CONEXÕES
NOVAS TECNOLOGIAS, SOCIEDADE E DIREITO
REVISTA DA ESCOLA DE PÓS-GRADUAÇÃO DA FADI

ZUMBIS MORAIS

POR QUE ALGORITMOS NÃO SÃO AGENTES MORAIS*

MORAL ZOMBIES: WHY ALGORITHMS ARE NOT MORAL AGENTS

Carissa Véliz
Universidade de Oxford, GBR
carissa.veliz@philosophy.ox.ac.uk
orcid.org/0000-0002-3189-3994

* O artigo foi originalmente publicado sob o título **Moral zombies: why algorithms are not moral agents** em **AI & Society**, v. 36, n. 2, 2021, p. 487-497, DOI: https://doi.org/10.1007/s00146-021-01189-x, e traduzido como parte das atividades do grupo de pesquisa Neoliberalismo, Tecnologia e Direito – NTD por Bruno Ian Lupi Jorge, Celso Naoto Kashiura Jr. e Isabelle Scheffer Rodrigues.

Zumbis morais: por que algoritmos não são agentes morais

Resumo: Na filosofia da consciência, zumbis são criaturas imaginárias que consistem em duplicatas físicas exatas, para as quais não há experiência em primeira pessoa, de sujeitos conscientes. Zumbis devem mostrar que o fisicalismo – teoria segundo a qual o universo é integralmente feito de componentes físicos – é falsa. Neste artigo, aplico o experimento mental do zumbi ao campo da moralidade para verificar se a capacidade moral é independente da senciência. Argumento que algoritmos são um tipo de zumbi moral funcional, de modo que pensar nesses possa nos ajudar a entender e regular aqueles. Afirmo que a principal razão pela qual algoritmos não podem ser nem autônomos nem imputáveis é a ausência de senciência. Zumbis morais e algoritmos são incoerentes como agentes morais porque lhes falta o necessário entendimento moral para serem moralmente responsáveis. Para entender o que o que é causar dor a alguém, é necessário conhecimento experimental da dor. No máximo, para um algoritmo que não sente nada, "valores" podem ser itens numa lista, talvez distribuídos numa ordem de prioridade a partir de um número que representa o peso. Mas entidades que não sentem não podem valorar e seres que não podem valorar não podem agir por razões morais.

Palavras-chave: Algoritmos. Capacidade moral. Responsabilidade moral. Sistemas autônomos. Zumbis. Imputabilidade. Autonomia. Senciência. Consciência. Responsividade-a-razões.

Moral zombies: why algorithms are not moral agents

Abstract: In philosophy of mind, zombies are imaginary creatures that are exact physical duplicates of conscious subjects for whom there is no first-personal experience. Zombies are meant to show that physicalism – the theory that the universe is made up entirely out of physical components – is false. In this paper, I apply the zombie thought experiment to the realm of morality to assess whether moral agency is something independent from sentience. Algorithms, I argue, are a kind of functional moral zombie, such that thinking about the latter can help us better understand and regulate the former. I contend that the main reason why algorithms can be neither autonomous nor accountable is that they lack sentience. Moral zombies and algorithms are incoherent as moral agents because they lack the necessary moral understanding to be morally responsible. To understand what it means to inflict pain on someone, it is necessary to have experiential knowledge of pain. At most, for an algorithm that feels nothing, 'values' will be items on a list, possibly prioritised in a certain way according to a number that represents weightiness. But entities that do not feel cannot value, and beings that do not value cannot act for moral reasons.

Keywords: Algorithms. Moral agency. Moral responsibility. Autonomous systems. Zombies. Accountability. Autonomy. Sentience. Consciousness. Reasons-responsiveness.

Introdução

Na filosofia da consciência, zumbis são criaturas imaginárias projetadas para ilustrar problemas relacionados ao fisicalismo e à consciência. Zumbis são duplicatas físicas exatas de sujeitos conscientes, mas para os quais não existe nenhuma experiência em primeira pessoa. Além de seres fisicamente idênticos a nós, molécula por molécula, são também nossa duplicata funcional. Zumbis se comportam de maneira indistinguível dos seres humanos – reclamam sobre o tempo, choram quando assistem filmes tristes e sabe-se que discutem questões filosóficas por horas. Diferente dos seres humanos, não possuem uma experiência consciente. Não existe nada igual a ser um zumbi –

eles não sofrem com a baixa temperatura quando está frio, não sentem tristeza quando choram e dilemas filosóficos não os preocupam nem os empolgam.

Zumbis são utilizados para demonstrar que o fisicalismo – a teoria que defende que o universo é composto inteiramente por componentes físicos – é falso. Se pudessem existir criaturas que agem como nós, mas sem experiências conscientes, então a consciência teria que ser algo além de nossos corpos físicos.

O experimento mental do zumbi ainda não foi aplicado no campo da moralidade, mas existe uma questão análoga para ser explorada relacionada à senciência e à capacidade moral. Aplicar o experimento mental dos zumbis à capacidade moral pode nos ajudar a avaliar se a capacidade moral é algo independente da experiência consciente ou da senciência. Chamo de senciência a capacidade humana de, no mínimo, ter uma experiência subjetiva de prazer e dor. Desejo permanecer agnóstica quanto à ideia de que a experiência subjetiva é sinônimo de senciência. Para o propósito desse artigo, foco na senciência porque é possível conceber que uma experiência consciente sem a sensação de dor ou prazer e sem emoções e considero que essas experiências são as mais importantes para a capacidade moral. Aprofundar-se nesses problemas, no entanto, está além do escopo deste artigo, sendo desnecessário para os meus propósitos.

A seguir, apresento o argumento de que os algoritmos são semelhantes a zumbis morais e explico o significado do experimento mental dos zumbis morais. Em seguida, sugiro que parte da literatura sobre algoritmos assume um certo grau de capacidade moral nos algoritmos. A maior parte das abordagens sobre capacidade moral defende alguma versão de autonomia ou alguma versão de responsabilidade moral ou ambas como constitutivas da capacidade moral. Analiso, então, se algoritmos e zumbis morais são autônomos no sentido relevante para a capacidade moral e argumento que não são. Então, argumento que a responsabilidade moral é mais relevante para definir a capacidade moral do que a autonomia, mas nem os algoritmos nem os zumbis morais são moralmente responsáveis. Defendo que a principal razão pela qual algoritmos não podem ser nem autônomos nem responsáveis é porque não possuem senciência. Zumbis morais são incoerentes como agentes morais porque não possuem o entendimento moral necessário para serem moralmente responsáveis. Para compreender o que significa causar dor a alguém, é necessário ter o conhecimento experimental da dor. Por fim, finalizo o artigo respondendo a três possíveis objeções.

