consciência das outras pessoas. Então por que julgo que elas são autoconscientes? Talvez porque tomo a mim mesmo como uma espécie de protótipo de um agente autoconsciente, e assim estabeleço, com base no meu comportamento, qual é o comportamento *típico* de um agente autoconsciente. Como as outras pessoas apresentam esse comportamento típico, considero por isso que elas são autoconscientes. Mas, se isso é assim, que motivo eu teria, adotando a atitude sugerida pelo argumento da sala chinesa, para negar aos agentes artificiais aquilo que concedo para as outras pessoas?

Em casos como esse, não há como não desrespeitar alguma de nossas intuições, e a única maneira de decidir entre elas será definir o conceito envolvido, e verificar o que resulta quando *assumimos* tal definição. No caso do nosso exemplo, se assumimos as definições cantorianas de igualdade e ordem entre os cardinais de dois conjuntos, somos levados a concluir, primeiro que o conjunto dos naturais não tem a mesma quantidade de elementos que o conjunto dos reais, em desacordo com a primeira das intuições mencionadas acima, e depois que o conjunto dos reais tem mais elementos que o dos naturais, coisa que, diga-se de passagem, não vamos concluir com base na segunda intuição mencionada (que não se aplicará, dadas as definições assumidas, a conjuntos infinitos).

ISSN 2179-9180
GUAIRACÁ REVISTA DE FILOSOFIA, GUARAPUAVA-PR, V34, N1, P. 1-14, 2018

Assim, parece que temos intuições que nos levam a crer que um agente que passe pelo teste de Turing *deve* ser considerado como uma entidade autoconsciente, e outras que nos levam a crer que ele *não deve* ser considerado assim. Como dirimimos esse conflito?

Vamos considerar com mais cuidado a indagação que nos fizemos mais acima: por que negaríamos aos agentes artificiais aquilo que concedemos às outras pessoas, ou seja, considerá-las autoconscientes ou não com base em seu comportamento? Parece que não temos nenhum motivo razoável para fazer isso. Mas, sendo assim, vamos concluir que é o caso de aceitar como adequada a definição de consciência baseada no teste de Turing? Mas já havíamos considerado que já há nos dias de hoje *chatbots* capazes de passar por tais testes, e não parece nada razoável considerar que tais *chatbots* sejam agentes autoconscientes.

Podemos tentar desfazer esse aparente impasse, nos perguntando porque tendemos a não considerar esses *chatbots* como entidades autoconscientes. Talvez a razão disso resida no fato de que, embora apresentem um *comportamento verbal* semelhante ao nosso, esses *chatbots* não se assemelham a nós nos demais aspectos do comportamento. Por exemplo, os *chatbots* não possuem nenhuma agenda própria. Eles não têm objetivos que desejam alcançar, não fazem planos sobre como vão alcançar tais objetivos, e não iniciam nenhuma ação voltada ao alcance dos mesmos. A propósito, os *chatbots*, quando não recebem os *inputs* dos usuários, não iniciam qualquer ação que seja. Eles não executam nenhuma ação por *própria iniciativa*.

Desse modo, podemos pensar que o que há de errado com o teste de Turing, e com uma definição de autoconsciência baseada nele, é sua incompletude: ele considera apenas um aspecto do comportamento, a saber, o comportamento verbal. Nesse caso, ao negarmos a autoconsciência aos agentes artificiais capazes de passar pelo teste de Turing, não estamos, de fato, negando a eles o que concedemos a outras pessoas, porque nós não consideramos as outras pessoas autoconscientes apenas em função de seu comportamento verbal, mas sim em função de diversos aspectos do comportamento humano.

Então poderíamos propor uma definição de autoconsciência baseada em uma espécie de teste de Turing completo, de modo que consideremos como autoconscientes aqueles agentes que apresentam um comportamento indistinguível do comportamento humano, mas nos mais diversos aspectos, e não apenas no que atine ao comportamento verbal.

