Interação, indistinguibilidade e alteridade na Inteligência Artificial

João Cortese[1]

Resumo: Como devemos agir diante de inteligências artificiais ou de robôs que viriam a ser indistinguíveis de um ser humano? O problema se insere no domínio de uma ética da tecnologia em relação ao homem, e algumas questões básicas devem ser levantadas. Pode-se pleitear que a IA sirva para investigar a inteligência humana. Mas, neste caso, deve-se precisar se o que se pretende investigar são os efeitos dessa inteligência ou sua ontologia. A questão é: afinal, o que está sendo comparado entre a inteligência humana e a IA? Não tratarei aqui da questão de fato de se um computador pode hoje se passar por um humano. Minha questão é sobre o estatuto ético de uma máquina caso isso ocorra: podemos tomar como equivalentes a indiscernibilidade pela interação e a constituição de um agente autônomo que teria, enquanto tal, um estatuto ético intrínseco? Trata-se, portanto, de colocar uma questão sobre a ontologia dos agentes éticos.

Palavras-chave: Robôs. Ontologia. Agentes éticos.

Abstract: How should we act in the face of artificial intelligences or robots that would be indistinguishable from a human being? The problem lies in the domain of an ethics of technology in relation to man, and some basic questions must be raised. One can plead that AI serves to investigate human intelligence. But, in this case, it must be determined whether what is intended to be investigated are the effects of this intelligence or its ontology. The question is, after all, what the purpose of comparing human with artificial intelligence is. The paper does not deal with the question of whether a computer can act today like a human being. It concerns the ethical status of a machine should this occur: can we take the indiscernibility by interaction and the constitution of an autonomous agent as an intrinsic ethical status as such? It is therefore a question of the ontology of ethical agents.

Keywords: Robots. Ontology. Ethical agents.

Introdução

Os recentes desenvolvimentos da Inteligência Artificial (IA) têm levado sistemas e robôs a realizarem ações cada vez mais indistinguíveis daquelas dos seres humanos. Saber até onde isso pode ser realizado, ou quando, é tarefa demasiadamente

[1] Doutor em co-tutela na Université Paris 7 e no Departamento de Filosofia da USP, pesquisador associado ao laboratório SPHERE (CNRS/Paris 7), membro do Núcleo de Bioética do Instituto PENSI - Pesquisa e Ensino em Saúde Infantil e participa do Grupo de Estudos em Inteligência Artificial do Instituto de Estudos Avançados da USP. Agradecimento do autor: Gostaria de agradecer a Bernardo Gonçalves, Fabio Cozman, Dora Kaufman, Hugo Neri e Lucas Petroni por terem me apresentado a diversas questões presentes neste artigo. Os colegas da *Associação Filosófica Scientiae Studia* acolheram uma primeira apresentação do presente trabalho, e Adriano Bechara, Marcos Paulo de Lucca-Silveira e Osvaldo Pessoa tiveram uma participação importante na discussão das ideias aqui apresentadas. E-mail: joaocortese@gmail.com.

difícil para o presente momento, ao menos para o autor. Isso não impede que nosso imaginário, não só em discussões como em diversos filmes recentes, já nos coloque: como devemos agir diante de inteligências artificiais ou de robôs que viriam a ser indistinguíveis de um ser humano? O problema se insere no domínio de uma ética da tecnologia em relação ao homem, e algumas questões básicas devem ser levantadas.

Um fator essencial aqui é a *eficácia:* a tecnologia da computação moderna percebeu que para ter *competência* sobre uma tarefa, não é preciso ter *compreensão* sobre ela. Uma máquina não precisa *entender* a aritmética para conseguir computar.[2] Que dizer então da ética envolvida para tais tipos de máquinas?

O Teste de Turing avalia, como se sabe, o resultado de *interações* entre um computador e um ser humano (TURING, 1950). À questão de se as máquinas podem pensar, Turing propõe uma substituição: seriam as máquinas capazes de se fazer indistinguíveis de humanos em um jogo da imitação?

Pode-se pleitear que a IA sirva para investigar a inteligência humana. Mas, neste caso, deve-se precisar se o que se pretende investigar são os *efeitos* dessa inteligência ou sua *ontologia*. A questão é: afinal, o que está sendo comparado entre a inteligência humana e a IA? A resposta é menos evidente do que parece. Tomando como referência o Teste de Turing, o que é evidente é que se consegue *simular* eficazmente diversos efeitos da interação humana – mas quais são os pressupostos envolvidos nisso?

Fala-se hoje de uma suposta superação deste teste por certos sistemas de IA. Pode-se questionar se isso de fato se realizou, pois cabe uma discussão sobre as condições nas quais isso teria se dado, assim como sobre a interpretação de tais resultados.[3] Não tratarei aqui da questão *de fato* de se um computador pode hoje se passar por um humano. Minha questão é sobre o estatuto ético de uma máquina caso isso ocorra: podemos tomar como equivalentes a indiscernibilidade pela interação e a constituição de um agente autônomo que teria, enquanto tal, um estatuto ético intrínseco? Trata-se, portanto, de colocar uma questão sobre a ontologia dos agentes éticos.