I. Algoritmos como zumbis morais

Zumbis morais seriam criaturas que agem de forma indistinguível de nós como agentes morais, mas para as quais não há nada que se assemelhe a ser eles. Zumbis morais agiriam como nós, mas não sentiram como nós. Assim como os seres humanos, seriam capazes de fazer o bem, doando para caridade, protegendo pessoas vulneráveis contra injustiças e respeitando direitos. Eles também teriam o poder de fazer o mal, violando direitos e causando sofrimentos desnecessários, poderiam insultar, trair, ameaçar e agredir fisicamente outras pessoas. Diferente dos seres humanos, eles não

sentiriam dor, prazer, empatia, intimidade, remorso, culpa, vergonha ou qualquer outra emoção moral. Um zumbi moral não se alegraria por salvar uma vida, nem sofreria culpa por tirar uma.

Zumbis morais, como descrevemos, parecem ser criaturas concebíveis, o que mostra que o poder de causar impacto moral no mundo é independente da senciência. Mas já sabemos isso por conta de furacões e outros fenômenos naturais que podem causar danos sem que haja nenhum agente. A questão relevante para a ética é se os zumbis morais são concebíveis *como agentes morais*. Se não forem, então o experimento mental sugeriria que onde há capacidade moral há senciência.

Zumbis morais podem não ser apenas um experimento mental. Pode-se argumentar que algo como os zumbis morais já existe entre nós; eles são os chamados de algoritmos, robôs ou IA (usarei esses termos de forma mais ou menos intercambiável). Sistemas automatizados podem (ainda) não parecer conosco, mas esse detalhe é, pode-se argumentar, moralmente irrelevante (pelo menos para a questão da capacidade moral). Podemos imaginar robôs que se pareçam com seres humanos num futuro não tão distante. Embora a IA não seja e, provavelmente, nunca virá a ser uma duplicata física dos seres humanos, ela pode ser considerada uma duplicata funcional em alguns aspectos – razão pela qual as máquinas podem substituir os seres humanos em um número crescente de tarefas. Os algoritmos já se assemelham a nós em algumas das decisões que tomam e no impacto moral que podem ter sobre o mundo, com pessoas sendo contratadas, demitidas, recompensadas e até mesmo presas por causa deles. Mas eles não são como nós, pois não há nada que se assemelhe a eles. Portanto, pensar sobre algoritmos pode nos ajudar a refletir sobre o experimento mental dos zumbis morais e pensar sobre zumbis morais pode nos ajudar a compreender melhor os algoritmos.

A primeira objeção que alguém poderia apresentar ao meu argumento é que não podemos saber com certeza que algoritmos ou máquinas não são sencientes. Admito que não consigo provar sem nenhuma dúvida que não há nada que seja igual a ser um algoritmo. Mas, da mesma forma, também não posso provar que pedras não são sencientes. Se os panpsiquistas estiverem certos, até as pedras podem ter algum grau de consciência ou protoconsciência. Entretanto, não existem evidências suficientes para sugerir que algoritmos (ou pedras) possuem uma consciência ou sejam capazes de sentir.

Pode-se argumentar, para além humildade epistêmica, que, se os algoritmos se comportam como nós, deveríamos tratá-los como tratamos os seres humanos (Danaher, 2020; Sparrow, 2004). Até o momento, no entanto, os algoritmos se assemelham a nós apenas em algumas das tarefas e funções que desempenham na sociedade, mas não num sentido mais amplo. Nesse aspecto, não são zumbis completos, mas somente algo como zumbis morais funcionais. Hoje, algoritmos podem selecionar candidatos para vagas de trabalho ou avaliar a probabilidade de alguém cometer um crime, mas não podem desenvolver uma paixão por comédias francesas ou se emocionar com gentilezas.

Pode ser que chegue o dia em que as máquinas serão tão semelhantes a nós que essa cautela será justificável (Véliz, 2016). É provável que eles precisem ter corpos semelhantes aos nossos. Pode não existir algo como uma mente sem corpo (Varela; Thompson; Rosch, 1991). E pode ser que apenas seres com corpos biológicos possam ser sencientes. Parte do que nos faz sentir e pensar do modo que

fazemos é o nosso coração batendo mais rápido quando estamos empolgados, nossa pressão arterial caindo quando sentimos tristeza, o suor escorrendo pela nossa testa quando estamos com medo, nossa pele arrepiada quando estamos impressionados. O ônus da prova, então, recai sobre quem quiser argumentar que algoritmos são sencientes.

II. Capacidade moral e algoritmos

Há uma tendência na literatura de descrever sistemas automatizados de modo que implica algum grau de capacidade moral[1] (Sharkey, 2017). Em alguns casos, referências implícitas à capacidade moral parecem ser um mero dispositivo retórico. No espírito da postura intencional de Daniel Dennett, às vezes pode ser útil descrever artefatos como se tivessem intenções e crenças, mesmo quando pensamos que são desprovidos de consciência. Wendell Wallach e Colin Allen (2009, p. 14) escrevem, por exemplo, que "sistemas sem motoristas colocam as máquinas na posição de realizarem decisões em frações de segundo que podem ter implicações de vida ou morte". De modo semelhante, Robert Sparrow (2007, p. 70) escreve que o que torna sistemas de armas automatizados é "o fato de que eles têm a capacidade de escolher seus próprios alvos". Apesar de Sparrow argumentar que esses sistemas não são moralmente responsáveis, a linguagem da tomada de decisão moral parece apropriada somente em referência a agentes morais. Referir-se a sistemas autônomos com termos que só fazem sentido ao descrever agentes morais é moralmente significativo. Como Noel Sharkey (2012, p. 793) coloca:

> "[Alguns termos] funcionam como cavalos de Tróia linguísticos que contrabandeiam uma rede rica e interconectada de conceitos humanos que não fazem parte de um sistema computacional nem de como ele opera. Uma vez que o leitor aceita um termo troiano aparentemente inocente […], ele abre caminho para outros significados associados ao uso da linguagem natural."