Mas uma tal definição de autoconsciência seria adequada? Parece que ainda podemos fazer objeções bastante razoáveis à adequação de uma tal definição. Por exemplo, é perfeitamente pensável que uma espécie que tivesse evoluído em um outro planeta pudesse ser constituída de indivíduos autoconscientes (no sentido intuitivo), e até mais desenvolvidos que nós em termos de civilização e tecnologia,

ISSN 2179-9180
GUAIRACÁ REVISTA DE FILOSOFIA, GUARAPUAVA-PR, V34, N1, P. 1-14, 2018

sem, contudo, apresentarem comportamento semelhante ao nosso. Nesse caso, usando a nossa proposta de definição, consideraríamos tais indivíduos como destituídos de autoconsciência, o que não parece nada razoável.

COMPORTAMENTO RACIONAL

Uma maneira de resolver essa dificuldade foi proposta por P. Norwig e S. Russell em seu livro *Artificial intelligence: A modern approach*[8]. A ideia deles consiste em abandonar não a definição de autoconsciência com base em comportamento, mas uma definição com base em *comportamento humano*. Em lugar disso, eles propõem uma definição de autoconsciência com base em um conceito de *comportamento racional*.

Obviamente, alguém poderia aqui pensar que, ao adotarmos uma tal proposta, estaríamos trocando um problema espinhoso – o de definir autoconsciência – por um outro tão ou mais espinhoso – o de definir comportamento racional. No entanto, como estamos falando em *comportamento* racional, há um modo bastante razoável e intuitivo de se definir esse conceito, que é precisamente o que Norwig e Russell adotam. Para falar a esse respeito, começamos com uma descrição estrutural e funcional de um agente qualquer, humano, outro animal qualquer, ou artificial.

Normalmente, um agente é constituído pelos seguintes componentes: a) sensores, mediante os quais ele recebe dados do meio externo; b) processador, empregado pelo agente para: b1) processar os dados brutos recebidos pelos sensores, de modo a formar uma compreensão do estado do meio externo; b2) planejar, com base no programa (item c), as ações que deveriam ser executadas de modo a modificar o meio externo para conformá-lo aos objetivos do agente, estabelecidos no programa; c) programa, ou agenda, que consiste em uma lista mutável dos objetivos que o agente pretende alcançar; d) atuadores, que são os dispositivos mediante os quais o agente pode atuar sobre o meio externo, de modo a modificá-lo de acordo com seus objetivos.

A essa descrição estrutural do agente, podemos acrescentar a seguinte descrição funcional, que é bastante óbvia: o agente possui sua agenda própria, ou seu programa, que contém seus objetivos. Mediante o uso de seus sensores, o agente obtém dados do meio, que são processados e lhe permitem formar um quadro da configuração do mesmo em um dado momento. Se o agente percebe que a configuração do meio naquele momento está em desacordo com seus objetivos, ele planeja ações que possam alterar essa configuração, de modo a adequar o meio

8. Cf. NORWIG & RUSSELL, 2010, pp. 1-5.

ISSN 2179-9180
GUAIRACÁ REVISTA DE FILOSOFIA, GUARAPUAVA-PR, V34, N1, P. 1-14, 2018

a tais objetivos. Esse planejamento é então posto em execução mediante o uso dos atuadores. Conforme os objetivos no programa vão sendo alcançados, ou vão se mostrando inviáveis, o agente pode modificar o programa, de modo mais ou menos arbitrário[9].

Essa descrição estrutural e funcional do agente é bastante simples, mas captura razoavelmente as características básicas de quaisquer agentes. No caso de nós humanos, por exemplo, poderíamos pensar em nossos órgãos sensoriais como nossos sensores, em nossos cérebros como nossos processadores (com o córtex visual, o córtex auditivo, etc., realizando a função b1 acima, e o neocótex realizando a função b2), em nossas agendas individuais e coletivas como nosso programa, e em nossos membros e nossa tecnologia como nossos atuadores. O nosso ponto aqui, no entanto, é que essa descrição estrutural e funcional do agente nos sugere uma definição bastante intuitiva de comportamento racional, que é a seguinte: o agente se comporta de forma racional quando as ações que resultam de seu planejamento maximizam, até onde o agente pode ter conhecimento, a probabilidade de que os objetivos de sua agenda sejam alcançados.