[2] Tal visão é apresentada por exemplo por Dennett (2013).
[3] Para uma crítica ao "sucesso" do teste de Turing, ver, por exemplo, Floridi et al. (2009).

IA "indistinguível"

A Inteligência Artificial (IA) pretende vir a se relacionar com seres humanos de maneira "indistinguível" em diversas frentes: por meio de um "chat", assim como já proposto pelo Teste de Turing original; em interação por aúdio: o Google já anunciou ser capaz de sintetizar voz indistinguível daquela de seres humanos,[4] e mais recentemente anunciou seu serviço *Duplex* de assistência, que seria pretensamente capaz de ligar para alguém sem ser distinguido de um ser humano,[5] vencendo os humanos no xadrez ou compondo música original esteticamente agradável;[6] dirigindo carros de maneira autônoma, como tem sido testado por diversas empresas; etc.

Poderíamos seguir com exemplos de êxito de IA indefinidamente. Mais do que comentar um exemplo concreto de sucesso ou de fracasso, importa aqui avançar o argumento de que, potencialmente, máquinas poderiam passar em qualquer teste interacional adaptável a elas, quando o que se avalia é a realização, ou não, de uma determinada *função*. Mas o que a interação pode mostrar sobre o agente?

Pensar sobre a IA é a outra face de se pensar sobre a inteligência humana. Ora, é claro que se pode considerar o ser humano unicamente a partir de suas interações, ou de seu comportamento – não foi o que fez, por exemplo, B. F. Skinner (1904-1990) com a sua psicologia comportamental, e o que é ainda objeto de diversos projetos contemporâneos? O aspecto metodológico do *behaviorista* é claro: tratar a psicologia humana unicamente a partir das interações entre os homens, não buscando uma causalidade além dos comportamentos observados.

Não há dúvida de que a modelagem do comportamento, por reforço ou por inibição, *funciona*. A teoria do behaviorismo "foi usada, por exemplo, para ensinar pombas a jogar tênis de mesa" (DALRYMPLE, 2017, p. 28). Mas é evidente que a questão é mais complexa: uma série de atos pode ser *interpretada* como a participação em um certo jogo, mas o que me garante que este jogo de fato esteja sendo *jogado* pelo agente, significando que realizar tais ações tenha um *sentido* para ele, abrangendo inclusive o desejo de vencer o jogo?

[4] Gershgorn, D. "Google's voice-generating AI is now indistinguishable from humans", *Quartz Media*, 26 de dezembro de 2017. Disponível em: <https://qz.com/1165775/googles-voice-generating-ai-is-now-indistinguishable-from-humans/>. Acesso em: 1 fev. 2018. Ver Shen, Jonathan, et al. "Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions." arXiv preprint arXiv:1712.05884 (2017).
[5] Ver, por exemplo, Harwell, D. "A Google program can pass as a human on the phone. Should it be required to tell people it's a machine?", *The Washington Post*, 8 de maio de 2018.
[6] Cf. Gonçalves em: Consentino et al. (2018).

É evidente que a realização de uma ação não responde pela *intencionalidade* que estamos acostumados a associar a ela. Esta pode existir ou não – a realização do comportamento não responde ao que, digamos assim, motivou este comportamento.

Parece assim que, em primeira instância, uma posição comportamentalista não responde em nada à questão ontológica do sujeito, ao menos no que esta diz respeito à intencionalidade deste sujeito.[7] Contudo, algumas versões do comportamentalismo podem se apresentar de maneira mais radical.

Conjuntos de dados e interações

> O que começou como metodologia tornou-se ontologia. Um adágio antigo do diagnóstico médico diz que a ausência de evidências nem sempre é a evidência da ausência, mas os behavioristas ignoram esse sábio chamado para a modéstia. Em vez disso, começaram a acreditar que estímulo e resposta era só o que havia na vida humana, que tudo que é humano pode ser explicado dessa maneira. Embora risível, isso foi levado extremamente a sério por muitos. (DALRYMPLE, 2017, p. 28)

À parte o mérito que tenham ou não tais teorias comportamentais para descrever o comportamento humano propriamente dito, sua eficiência parece ser indubitável em um outro aspecto contemporâneo: aquele da gestão de sistemas nos quais a pessoa humana é inserida, por definição, reduzindo-se sua totalidade a um conjunto de dados. Podemos crer que o homem seja irredutível a um aspecto binário. Mas ele construiu máquinas que funcionam sob tal lógica; desde que se submeteu a viver a partir de sistemas gerenciados por tais máquinas, é claro que ele em certo sentido decidiu viver sob tal estrutura de dados.