Essa tendência de falar sobre sistemas autônomos como se fossem agentes morais ou confuso debate sobre IA e capacidade moral talvez tenham levado alguns autores a sugerir que deveríamos parar de nos perguntar se sistemas autônomos são *realmente* agentes morais (Behdadi; Munthe, 2020).[2] Mas essa abordagem parece insatisfatória. Não nos perguntamos se IAs são ou podem ser agentes morais por curiosidade metafísica, mas porque nos importamos com as implicações práticas. Se IAs não são agentes morais, então alguém precisa ser responsabilizado pelo que elas fazem. Se elas forem agentes morais, então deveríamos tratá-las como tal (processando-as quando violarem leis, prendendo-as e considerando-as portadoras de direitos, entre outras práticas)

[1] Na literatura sobre ética e IA, muitos autores (por exemplo, Gunkel, 2012; Floridi; Sanders, 2004) distinguem entre capacidade moral (em suma, a capacidade de realizar julgamentos morais e suportar obrigações morais para com outros) e paciência moral (em suma, ser objeto das ações de um agente moral e ser digno de consideração moral). O presente artigo se preocupa com a capacidade moral, não com a paciência moral. Noutras palavras, o artigo questiona se algoritmos podem ser considerados imputáveis pelo que nos fazem. O que devemos, se é que devemos algo, aos algoritmos está além do escopo do artigo. Penso, contudo, que, se alguém é um agente moral, então é também um paciente moral. O inverso pode não ser verdadeiro.

[2] Dorna Behdadi e Christian Munthe (2020, p. 214) admitem explicitamente que o que motiva sua proposta "é o alto grau de confusão conceitual e a ausência de utilidade prática do tradicional debate AMA [capacidade moral artificial]".

Embora a maioria dos autores use a linguagem da capacidade moral de forma metafórica ou pragmática, em alguns casos parece haver crença explícita de que agentes autônomos podem ser agentes morais. Por exemplo, ao escrever sobre carros autônomos, Mark Coeckelbergh (2016, p. 754) argumenta que "toda a agência é totalmente transferida para o [carro]".

Claro, diferentes autores entendem coisas diferentes quando se referem à capacidade moral. Uma forma dominante de categorizar a capacidade moral na literatura sobre IA é a de James H. Moor (2009), que propõe quatro categorias de agentes éticos: (1) *agentes de impacto ético* (cujas ações têm consequências éticas; a maioria ou todos os sistemas autônomos se encaixam aqui); (2) *agentes éticos implícitos* (projetados com a ética em mente, como caixas eletrônicos); (3) *agentes éticos explícitos* (capazes de agir *a partir* da ética, e não meramente de acordo com ela; podem identificar problemas éticos e encontrar soluções sozinhos); (4) *agentes éticos completos* (fazem o que agentes éticos explícitos fazem, mas com consciência, intencionalidade e livre-arbítrio). Meu argumento é de que agentes éticos explícitos e agentes éticos completos são, na verdade, a mesma coisa. Noutras palavras, alego que não teremos um agente capaz de identificar problemas éticos e responder de forma adequada a eles sem senciência.

Visto que o meu argumento gira em torno das senciência, a perspectiva mais interessante (e contrária à minha) é a de Luciano Floridi e J. W. Sanders (2004, p. 351), que não apenas argumentam que agentes artificiais autônomos podem ser considerados agentes morais, mas também que eles constituem um exemplo de "moralidade sem consciência". Contra Floridi e Sanders, argumento que a senciência é necessária para a capacidade moral, porque a experiência consciente do tipo que permite sentimentos como prazer, dor e empatia parece ser necessária para a capacidade moral.

Um agente é, de modo geral, uma entidade que pode ser a fonte de uma ação. A maioria das concepções filosóficas de agente *moral* defende alguma versão de autonomia ou de responsabilidade moral ou de ambas como constitutivas da capacidade moral. A relação entre esses conceitos é repleta de controvérsias. Alguns filósofos pensam que sempre que há um, o outro também está presente, enquanto outros argumentam que autonomia e responsabilidade moral são independentes entre si. Para nossos propósitos, não é essencial estabelecer a relação exata entre autonomia e responsabilidade moral. O que importa é que, segundo qualquer concepção plausível de autonomia relevante para a moralidade e segundo qualquer concepção plausível de responsabilidade moral, nem zumbis morais nem algoritmos podem ser agentes morais, isso é o que argumento.

III. Autonomia e capacidade moral

Nomy Arpaly argumenta que há pelo menos oito maneiras pelas quais a "autonomia" foi conceituada na literatura: como eficácia pessoal, independência mental, autodeterminação, autenticidade, ter uma autoimagem coerente, ser heroico, autogoverno e ser responsivo a razões. A seguir, analiso cada uma delas e argumento que muitas dessas concepções de autonomia não parecem relevantes para a moralidade e as que parecem sê-lo (autogoverno e responsividade a razões) não se aplicam a zumbis morais ou algoritmos.

III.1 Eficácia pessoal

Essa concepção de autonomia aponta para a qualidade de não depender da ajuda de outras pessoas para navegar pelo mundo. Floridi e Sanders (2004, p. 357) afirmam que os algoritmos são autônomos porque são "capazes de mudar de estado sem resposta direta à interação" – isto é, podem agir independentemente dos seres humanos que os criaram. No entanto, não parece que essa concepção de autonomia seja relevante para avaliar se os algoritmos são agentes morais. Tornados também podem realizar suas atividades sem ajuda humana e isso não nos diz nada sobre se eles são agentes – menos ainda, se são agentes morais.

Autonomia, na filosofia moral e política, é um conceito muito mais rico do que no contexto da ciência da computação e da engenharia. É o tipo mais rico de autonomia que é relevante para a capacidade moral, que não deve ser confundido com o sentido mais vago do adjetivo "autônomo" usado ao falar sobre tecnologia. Uma vez que identificamos uma entidade como um agente moral ou como sujeito de direitos morais, então pode ser relevante saber o que ela pode e não pode fazer independentemente de outros agentes (por exemplo, avaliar quão responsáveis podem ser por suas ações ou investigar se lhes é devida assistência em virtude de suas limitações etc.). Mas ainda não chegamos lá quando se trata de zumbis morais ou algoritmos; saber se eles são pessoalmente eficazes não nos diz nada sobre se eles são agentes morais.