Assim, se o agente tem uma agenda constituída pelo conjunto de objetivos O, mas executa outras ações que não aquelas que, até onde ele sabe, são as que lhe dão maior probabilidade de alcançar os objetivos em O, então ele não se comporta de modo racional. Essa definição de comportamento racional é testável? Não se mantivermos a exigência de que o agente deve executar as ações que ele *acredite* serem as que maximizam as chances de alcançar seus objetivos, já que, é claro, não temos acesso direto às crenças de outros agentes que não nós mesmos. Mas podemos substituí-la por uma outra definição, próxima dela, que é a seguinte: vamos considerar racionais aqueles agentes que aparentam ter sua própria agenda, e executar ações com vistas ao alcance dos objetivos que compõem a agenda em questão.

Dada essa definição de agente racional, podemos considerar que a racionalidade assim definida seja condição suficiente para considerarmos um agente como sendo autoconsciente. Não parece ser o caso de considerarmos que isso seja também uma condição necessária para a autoconsciência, já que, do ponto de vista intuitivo, um agente autoconsciente pode agir de forma irracional. Por exemplo, quando uma pessoa age de forma irracional, contra seus próprios objetivos, nem por isso dizemos que ela deixou de ser autoconsciente. Mas o fato é que, se considerarmos a racionalidade como uma condição suficiente para a autoconsciência, então temos um meio de resolver nosso problema hipotético dos membros do comitê que vai

9. A modificação do programa pode ser feita pelo agente de modo totalmente arbitrário, ou pode ser feita com base em um conjunto, que pode ou não ser imutável, de critérios, que podemos chamar de *princípios*.

ISSN 2179-9180
GUAIRACÁ REVISTA DE FILOSOFIA, GUARAPUAVA-PR, V34, N1, P. 1-14, 2018

julgar as alegações de criação de agentes artificiais autoconscientes, que propusemos mais acima[10]. E essa solução para o tal problema parece ser tão intuitiva quanto a exigência da testabilidade parece permitir.

Suponha-se que adotemos essa definição de autoconsciência baseada no conceito de comportamento racional, e criemos um teste para determinar se um agente se comporta racionalmente e, portanto, se é autoconsciente. Nesse caso, mais uma vez, tal como fez Turing, estaríamos identificando o que chamamos de IA média com o que chamamos de IA forte, sendo apenas que agora tomamos o comportamento racional, e não mais o comportamento humano, como evidência da capacidade genérica de resolução de problemas. Mas então, a IA forte é possível, dada tal definição? Perguntando a mesma coisa de outra forma: um agente artificial pode agir racionalmente, dada nossa definição de racionalidade? A nosso ver, não há nenhuma razão *a priori* para pensarmos que a resposta para essa pergunta seja negativa, de modo que a questão deve permanecer em aberto, sendo definitivamente resolvida se e quando um tal agente artificial for produzido.

REFERÊNCIAS

DESCARTES, René. *Discurso do método*. Brasília: Editora da Universidade de Brasília; São Paulo: Ática, 1989.

NORVIG, Peter & RUSSELL, Stuart. *Artificial Intelligence: A modern approach*. 3.ed. Upper Saddle River, NJ: Prentice Hall, 2010.

SEARLE, John. *Mind, brains and science*. Cambridge, MA: Harvard University Press, 1985.

TARSKI, Alfred. 'The Concept of Truth in Formalized Languages'. In: *Logic, Semantics, Metamathematics*. Indianapolis: Hackett, 1983.

TURING, Alan. 'Computing machinery and intelligence'. In: *Mind*. n. 49, pp. 433-460.

10. Obviamente, poderíamos ter falsos negativos, já que lidamos com uma condição suficiente mas não necessária para a autoconsciência, mas não teríamos falsos positivos.

ISSN 2179-9180
GUAIRACÁ REVISTA DE FILOSOFIA, GUARAPUAVA-PR, V34, N1, P. 1-14, 2018