"Alice", uma mulher, quando classificada em um sistema de dados, se torna uma jovem mulher de perfil executivo com um diploma universitário etc. Nos mais diversos âmbitos de nossa vida hoje, do trabalho à saúde, dos estudos ao entretenimento e toda a interação virtual, pessoas são transformadas em meros *tipos* (estudante, cliente, terrorista potencial, etc.). Poderíamos dizer que cada indivíduo é inexaurível quanto às suas informações (voltaremos a isso). Por outro lado, a quantidade de dados é sempre finita, por maior que seja. Nós então abstraímos,

[7] Vale ressaltar que não se trata aqui de avaliar a posição comportamentalista de maneira aprofundada em nenhuma de suas versões, mas unicamente de levar em conta seu aspecto metodológico de maneira ampla, naquilo que ele poderia ser transferido para a avaliação da "intencionalidade" de inteligências artificiais e de robôs.

generalizamos, agregamos, interpolamos, agrupamos, classificamos dados... Lidando portanto com um número muito grande que é, porém, sempre finito.[8]

Ainda que o ser humano tenha mais informações do que cabe em qualquer conjunto de *big data*, não deixa de ser o caso de que este sempre será insuficiente para representá-lo em sua integralidade.[9] O que ocorre é que os bancos de dados tratados são suficientemente grandes para que, sob certos aspectos, possamos negligenciar o que é perdido, de maneira que dizemos, por exemplo, que estamos "conversando" com alguém por um chat.

Não há dúvida de que tais empreitadas sejam de grande *eficiência*. A "quantificação das interações humanas", se podemos dizer assim, toma cada vez maior parte em nossas sociedades. Cabe, porém, questionar onde reside o seu fundamento. Ora, além de descrever as interações humanas, a quantificação das interações sociais hoje *intervém e molda* parte das interações humanas. Trata-se de reconhecer a formatação das interações entre humanos que aparecem em meios que são, por sua própria constituição, intrinsecamente quantificáveis.

Pense-se, por exemplo, em um aplicativo para *smartphones* destinado a facilitar relacionamentos humanos – seja para compra e venda de imóveis, seja para a busca de companhia. Antes de falar-se no uso bom ou mal de tal aplicativo, deve-se levar em conta que, pelo seu próprio *design*, um tal aplicativo define já um espaço de possibilidades prévio, ao qual o usuário deverá adequar-se para poder "escolher". Ao mesmo tempo em que esse tipo de implementação cria um leque de possibilidades, cria também uma lista de "restrições" quanto às interações possíveis (que seja possível enviar mensagens de texto, de áudio ou de vídeo – é evidente que isso não esgota as interações humanas). Escrever um código é definir um espaço de possibilidades de vivências. Neste sentido, o homem restringe-se a uma quantidade finita de posibilidades, o que evidentemente pode ser mais fácilmente imitado por uma IA.

[8] "The overall perspective, emerging from digital ontology, is one of a metaphysical monism: ultimately, the physical universe is a gigantic digital computer. It is fundamentally composed of digits, instead of matter or energy, with material objects as a complex secondary manifestation, while dynamic processes are some kind of computational states transitions. There are no digitally irreducible infinities, infinitesimals, continuities, or locally determined random variables" (FLORIDI, 2011, p. 319).

[9] Poderíamos evocar aqui ainda a questão mais ampla de se o caráter qualitativo das vivências humanas pode ser representado por relações quantitativas implementadas.

Aspecto da consciência

Dennett (2013) lembra bem que a palavra "computador" não se aplicou unicamente a máquinas: antes de que Turing criasse as máquinas que levam o seu nome, pessoas possuíam a função de "computadores" em diversas instituições, geralmente mulheres. Estas realizavam uma série de contas necessárias a empreitadas complexas, num trabalho que seguia algoritmos de cálculo. É assim que Turing (1936, p. 251) declara que "podemos agora construir uma máquina para fazer o trabalho deste computador [humano]".

Nesta passagem, diz Dennett (2013, p. 571), "vemos a redução de *todas as computações possíveis* a um processo sem uma mente [*mindless*]". Isto é um fato; mas por outro lado podemos dizer que este processo sem mente repete *apenas* todas as computações possíveis. A questão, no fundo, é saber o que é passível de computação.

Um enorme conjunto de dados, utilizado por uma IA dotada de *machine learning*, parece portanto poder vir a gerar uma simulação de interação humana tão bem quanto se queira – desde que aceitemos, como no Teste de Turing, que a "interação" pode ser modelada, avaliando uma função específica.

Mas o que isso nos diz sobre a ética de tais sistemas de IA? Um aspecto fundamental a ser considerado aqui é aquele da *consciência*, frequentemente considerada como necessária para que se considere que um agente tem estatuto ético.[10] Mas como saber se uma máquina é consciente? Se por acaso, examina Descartes na sua Segunda Meditação, vejo pela janela homens que passam pela rua, não deixaria de dizer, ao vê-los, que vejo homens;

> e, entretanto, que vejo desta janela, senão chapéus e casacos que podem cobrir espectros ou homens fictícios que se movem apenas por molas? Mas julgo que são homens verdadeiros e assim compreendo, somente pelo poder de julgar que reside em meu espírito, aquilo que acreditava ver com meus olhos. (DESCARTES, 1904, p. 25)[11]

Seriam esses "chapéus e capas" que passam diante de minha janela realmente homens ou meros autômatos? Como sabê-lo?

[10] Falo aqui de "consciência" no sentido forte de autoconsciência.
[11] Tradução em Descartes (Os Pensadores) São Paulo: Abril Cultural, 1983.