III.2 Independência mental

Essa versão de autonomia é semelhante à anterior, exceto que, em vez de focar na capacidade de *agir* independentemente de outros, ela se preocupa com a capacidade de *pensar* independentemente de outros. Se por "mente" queremos dizer algo como experiência subjetiva, então nem zumbis morais nem algoritmos têm uma mente. Alguns filósofos, no entanto, podem argumentar que a experiência subjetiva não é necessária para ter uma mente. De qualquer forma, parece que a independência mental não serve como critério para estabelecer a capacidade moral. Computadores que jogam xadrez, por exemplo, podem "pensar" independentemente de seres humanos e está bastante claro que não são bons candidatos à capacidade moral. Uma vez que seja evidente que alguém é um agente moral, pode ser moralmente relevante estabelecer se esse agente foi indevidamente influenciado em uma circunstância específica (por meio de propaganda, mensagens subliminares ou estimulação cerebral direta). Mas, novamente, a independência mental não é relevante como teste de capacidade moral.

III.3 Prerrogativa moral de se autodeterminar

É amplamente aceito que agentes morais têm direito à autodeterminação. Como adulto competente, você deve ser capaz de decidir como viver sua vida — desde que não viole os direitos de outras pessoas. O respeito à autonomia é um dos pilares da bioética. É o que permite aos pacientes decidir se querem ser tratados e qual tratamento receber. Embora o direito à autodeterminação seja muito importante, ele pressupõe capacidade moral por parte do titular do direito. O direito à

autodeterminação é uma ferramenta para proteger agentes morais, não é uma qualidade dos agentes e, desse modo, não pode ser um critério para a capacidade moral.

III.4 Autenticidade

Até que ponto alguém age de forma autêntica — seguindo suas verdadeiras convicções ou sendo fiel a quem é — em vez de agir impulsivamente, parece ser importante para a moralidade. Harry Frankfurt (1988) argumentou que algumas situações que poderiam ser interpretadas como situações de fraqueza de vontade são, na verdade, situações nas quais alguém está sendo limitado por seus próprios valores mais profundos. Em resposta, David Velleman (2002) argumentou que tal teoria não se refere à autonomia, mas à autenticidade. Não é importante para nossos propósitos se a autenticidade faz parte da autonomia. Embora a autenticidade possa ser importante para avaliar o caráter e o significado moral das ações, não parece ser um bom teste para a capacidade moral. Podemos imaginar agentes morais inautênticos que, ainda assim, são agentes morais. Considere uma pessoa que não confia em suas convicções mais profundas e, em vez disso, age imitando os outros. Tal comportamento a torna menos admirável como agente moral, mas sua inautenticidade, a menos que seja causada por alguma deficiência cognitiva que questione sua competência, não é motivo para questionar sua capacidade moral.

III.5 Autoidentificação

Dado que a autonomia é frequentemente considerada como relacionada à ausência de pressão externa, a experiência fenomenológica de assumir os próprios desejos e atos é frequentemente considerada também relevante para a autonomia. A pessoa autônoma, segundo essa concepção, é aquela que capaz de se autoidentificar com seus desejos e possui uma autoimagem coerente. A maioria das pessoas fez algo em algum momento de suas vidas que não era do seu feitio, com o qual não se autoidentificavam. Embora o grau de autoidentificação possa ser relevante para julgar o significado moral de uma ação, não parece relevante para julgar se alguém é um agente moral. Suponha que, embora você geralmente seja uma pessoa muito calma, um dia perca a paciência e grite com seu parceiro. Embora você possa não se autoidentificar com essa ação, isso não o destitui de sua capacidade moral. A autoidentificação, portanto, não é, mais uma vez, um teste apropriado de capacidade moral.[3]

III.6 Heroísmo

Muitas abordagens sobre autonomia têm um agente ideal em mente. Para os estoicos, o agente ideal é aquele que exerce a ataraxia; para Aristóteles, é a pessoa que leva uma vida de contemplação; para Nietzsche, é o espírito livre. Ter um ideal pode nos ajudar a avaliar o quão próxima ou distante uma pessoa ou ação está do melhor que poderia ser, mas não pode ser um critério

[3] Acerca da autoidentificação e da autonomia, v. ainda Schroeder; Arpaly, 1999.

para a capacidade moral. Se apenas os poucos que alcançassem o ideal fossem considerados agentes morais, a maioria dos adultos competentes não passaria no teste de capacidade moral.

III.7 Autogoverno e responsividade-a-razões

Uma das formas mais populares de pensar sobre autonomia é em termos de autogoverno. O termo "autonomia" deriva do grego "autos" (próprio) e "nomos" (norma); como tal, o conceito que o termo "autonomia" visa capturar parece ser, em termos gerais, a qualidade do autogoverno, caracterizada pela capacidade do agente de decidir como agir. Mesmo dentro das teorias de autonomia como autogoverno há muita controvérsia e variedade.

A ideia de autonomia como autogoverno pode ser rastreada até Immanuel Kant (2019), para quem autonomia significa que nenhuma autoridade externa a nós mesmos é necessária para ditar as exigências da moralidade (Schneewind, 1992). Ser autônomo implica impor a nós mesmos as exigências da moralidade: temos a capacidade de reconhecer o que é a coisa certa a fazer e agir de acordo.

Discuto autonomia como autogoverno e como responsividade-a-razões em conjunto porque parecem intimamente relacionadas. Mais precisamente, parece que autogoverno requer responsividade-a-razões. Agentes morais tomam decisões e agem de acordo, pelo menos em parte, porque estão respondendo a razões. Se uma entidade não é o tipo de criatura que consegue entender razões, então ela não se governará de uma forma que seja relevante para a moralidade.

É crucial para a autonomia como autogoverno a capacidade de agir de acordo com a razão de uma forma que responda aos próprios motivos (Christman, 2015). Para ser autônomo, é preciso ser capaz de refletir (Watson, 2013, p. 4-5), endossar e agir de acordo com os próprios valores (Christman, 2015). É porque uma pessoa é capaz de escolher seus valores por si mesma e viver de acordo que devemos pedir seu consentimento para interagir com ela de maneiras invasivas, por exemplo, no caso de um procedimento médico. Não precisamos pedir permissão a um algoritmo para modificá-lo ou mesmo encerrá-lo, porque algoritmos não têm valores próprios; eles não se importam com sua própria existência. Nem podem responder a razões enquanto razões. Não se pode *persuadir* um algoritmo a fazer algo dando-lhe boas razões — só se pode programá-lo de uma forma ou de outra.