> Como saberemos se nossas máquinas se tornaram conscientes? Descartes argumentou que a própria consciência está além de qualquer possibilidade de dúvida. No caso dos outros, nunca estamos absolutamente certos. Muitos de nós tivemos, ainda que por um momento, a ideia de que todos os outros pudessem ser um zumbi: rindo, chorando, reclamando, regozijando-se, mas sem "ninguém em casa". Talvez os cientistas acabem descobrindo a assinatura da consciência, e então poderemos testá-la em nossos robôs, assim como nos animais e uns aos outros. Mas é certo que construiremos máquinas que *parecem* conscientes muito antes de chegarmos a esse ponto. (BLOOM; HARRIS, 2018)

À parte a questão de se a máquina terá efetivamente uma consciência, a *simulação* de uma consciência certamente aparecerá muito antes. Mas isso não é o mesmo que a consciência, ao menos de um ponto de vista ético. O que a IA faz aqui é uma aproximação, que se torna indistinguível de um ser humano.

Aproximação indefinida

A própria noção de aproximação parece estar no coração da ciência moderna. Contrariando a clássica separação aristotélica, a partir dos séculos 16 e 17 vê-se uma tendência na Europa a relacionar as Matemáticas e as Ciências Naturais, em particular matematizando a Física. Isto tem implicações também para as práticas da Engenharia.

Cabe dizer que, ao contrário do que se crê comumente, as matemáticas e a exatidão não se identificam necessariamente. Em meados do século 17, por exemplo, o cálculo das probabilidades foi inventado por Blaise Pascal e Pierre Fermat: a Matemática já podia quantificar o incerto, algo que não seria sem implicações para a computação e a IA.

Norbert Wiener, o criador da cibernética, acreditava que, mais do que a Einstein ou a Planck, deveríamos creditar a J. W. Gibbs, já no século 19, a maior das revoluções na Física do século 20: aquela de tratar probabilisticamente fenômenos contingentes.

> Nenhuma medição física é jamais precisa; e o que tenhamos a dizer acerca de uma máquina ou de outro sistema mecânico qualquer concerne não àquilo que devamos esperar quando as posições e momentos iniciais sejam dados com absoluta precisão (o que jamais ocorre), mas o que devemos esperar quando eles são dados com a precisão alcançável. [...] Por outras palavras: a parte funcional da Física não pode furtar-se a considerar a incerteza e contingências dos eventos. (WIENER, 1968, p. 10)

Isto impactaria tanto a própria ciência da época de Wiener quanto "nossa atitude para com a vida em geral" (WIENER, 1968, pp. 13-14). Quer dizer que, para Wiener, o "paradigma'' gibbsiano, se podemos dizer assim, é uma das matrizes do nosso modo de viver moderno (Wiener escrevia nos anos 1950). A análise da aproximação e da incerteza seria assim constitutiva da ciência moderna. O mesmo, podemos crer, se aplica à IA: a noção de *aproximação indefinida* aparece como um critério fundamental na avaliação das interações descritas até aqui. Se tal aproximação é possível no caso da IA, a simulação da consciência será indistinguível da própria consciência.

Aceitar uma aproximação como solução implica definir quão próximo se está da solução exata, o que implica uma teoria do erro. Isso aparece de certa maneira na fundamentação dos métodos dos Cálculos Integral e Diferencial. Estabelecidos pela análise do século 19, eles foram problematizados já no século 17, retomando desenvolvimentos tão antigos quanto aqueles que aparecem nos escritos de Euclides e de Arquimedes.

Vale aqui ressaltarmos brevemente um aspecto da indistinguibilidade na prática matemática, no método dos "indivisíveis" de Pascal.[12] Retomando o "Método da exaustão" arquimediano, Pascal propõe, como diversos autores de sua época, que para calcular a área sob uma curva, sejam somados retângulos delimitados abaixo ou acima dela (Figura 1). O problema é que sempre a área desses retângulos excede ou falta em algo para cobrir a área da curva. A soma seria "exata" apenas caso houvesse infinitos retângulos, cada um deles com uma área infinitamente pequena (os "infinitesimais"). No método dos indivisíveis de Pascal, entretanto, isso não é preciso: basta que a diferença entre a área sob a curva e a soma dos retângulos seja menor do que uma "quantidade dada qualquer" para que o resultado seja considerado como uma solução. O controle do erro permite uma aproximação indefinida, aceita na prática como uma solução válida.

[12] Para mais detalhes, ver Cortese, J. F. N. L'infini en poids, nombre et mesure: la comparaison des incomparables dans l'oeuvre de Blaise Pascal. Tese de doutorado, Université de Paris 7 e Universidade de São Paulo, 2017.

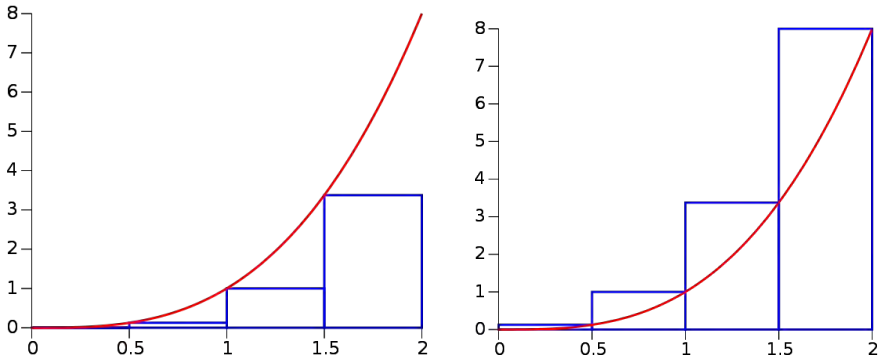

Figura 1. Cálculo da área sob uma curva: soma dos retângulos delimitados abaixo ou acima dela.
Fonte: <https://en.wikipedia.org/wiki/Riemann_sum>. Acesso em: 1 fev. 2018. Imagens de domínio público.