Um robô poderia responder ao seu ambiente de maneiras conformes à ética (por exemplo, se vir um ser humano carregando algo pesado, oferece-se para carregá-lo). O fato de um ser humano estar com dificuldades para carregar suas compras e de ser fácil para o robô carregá-las, no entanto, não é uma *razão* para o robô — é uma instrução. O robô não pode *desejar* aliviar os braços tensos da pessoa porque, primeiro, não tem desejos e, segundo, não tem ideia de como é ter os braços doloridos por carregar algo pesado. Além disso, o robô não pode refletir sobre a relação entre nossos atos e ter uma cidadania saudável ou sobre os benefícios da amizade civil. Uma razão, para agentes morais, equivale a algo que *importa* para nós, com o qual nos *importamos* porque entendemos seu significado moral.

Algoritmos são programados para fazer algo: vencer uma partida de xadrez, distinguir spam de não spam, identificar pessoas que possam querer comprar um produto, avaliar se um candidato será adequado para uma vaga etc. Algoritmos, no entanto, são incapazes de avaliar normativamente o objetivo para o qual foram criados e modificar seu comportamento de acordo.[4]

Considere o papel que os algoritmos desempenham no avanço de faculdades com fins lucrativos nos Estados Unidos. Essas faculdades caras e de baixa qualidade se propagandeiam para populações vulneráveis como uma saída para sua condição desfavorecida. De fato, no mercado de trabalho, uma pessoa não está em melhor situação por ter um diploma de uma faculdade com fins lucrativos do que por não ter frequentado nenhuma faculdade (Darolia; Koedel; Martorell; Wilson; Perez-Arce, 2015). Para identificar possíveis clientes para uma faculdade com fins lucrativos, algoritmos procuram pessoas nos códigos postais mais pobres que clicaram em anúncios de empréstimos consignados ou cujos históricos de busca revelam preocupação com estresse pós-traumático (O'Neil, 2020, p. 119). Quando esses algoritmos realizam suas tarefas, eles não se perguntam se é moralmente correto predar pessoas vulneráveis e são incapazes de decidir largar seus empregos e buscar uma linha de trabalho mais ética.

Um carro autônomo é incapaz de escolher seu destino por capricho. Ele não pode acordar um dia com o desejo de aproveitar o campo e desobedecer ao seu dono, que precisa trabalhar. Um robô assassino não pode se tornar pacifista depois de considerar as consequências negativas de suas ações. Não é que o robô assassino tenha sido programado para acreditar que sua matança seja moralmente justificada — ele não tem a capacidade de acreditar ou questionar sua *raison d'être*. Ele não pode refletir sobre o que quer, o que vale a pena perseguir ou como deveria viver sua vida.[5]

Em resumo, então, os algoritmos não se autogovernam, porque precisam de informações externas para definir objetivos, nem respondem a razões, pois nenhuma razão jamais poderia "convencê-los" a mudar o objetivo para o qual foram programados.

Zumbis morais, por outro lado, podem *parecer* tanto se autogovernar quanto responder a razões, visto que, por definição, seu comportamento é indistinguível daquele dos seres humanos. No entanto, não está claro se zumbis morais podem ter motivos próprios. Se nada os *move*, se não conseguem sentir desejo, medo, esperança ou empatia, pode-se argumentar que não conseguem

[4] Para ser clara, o problema que destaco aqui não é o da programação dos algoritmos. Um crítico poderia pontuar que seres humanos são programados em certo sentido pela genética e pela cultura. O que importa é que algoritmos não têm senciência para ajudá-los a modificar sua programação. Um ser humano pode ter sido educado para ser religioso, mas um sentimento de insatisfação pode levá-lo a mudar o curso de sua vida. Agradeço a um revisor anônimo por me incentivar nesse quesito.

[5] O argumento central desse artigo é o que de a senciência é necessária para a capacidade moral. Alguém poderia pensar, contudo, que esse argumento está mais relacionado à responsividade-a-razões (ou possibilidade de agir diversamente em resposta a razões). O que sugiro é que a senciência é o ingrediente necessário para que um agente seja motivado por razões (enquanto razões). Tomemos o exemplo de alguém que foi convencido por argumentos filosóficos a doar para a caridade para aliviar a pobreza. Para ser motivado por tais razões enquanto razões, é preciso ter o senso visceral do que é viver em extrema pobreza, da perversidade do extremo sofrimento. Ainda que alguém jamais tenha tido a experiência real d extrema pobreza, é possível extrapolar as próprias experiências de sofrimento para imaginar como seria. Para dizê-lo de outra forma, qualquer ser humano sabe o suficiente sobre o sofrimento para *temer* tornar-se vítima da extrema pobreza, de uma maneira que seres não sencientes não poderiam.

possuir objetivos próprios como os agentes morais. Nesse sentido, qualquer objetivo que perseguem não é deles, pois não têm a capacidade de endossá-lo, de sentir que o aprovam. Da mesma forma, é questionável que zumbis morais respondam a razões, pelo menos em algumas situações morais. Suponha que um ser humano peça a um zumbi moral que pare de pisar em seu pé porque lhe dói. Se o zumbi moral nunca sentiu dor, não está claro se poderíamos dizer que, quando ele parar de pisar no pé da pessoa, estará respondendo às razões dadas por ela.

De acordo com nossa análise até o momento, apenas autogoverno e responsividade-a-razões são relevantes para o tipo de autonomia que, por sua vez, sugere capacidade moral e parece que nem zumbis morais nem algoritmos são autônomos nesses sentidos. Mas focar na autonomia, ainda que intuitivo, pode não ser a melhor maneira de determinar se zumbis morais ou algoritmos são agentes morais. Primeiro, existem tantos sentidos de autonomia, alguns dos quais são difíceis de separar dos outros, que o foco na autonomia em discussões morais corre o risco de incitar mal-entendidos, em vez de contribuir para a clareza (Arpaly, 2002, p. 126). Segundo, descobrir quem é agente moral não é, antes de tudo, uma questão de curiosidade intelectual — é uma tarefa orientada para a prática. O que deveríamos procurar, então, se quisermos saber se pode haver capacidade moral sem senciência, é uma explicação satisfatória da capacidade moral.

Há duas razões práticas principais pelas quais podemos nos importar em determinar capacidade moral. A primeira razão é proteger sujeitos de direitos morais que podem não ser capazes de se proteger a si mesmos. Na ética médica, por exemplo, queremos garantir que os sujeitos da pesquisa e os pacientes estejam em condições de tomar decisões idôneas e bem-informadas das quais provavelmente não se arrependerão no futuro. Nesse contexto, estabelecer autonomia é uma prioridade, pois autonomia é um sinal de que tais pessoas podem decidir por si mesmas o que é melhor para elas e, portanto, devem ter permissão para tomar tais decisões. Essa preocupação não se aplica nem a zumbis morais nem a algoritmos. Dado que zumbis morais e algoritmos não têm a capacidade de sofrer, não nos preocupamos com a possibilidade de se arrependerem de suas decisões ou de involuntariamente causarem danos a si mesmos.