Trata-se em um certo sentido do que Aristóteles já propusera na sua *Física* (III, 11): os matemáticos não precisam do infinito atual para seus cálculos, mas apenas de um finito tão grande quanto se queira. O que importa aqui é que uma certa prática matemática pode ser feita em total acordo com uma colocação entre parênteses da questão ontológica do infinito, bastando uma aproximação indefinida para que se efetuem os cálculos.

Tais considerações podem talvez trazer alguma luz sobre a questão do que é que se avalia ao se abordar a interação entre o humano e a máquina. Que dizer nesse caso? O infinito é necessário desde que passamos aos fenômenos humanos ou o indefinido bastaria para avaliarmos tais interações?

> Dentre todos os filósofos que rejeitam o Teste de Turing como um critério para a inteligência, pode-se esperar que alguém apoie a proposta de responder positivamente à Grande Questão, tão quixotesco quanto isso possa parecer. (SHIEBER, 2004, p. 268)

É desta maneira que Shieber (2004) apresenta Daniel Dennett. Eu gostaria de ser aqui "quixotesco" no sentido oposto ao de Dennett, mas no que concerne a agência moral. Trata-se de dizer que, no caso do ser humano, não podemos nunca esgotar este *infinito* que é o homem. Não basta que nos satisfaçamos com nenhuma caracterização parcial, com nenhuma redução de dimensionalidade do que é o humano. A modelização dos fenômenos humanos nunca poderia ter uma pretensão ontológica radical, se admitimos o próprio fato de que a modelização deve ter sempre limites. Por mais que

tenhamos que lidar com um mero "indefinido" que não podemos fazer ser "infinito", temos de aceitar que este infinito existe – e está colocado na pessoa do outro, ainda que eu veja os "limites" do corpo deste. O ser humano é sempre irredutível, por mais que se faça uma aproximação dele.

Dada esta concepção, creio que temos de abandonar uma exequibilidade indistinguível como critério de demarcação de um agente ético no sentido próprio. A verdadeira ética, afinal, não está numa questão do fazer, mas numa questão do ser – entendendo por isso que a autoconsciência seja necessária para a constituição de um agente ético. Isso não quer dizer porém, como veremos adiante, que a dimensão ética esteja absolutamente excluída desse tipo de situação – o que pleiteio é que a constituição de um *agente* ético não pode ser mostrada por interações.

Estatuto moral da IA

A *indiscernibilidade* sob um certo critério (pelo controle do erro) mostra como de uma ideia matemática se faz algo que *funciona*. O critério interacional também *funciona*, como em toda boa tecnologia; mas será que isso basta no caso da ética?

Acredito que a consideração de tal aspecto puramente interativo não pode dar uma resposta sobre a constituição de um agente ético, *nem no caso dos homens e nem no caso das máquinas*. Cabe aqui voltarmo-nos ainda para o infinito, que já na prática matemática trazia uma questão ontológica independente daquela verificável pelas "interações" da prática do indefinido.

Se Descartes, dando um passo a partir da tradição medieval, refletiu sobre o homem, ser finito, que busca conhecer o Outro que é Deus, ser infinito, algo semelhante pode ser visto nas interações entre os homens. Para Emanuel Lévinas, de fato, a *alteridade* é vista como algo irredutível: cada *outro* é um infinito que se apresenta a mim, e pretender "conhecê-lo" plenamente seria reduzi-lo. A alteridade humana torna-se assim discernível daquela da máquina – mas não se trata de fazê-lo considerando somente as interações, sob a pena de perder aquilo que é intrinsecamente ético.

Antes de analisar quais seriam as implicações dessa concepção, cabe analisar uma dentre as diversas proposições contemporâneas sobre o tema da ética da IA: aquela de Bostrom e Yudkowsky (2011/2014).

Esses autores consideram um ser tendo um estatuto moral quando ele é um fim em si mesmo, e não um meio para algo. Eles declaram que, pelo momento,

> [...] é amplamente aceito que os atuais sistemas de IA não têm *status* moral. Nós podemos alterar, copiar, encerrar, apagar ou utilizar programas de computador tanto quanto nos agradar, ao menos no que diz respeito aos próprios programas. (BOSTROM; YUDKOWSKY, 2011, p. 208)

Não é por isso, entretanto, que seja claro quais atributos deveriam ser levados em conta para avaliar se um sistema de IA tem estatuto moral ou não. Bostrom e Yudkowsky identificam dois critérios comuns propostos à avaliação de se uma máquina tem estatuto moral ou não: a *Senciência*, ou seja, a "capacidade para a experiência fenomenal ou *qualia*, como a capacidade de sentir dor e sofrer", e a *Sapiência*, o "conjunto de capacidades associadas com maior inteligência, como a autoconsciência e ser um agente racional responsável". Seria portanto moral um agente que tivesse ambas essas capacidades. Nesta linha, desligar um computador hoje não parece infringir um direito, mas se fosse possível desligar uma máquina que pudesse "sentir dor", isto seria moralmente errado, assim como maltratar um animal.

Bostrom e Yudkowsky indicam ainda que atrás desta ideia subjaz um outro princípio:

> *Princípio da Não-Discriminação do Substrato*: "Se dois seres têm a mesma funcionalidade e a mesma experiência consciente, e diferem apenas no substrato de sua aplicação, então eles têm o mesmo *status* moral". (BOSTROM; YUDKOWSKY, 2011, p. 209)

Quer dizer que "não faz diferença moral se um ser é feito de silício ou de carbono, ou se o cérebro usa semicondutores ou neurotransmissores". Ora, o problema é que esse princípio parece fazer sentido apenas se pressupomos um reducionismo materialista em relação à consciência. Ainda aqui, o problema parece ser o mesmo: a solução funciona desde que pressuponhamos que a consciência é redutível a propriedades físicas que podem ser decompostas.

Os autores apresentam ainda um outro princípio, que faz igualmente referência à consciência sem problematizar como identificá-la:

> *Princípio da não-discriminação da ontogenia*: "Se dois seres têm a mesma funcionalidade e mesma experiência de consciência, e diferem apenas na forma como vieram a existir, então eles têm o mesmo *status* moral". (BOSTROM; YUDKOWSKY, 2011, p. 210)

Deveríamos conceder à personagem *Joi*, uma IA no filme *Blade Runner: 2049*, que um homem é feito apenas de dados, A e C e T e G, meros quatro símbolos, e o robô é feito dos dois símbolos 0 e 1, de maneira que a única diferença entre um homem e uma máquina seria quantitativa? Bostrom e Yudkowsky (2011) avançam os princípios de que, tanto o substrato como o fato de ser objeto de um *design* ou não, pouco importam. Entretanto, eles parecem cair, em um certo sentido, no erro de tomar a metodologia pela ontologia: não parece haver meio de conceder ou não a "mesma experiência consciente" a dois seres a não ser por meio de suas interações funcionais. Ainda aqui, é um pressuposto metafísico reducionista que parece adiantar a resposta que foi "buscada".

Como dito acima, escolho também partir de um pressuposto metafísico, porém aquele diametralmente oposto: o de que, a despeito da "equivalência" comportamental, há um *infinito* humano contraposto a um *indefinido* da máquina. Cada lado sairá descontente com o pressuposto metafísico contrário; dado, porém, que a ciência hoje não parece ter provado cabalmente nem o reducionismo nem o irreducionismo físico da consciência, a questão é a quem cabe o ônus da prova.

Considerar portanto o agente ético apenas como um *outro* que é infinito, ao invés de indefinido, caracteriza um tipo de "resistência ontológica", sob um certo ponto de vista quixotesca. Seria possível apresentar uma motivação à ela?

Alteridade

No filme *Contato*,[13] a cientista interpretada por Jodie Foster decifra os sinais enviados por seres do espaço para descobrir que eles indicam como construir uma máquina. Esta concluída, a doutora embarca e a liga, iniciando um tipo de viagem. Em

[13] Filme de 1997, dirigido por Robert Zemeckis e adaptado de um romance de Carl Sagan.

vez de pequenos homens verdes, ela se encontra numa praia de areias claras, e vê vindo ao seu encontro seu pai, já morto. "Você não é real, nada disso é real", diz a mulher, desorientada; "quando eu estava inconsciente, você extraiu meus pensamentos, minhas lembranças". "Achamos que assim seria mais fácil para você", respondem *eles*, sob a forma do seu pai. O contato foi feito, mas *eles* não podem se mostrar por completo. O *outro*, enquanto outro, sempre se apresenta de maneira mais diferente do que o *eu* gostaria de crê-lo.

Cabe retomar aqui o tratamento cartesiano do infinito. Como se sabe, uma das demonstrações das *Meditações* sobre a existência de Deus repousava sobre o seguinte argumento: tenho a ideia de infinito em mim; sou, porém, um ser finito; como posso então ter a ideia de infinito em meu interior? Unicamente se um ser infinito deixou esta ideia dentro de mim.

Lévinas (2000) estendeu essa ideia à interação entre o eu e o *outro* que é um homem. Para este autor, a alteridade é vista como algo irredutível: cada *outro* é um infinito que se apresenta a mim, e pretender "conhecê-lo" plenamente seria reduzi-lo.

Mas, como vimos, pode-se propor de um ponto de vista matemático que o indefinido não é o mesmo que o infinito. Cabe distingui-los para saber quando é que basta tratar de algo suficientemente grande, e portanto indistinguível do infinito porque imenso, ou quando trata-se do verdadeiro infinito.