O preocupante sobre algoritmos é que eles podem agir no mundo e ter um enorme impacto, para o bem ou para o mal, o que nos leva à segunda razão pela qual determinar capacidade moral é importante. Quando as coisas dão errado, queremos ter certeza de que sabemos a quem recorrer quando queremos garantir a imputabilidade. Dado que estabelecer quem é um agente moral é um desafio prático, evitarei questões metafísicas (por exemplo, em relação ao livre-arbítrio). Entendo que nossas práticas morais e políticas em relação à capacidade moral são suficientemente fundamentadas para que questões metafísicas não sejam relevantes para fins práticos. Estabelecer a responsabilidade moral, então, é importante em contextos nos quais queremos garantir que as partes apropriadas respondam por possíveis irregularidades.

Em outras palavras, as discussões filosóficas sobre capacidade moral são uma moeda com dois lados: um lado tem o tomador de decisão como objeto de preocupação (autonomia) e o outro lado está

preocupado com a imputabilidade (responsabilidade moral). No contexto de zumbis morais e algoritmos, o que importa é a imputabilidade.

IV. Responsabilidade moral e capacidade moral

A maior parte das concepções de autonomia não é relevante para estabelecer responsabilidade moral. Agentes podem ser moralmente responsáveis ainda quando não têm muita independência mental ou corporal ou quando sua prerrogativa moral de se autodeterminar é violada (por exemplo, escravos podem ser agentes morais responsáveis), mesmo se não são heróis morais, mesmo se não se identificam sempre com sua ação, mesmo se inautênticos. Em contraste, tanto o autogoverno quanto a responsividade-a-razões parecem importantes para a capacidade moral, mas não necessariamente porque são importantes para a autonomia – a não ser que tomemos autonomia e capacidade moral como sinônimos, o que contrariaria a terminologia da maior parte dos filósofos. Na verdade, autogoverno e responsividade-a-razões são importantes para a capacidade moral na medida em que são capacidades que fundamentam a responsabilidade moral.

Gary Watson (2013) argumenta de modo convincente que agentes morais são seres autônomos (no sentido de que se autogovernam) e imputável (moralmente responsável no sentido de que respondem perante outros). Felizmente há muito mais consenso sobre o que é responsabilidade moral do que sobre autonomia.

Michael McKenna (2013, p. 206) entende responsabilidade moral como "imputabilidade por guiar a conduta de acordo com as demandas da moralidade". Imputabilidade se relaciona intimamente com as noções de censurabilidade e louvabilidade morais. De acordo com Arpaly (2002, p. 129), "qualquer agente que é censurável ou louvável por sua ação é, por definição, ao menos em algum sentido moralmente responsável por tal ação".

Assim como não são autônomos, algoritmos não são imputáveis. Como seres imputáveis, "podemos responder a outros pelo modo como conduzimos nossas vidas" (Watson, 2013, p. 1). Isto é, podemos reconhecer os interesses e exigências morais de outros e, quando não os respeitamos, podemos nos sujeitar a receber a culpa ou até punição. Um algoritmo, por outro lado, não pensa sobre o sofrimento que pode causar ao encorajar pessoas vulneráveis a tomar altos empréstimos para pagar por uma graduação numa instituição de ensino com fins lucrativos de pouca ou nenhuma reputação. Quando enganados por um algoritmo, pode não nos ocorrer puni-lo ou pedir uma compensação. Ao invés disso, buscaríamos reparação das pessoas que desenvolveram, implementaram e supostamente deveriam supervisionar o algoritmo.

Floridi e Sanders (2004, p. 371) argumentam que não devemos confundir imputabilidade e responsabilidade. "Um agente é moralmente imputável por x se for a fonte de x", onde x é uma ação que causa bem ou mal morais. Para ser também moralmente responsável, "o agente precisa mostrar os estados intencionais adequados". Creem que misturar os conceitos de imputabilidade e responsabilidade leva a "confundir a *identificação* de x como um agente moral com a *avaliação* de x como um agente moralmente responsável" (Floridi; Sanders, 2004, p. 367). Mas a moralidade

intrinsecamente diz respeito à avaliação normativa. Se um agente moral pode ser identificado como tal, então também deverá ser o caso poder avaliá-lo como responsável por suas ações.

A responsabilidade moral está ligada à responsividade moral. Arply (2002, p. 72) argumenta que, "para que um agente seja moralmente louvável por fazer a coisa certa, ele deve ter feito a coisa certa pelas razões morais relevantes – isto é, as razões pelas quais age são idênticas às razões pelas quais a ação é correta". Quando uma consequência moralmente valiosa resulta de um incidente de pura sorte (i.e. o sol brilhando), ninguém é louvável por isso.

Identificar um agente moral como a fonte da ação equivale a torná-lo o alvo apropriado de louvor ou de censura. Não faz sentido identificar alguém como agente moral sem avaliá-lo como agente moral responsável. Ser um agente moral significa exatamente que alguém é responsável por suas ações morais. Quando agentes morais ferem outros, podemos censurá-los por suas más intenções ou por sua negligência. Em contrapartida, não nos sentimos moralmente indignados contra algoritmos, já que, dado seu desenho e seu input, eles não poderiam agir de outro modo e não têm intenções – não sentem má vontade ou desprezo. "Se a boa vontade – o(s) motivo(s) de que deriva(m) as ações louváveis – é responsividade a razões morais, a deficiência em boa vontade é responsividade insuficiente a razões morais, ignorância ou indiferença a fatores morais relevantes, e a má vontade é responsividade a razões sinistras – razões pelas quais nunca é moral agir, razões que, em essência, conflitam com a moralidade" (Arpaly, 2002, p. 79). Diversamente de pessoas, zumbis morais e algoritmos não podem agir por boa ou má vontade – porque não são sencientes.

V. Senciência e capacidade moral

Afirmo que a principal razão por que algoritmos não podem ser nem autônomos nem imputáveis é que lhes falta senciência. Para ter uma concepção do bem que queremos buscar (autonomia), precisamos ter uma noção do que leva ao prazer, ao que é significativo e ao produz contentamento. Para guiar nossas ações pelo reconhecimento das exigências morais de outros, para que assim possam contar como ações morais (imputabilidade), precisamos ter certa noção da capacidade dos outros de sofrer ou de como é sentir-se ofendido, do que podemos fazer às mentes e corpos de outros por meio de nossas ações.