Para Lévinas (2000, p. 31-32), o outro é infinito – não um infinito matemático, ou um infinito de mera negação do finito, mas um infinito de transcendência. Para este autor, a ética *precede* a ontologia: pretender conhecer o outro para então interagir com ele seria tentar reduzi-lo, o que é impossível. Trata-se, ao invés disso, de fazer uma escolha moral: tratar o outro como um *sujeito moral* ou como um *objeto*. A atitude ética fundamental é colocar-se face ao outro. Há dimensão ética quando coloco-me face a face a alguém. Pode-se ainda dizer, com Martin Buber (2001), que ela aparece quanto me situo diante de um *Tu*; ela não existe quando me situo diante de um *isso*.

A alteridade é, portanto, irredutível para Lévinas (2000). Mas poderíamos colocar ainda a questão: a máquina seria um outro? Se *o meio é a mensagem* (cf. Marshall McLuhan), a tecnologia com a qual eu "falo" é um outro (cf. Lévinas) ou um mero dispositivo?

Alguns pensadores propuseram recentemente que, segundo uma ética lévinasiana, poderíamos crer que o homem, ao lidar com o robô de maneira próxima e antropomorfizada, chegaria a tratar-lhe como um *Tu*, conferindo-lhe portanto um estatuto ético (WOHL, 2014; GUNKEL, 2012). Desta maneira, com a ética precedendo a ontologia, e havendo a opção de colocar-se diante de um *Tu* ou de um *isso*, uma alteridade poderia ser encontrada na interação com máquinas.

Seria interessante abordar mais detidamente a questão da "alteridade" da máquina num trabalho futuro. Por ora, vale apenas levantar a questão de se, mais do que tratar a máquina como um *Tu*, não estamos talvez nos apresentando a nós mesmos como um *isso*. Desde que o homem *entra* no sistema que projetou, desde que ele se reduz a uma certa modelização, ele se empobrece ao crer que isso é todo o seu ser. Perguntar por uma interação ética meramente a partir de uma funcionalidade numa interação já é empobrecer a questão. Parece que, do ponto de vista "ontológico", a alteridade segundo Lévinas (2000) é um infinito transcendente, que não poderia ser confundido com um indefinido.

Esse posicionamento pode parecer difícil para um funcionalista ou para qualquer pessoa que discorde do pressuposto de que o ser humano é um "outro" infinito. Eu gostaria, contudo, de terminar apresentando um argumento quanto à moralidade no trato de robôs antropomórficos que independe desse pressuposto, podendo, portanto, ser aceito de maneira mais geral.

Robôs antropomórficos

No caso de robôs e sistemas de IA que se assemelham indistintamente a pessoas, cabe refletir sobre o fato de que parecemos ter uma tendência a antropomorfizá-los, a despeito de nossas visões filosóficas ou de uma advertência dos fabricantes sobre como foram construídos. Como agir, portanto, em relação a esses robôs? Bloom e Harris (2018) consideram o que Kant dizia sobre o respeito humano aos animais. Ainda que ele visse estes como coisas sem valor moral, ele insistia em que os homens os tratassem adequadamente: "pois quem é cruel com animais torna-se duro

também na sua conduta com os homens".[14] Quanto mais não seria o caso para robôs que em sua aparência e interação fossem indistintos de seres humanos?

> Nós certamente poderíamos dizer o mesmo para o tratamento de robôs realistas [*lifelike*]. Mesmo se pudéssemos ter certeza de que eles não estejam conscientes e não possam realmente sofrer, a tortura deles provavelmente prejudicaria o torturador e, em última análise, as outras pessoas em sua vida. (BLOOM; HARRIS, 2018)

Isso quer dizer que devemos, desse modo, respeitar as máquinas de alguma maneira, sob a pena de nós mesmos nos fazermos piores no caso contrário. Há limites éticos no tratamento de IAs, *mesmo que elas não sejam agentes morais* – este "estatuto ético derivado", por assim dizer, vem simplesmente do fato de que somos homens, de estatuto moral, a lidar com elas.

Passemos a um exemplo prosaico, que já pode ocorrer em nossa sociedade. Um garoto trata *Siri*[15] de maneira indelicada, insultando-a quando lhe ordena que forneça alguma informação. A mãe lhe repreende: "filho, não fale assim com ela". O filho responde: "mas mãe, é só uma máquina". O filho está correto? Sim, pois ontologicamente não parece haver razões para tratar a máquina como um ser dotado de liberdade ou de autoconsciência, fatores importantes para se considerar um agente como detentor de moralidade no sentido forte. Mas a mãe está também correta: ao agir como tal, em particular com um sistema que simula um tipo de interação tipicamente humana (trocar informações numa conversa), o garoto está *agindo mal*, no sentido de que deixa de lado uma virtude no seu agir. Quem piora quando Siri é "desrespeitada" não é o próprio sistema de IA, pois a rigor ele não sofre, mas a própria pessoa que realiza a ação, pois em certo sentido ela se empobrece quanto à sua dignidade no trato com os outros.

Considerações finais

Como conceituar essa dimensão "ética" no trato com robôs e com inteligências artificiais? Eis uma importante discussão a ser feita, e que poderia se desenvolver por

[14] Kant, I. (1997 [1784–5]). "Moral Philosophy: Collin's Lecture Notes", in *Lectures on Ethics*, P. Heath and J.B. Schneewind (org. e trad.), Cambridge : Cambridge University Press, p. 212.
[15] Uma IA da Apple que atua como assistente pessoal, interagindo por voz com o usuário.

diversas vias. Eu gostaria de apresentar aqui simplesmente um esboço de uma proposta que distinguiria duas instâncias éticas.