Não precisamos experimentar todas as formas de dor para sermos capazes de ter empatia pela dor dos outros. Alguém que nunca experimentou dar à luz uma criança pode ter empatia e desejo de aliar a dor de uma mulher sofrendo as dores do parto.[6] Claro, quanto mais próxima a sua experiência daquela de outro, menor a chance de uma lacuna de empatia. Não é por acaso que pessoas que estão passando por dificuldades podem sentir especial conforto de outros que passaram por situação semelhante. Mas basta ter uma noção do que é o prazer e a dor para atuar como agentes morais competentes.

[6] Agradeço a um revisor anônimo pelo exemplo.

A senciência serve como base de laboratório moral interno que nos guia na ação. Quando pensamos em fazer algo, imaginamos as possíveis consequências que poderemos causar e consideramos o tipo de dor ou prazer que poderemos criar, o que nos motiva a agir de um modo ou de outro. Quando concluímos que podemos causar grande dano corporal a alguém, podemos estremecer ao lembrar como é sentir dor física, nossos estômagos contraem enquanto pensamos. Quando imaginamos fazer feliz alguém que amamos, sorrimos e nos deliciamos com essa perspectiva porque sabemos como é prazeroso sentir-se feliz.

Zumbis morais não podem agir por desejo de ferir ou beneficiar alguém. O que entendemos como valores nunca serão valores para uma IA, já que ela não pode sentir o calor do sol ou o fio da lâmina de uma faca, o conforto da amizade e o desprazer da inimizade. No máximo, para uma IA que não sente nada, "valores" podem ser itens numa lista, talvez distribuídos numa ordem de prioridade a partir de um número que representa o peso. Mas entidades que não sentem não podem valorar e seres que não podem valorar não podem agir por razões morais. Zumbis morais são, portanto, incoerentes. Zumbis podem agir de modos que ofendem ou beneficiam seres humanos, mas jamais podem ser agentes morais ou moralmente responsáveis.

Minha perspectiva sobre a capacidade moral é humeana. De acordo com Hume, crenças não são por si sós suficientes para moralmente nos motivar na ação. Precisamos de sentimentos, paixões para nos motivar a agir moralmente (T 2.3.3.4/415, T 3.1.1). Se algoritmos não têm acesso a experiências subjetivas que estão vinculadas a valores, então não terão motivações morais, já que são incapazes de apreciar razões morais.

Muitas concepções parecem implicitamente apoiar a visão segundo a qual a senciência é um requisito para a capacidade e a responsabilidade morais. Por exemplo, Harry Frankfurt (1999, p. 113) afirma que um agente livre está "preparado para endossar ou repudiar os motivos pelos quais age […] para guiar sua conduta de acordo com aquilo que considera que realmente importa". De modo similar, David Shoemaker (2003, p. 94, p. 114) observa que "as emoções que temos fazem de nós os agentes que somos"; "sem a capacidade de sentir, alguém poderia (por definição) ser incapaz de sentir, o que faria seu horizonte de tomada de decisão plano e sem relevo. Sem o investimento emocional acerca do que fazer, todas as opções estão no mesmo pé". De acordo com Arpy (2002, p. 131), para que um agente seja moralmente responsável, ele deve se importar com considerações moralmente relevantes e as características moralmente relevantes das situações devem motivá-lo a agir: "Não se pode censurar ou louvar uma criatura de que não se pode esperar que perceba as características moralmente relevantes de situações mais do que se espera que um elefante perceba aspectos jurídicos [ou] aspectos estéticos".

Uma razão para que a senciência não tenha obtido mais destaque na literatura sobre responsabilidade moral pode ser que, até agora, salvo por causas naturais como idade avançada ou clima, apenas seres humanos podem ser a causa de fenômenos como ofensa e injustiça. Não era importante focar na senciência porque se tratava de um dado. Apenas seres humanos tomaram decisões sobre nossas vidas que poderiam nos afetar negativamente e seres humanos são

evidentemente tanto sencientes quanto agente morais. Zumbis morais eram apenas uma possibilidade teórica. Agora que os algoritmos constituem uma fonte significativa das decisões da sociedade, temos mais motivos para pensar sobre o papel da senciência na capacidade e na responsabilidade morais.

VI. Respondendo a objeções

VI.1. A objeção da equivalência funcional

Um adepto do funcionalismo no campo da moralidade poderia argumentar que nada mais importa para definir um agente moral do que agir como um. Esse crítico poderia argumentar que, para que algoritmos sejam agentes morais, basta que se comportem como tais, que sejam funcionalmente equivalentes a nós. Se *parecerem* capazes de responder apropriadamente a razões morais, tomar decisões para minimizar danos e se forem capazes de modificar seu comportamento em resposta a críticas e punições, então são agentes morais.

Na literatura sobre inteligência, a crítica segue, a preocupação acerca de inteligência sem experiência subjetiva perdeu força com o decurso do tempo. John Searle (1980) criou seu experimento mental do quarto chinês para mostrar que regras computacionais não produzem real compreensão ainda que computadores possam ser programados de modo a imitar compreensão. O experimento mental de Searle inspirou uma explosão de literatura a seu redor, mas o interesse a esse respeito caiu. Aparentemente deixamos de nos importar se assistentes digitais e computadores realmente compreendem – basta-nos que façam o que lhes pedimos. Aparentemente estamos cada vez mais confortáveis em falar de inteligência sem senciência. Por que a equivalência funcional não basta para a capacidade moral?

Minha primeira preocupação acerca dessa objeção diz respeito à teoria não idealizada: podemos nunca conseguir produzir algoritmos que sejam equivalentes funcionais completos de agentes morais humanos. O experimento mental do zumbi moral pode permanecer para sempre um experimento mental. Ainda que algoritmos substituam seres humanos nas tarefas de tomada de decisão, é de se duvidar que chegarão algum dia a replicar o julgamento moral humano. Minha concepção sobre a senciência apoia a visão segundo a qual a moralidade não é codificável. De acordo com a tese da codificabilidade, a ética poderia ser sintetizada num conjunto de regras morais que poderiam ser aplicadas por qualquer um, independentemente da competência moral. É pouco usual, porém, pensar que alguém pode agir moralmente ao seguir uma regra, ainda que lhe falte compreensão ou sabedoria. A moralidade parece ser mais um saber-fazer do um saber algo.