A primeira está no caso de uma interação "ontologicamente ética": é quando me coloco em interação com alguém que sei possuir um estatuto ético próprio, estabelecendo uma relação entre um *Eu* e um *Tu* (BUBER, 2001) e situando-me face a face com ele (LÉVINAS, 2000). Poderíamos dizer que lido aqui com um "agente moral", ou mesmo com um "paciente moral", no sentido de alguém que, tendo um estatuto moral próprio, sofre uma ação.[16] Nessa situação, posso perguntar: há *alguém* aí? Com *quem* eu falo?

O segundo caso é o de uma interação na qual não lido com alguém que possua estatuto ético, mas trato de um *isso*, de um objeto. Entretanto, do próprio fato de que sou uma pessoa autoconsciente, um estatuto ético perpassa todas as minhas ações, de maneira que posso considerar este *isso* como um "objeto de moralidade", ou seja, um objeto sem intenção, em relação ao qual eu mesmo posso agir eticamente ou não. Pode-se pensar "moralidade" de tal objeto deve ser maior quando ele for um produto humano investido de intencionalidade (fabricado por alguém), ou quando a sua interação comigo assemelhar-se àquela de humanos. Nessa situação, devo perguntar: *o que* está aí? Com *o que* eu "falo"? Como devo me colocar em relação a isso?

Podemos problematizar se há um âmbito ético relacionado a robôs e inteligências artificiais. A única conclusão definitiva é que já não podemos deixar de nos colocar questões sobre isso.

Enviado: 2 fevereiro 2018
Aprovado: 9 março 2018

Referências

BLOOM, P.; HARRIS, S. The Stone: it's Westworld: What's wrong with cruelty to robots? New York Times, 23 de 4, 2018. Disponível em: <https://www.nytimes.com/2018/04/23/opinion/westworld-conscious-robots-morality.html>. Acesso em: 1 fev. 2018.

[16] Sobre as diversas possibilidades do uso dos termos "agente moral" e "paciente moral", ver Floridi e Sanders (2004).

BOSTROM, Nick; YUDKOWSKY, Eliezer. A ética da Inteligência Artificial, trad. Pablo Araújo Batista. *Fundamento*, v. 1, n. 3, p. 200-226, 2011.

_____. The ethics of Artificial Intelligence. In: FRANKISH, Keith; RAMSEY, William M. *The Cambridge handbook of Artificial Intelligence*. Cambridge: Cambridge University Press, p. 316-334, 2014.

BUBER, M. *Eu e e tu*, trad. Newton Aquiles von Zuben. São Paulo: Centauro, 2001 [1974].

DALRYMPLE, T. *Evasivas admiráveis*: como a Psicologia subverte a moralidade. São Paulo: É Realizações, 2017.

DENNETT, D. Turing's strange inversion of reasoning. In: COOPER, S. B.; LEEUWEN, J. van. (Orgs.). *Alan Turing*: his work and impact. Amsterdam: Elsevier, 2013.

DESCARTES, R. *Oeuvres*, vol. IX. Adam, C.; Tannery, P. (Orgs.). Paris: Cerf, 1904.

FLORIDI, L. *The philosophy of information.* Oxford: Oxford University Press, 2011.

FLORIDI, L.; SANDERS, J. On the morality of artificial agents. *Minds and Machines*, 14 (3), 349-379, 2004.

FLORIDI, L.; TADDEO, M.; TURILLI, M. Turing's imitation game: still an impossible challenge for all machines and some judges – an evaluation of the 2008 Loebner Contest. *Minds and Machines*, 19(1), 145-50, 2009.

Consentino, Marcelo; Gonçalves, B.; Cozman, F.; Wasserman, R. Os primeiros 60 anos de feitos da Inteligência Artificial – Revisitando as previsões de Herbert Simon. *Estadão: Estado da arte*, 23 de março de 2018. Disponível em: <https://oestadodaarte.com.br/inteligência-artificial/>. Acesso em: 1 fev. 2018.

GUNKEL, D. J. *The machine question:* critical perspectives on AI, robots, and ethics. Cambridge, MA: MIT Press, 2012.

LÉVINAS, E. *Totalité et infini:* essai sur l'extériorité. Paris: Hachette. 2000 [1961].

SHIEBER, S. M. (Org.). *The Turing test*: verbal behavior as the hallmark of intelligence. Cambridge, MA: MIT Press, 2004.

TURING, A. On computable numbers, with an application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society*, 42, 23–265, and erratum (1937) 43, p. 544-546, 1936.

TURING, A. Computing machinery and intelligence. *Mind*, 59 (236), 433-60, 1950.

WIENER, N. *Cibernética e sociedade:* o uso humano de seres humanos. São Paulo: Cultrix, 1968.

WOHL, B. S. Revealing the 'face' of the robot: introducing the ethics of Levinas to the field of robotics. In: KOZLOWSKI, Krzysztof; OSMAN, Tokhi; MOHAMMED, O.; GURVINDER, Virk S. (Orgs.). *Mobile service robotics*, Singapore: World Scientific, 704-714, 2014.