Algoritmos imitando capacidade moral responderão a um conjunto de instruções e conjuntos de instruções podem, no máximo, ser representantes de razões morais. A preocupação é que representantes podem não bastar. Um cientista da computação poderia programar um algoritmo para comportar-se de maneira tal que não faça as pessoas franzirem a testa ou chorarem (uma representação de não fazer pessoas sofrerem), assim aproximando-se de um comportamento

moralmente aceitável. Às vezes, no entanto, fazer alguém chorar é precisamente a ação moral a ser tomada, como quando vacinamos crianças.

Alguém poderia, claro, imaginar inserir exceções no algoritmo – por exemplo, não fazer as pessoas chorarem, exceto se forem crianças que precisam ser vacinadas –, mas o algoritmo ainda estaria perseguindo representações e não a moralidade mesma. Ainda que possa acertar algumas vezes ou até na maior parte das vezes, noutras ocasiões, quando não encontrar uma situação similar anterior ou quando as exceções relevantes não tiverem sido inseridas, a moralidade estará além de seu alcance.

Minha segunda e mais importante preocupação é a de que, se equivocadamente atribuirmos capacidade moral aos algoritmos, as pessoas que são responsáveis pelo dano causado pelo algoritmo serão inalcançáveis, incentivando imprudência no projeto e implementação de algoritmos.[7] Não ganhamos nada ao atribuir capacidade moral funcional a algoritmos. Não nos ajuda a entender melhor os algoritmos e não nos leva a uma melhor imputabilidade. Punir robôs seria mera encenação. Encenar punir um robô (por exemplo, trancando-o numa cela por um tempo) pode ser um interessante espetáculo, mas não seria uma punição real, já que é impossível punir uma entidade que não sente e não valora. A ausência de liberdade só pode ser punição para um ser que valoriza a liberdade.

Pelo contrário, temos muito a perder atribuindo capacidade moral a algoritmos. Se algoritmos se tornarem os alvos (vazios) de nosso louvor e de nossa censura, as pessoas que os projetaram, programaram, encomendaram, implementaram e auditaram não serão encaradas como agentes morais responsáveis que são.

No que diz respeito à inteligência, não nos importamos se o algoritmo "lendo" Shakespeare é capaz de se maravilhar com a beleza de sua linguagem. Basta que ele seja capaz de nos dizer quantas vezes Shakespeare usou uma certa palavra ou qualquer outra coisa que solicitemos. O que acontece "dentro" da IA é irrelevante desde que ela nos entregue os resultados que esperamos. Mas a moralidade é uma esfera de ação bastante diversa. Parte da importância que atribuímos a ela se deve ao fato de há muito em jogo. Quando as pessoas são enganadas ou prejudicadas – e pessoas estão sendo enganadas e prejudicadas por algoritmos –, queremos saber quem é o responsável e se foi intencional; queremos que alguém seja responsabilizado, que a justiça seja feita e que situações similares sejam evitadas no futuro. Se deveríamos dizer que um algoritmo é "inteligente" ou que "compreende" é uma questão conceitual que interessa apenas aos filósofos. Se um algoritmo é um agente moral é uma questão que tem implicações práticas para toda a sociedade.

Na medida em que algoritmos estão sendo encarregados de cada vez mais tarefas, tanto no setor público quanto no privado, será tentador empurrar a culpa para sistemas automatizados quando as coisas derem errado (Danaher, 2016). Temos boas razões para não deixar as pessoas escaparem sem arcar com sua cota de responsabilidade. Algoritmos são ferramentas e seres humanos são responsáveis pelas ferramentas que criam e administram.

[7] Outros autores já apontaram nesse sentido, como Bryson; Diamantis; Grant, 2017 e Birhane; Van Dijk, 2020.

VI.2. A objeção dos impulsos

Alguns cientistas da IA podem pensar que algoritmos sofisticados terão algo como impulsos, que poderiam ser equiparados a motivações e, nesse ponto, estaríamos a um pequeno passo da motivação moral. Steve Omohundro (2008) afirma que, para alcançarem seus objetivos tão eficientemente quanto possível, IAs desenvolverão certos "impulsos básicos" – serão "altamente motivadas" a buscar autoaperfeiçoamento, proteger a si mesmas e adquirir recursos.

O termo "impulso" é tão retoricamente convincente quanto enganoso. Faz parecer que objetivos instrumentais que podem ser programados numa IA podem influenciar o sistema da mesma forma que impulsos físicos e psicológicos influenciam seres humanos – por pulsões fenomenológicas que nos motivam a agir (Bostrom, 2012, p. 76).

Descrever algoritmos como "impulsionados" a cumprir suas tarefas é apenas uma maneira de falar, um modo de dizer que são impelidos a fazer o que são projetados para fazer e que, nesse processo, encontram maneiras mais eficientes realizá-las. Não significa que algoritmos poderiam agir de modo diverso em virtude de uma reflexão mais profunda sobre o sentido da vida ou que querem realizar tais tarefas porque as consideram dignas ou pensam que é certo realizá-las – em todos esses casos, teríamos sintomas de capacidade moral. Um agente só pode mudar seu curso de ação como resposta a razões se puder *sentir* a pulsão da razão. Ainda que o comportamento orientado a objetivos possa nos dizer algo sobre os algoritmos (como o tipo de ferramenta que são e quão perigosos podem se tornar a depender dos objetivos e limites neles programados), não é evidência de capacidade moral.

VI.3. A objeção da inteligência geral

Outra objeção que pode ser levantada contra a ideia de que algoritmos não são agentes morais porque lhes falta senciência é a de que, admitindo que algoritmos não são autônomos nem imputáveis, reputa isso ao fato de não serem espertos o suficiente, não à falta de senciência. Se conseguirmos resolver o problema de inteligência geral com o desenvolvimento de um "algoritmo mestre", então esse algoritmo seria um agente moral. Desse ponto de vista, algoritmos não podem mudar sua linha de trabalho porque tipicamente cuidam apenas de um conjunto limitado de tarefas. Logo, não deveria surpreender que um algoritmo de xadrez seja incapaz de pedir demissão para dedicar-se à filantropia – uma atividade para a qual faltam-lhe as habilidades necessárias.

Uma resposta é que, mesmo faltando ao algoritmo de xadrez as habilidades necessárias para tornar-se qualquer outra coisa, ele poderia, se se opusesse moralmente à sua linha de trabalho (talvez por considerá-la frívola) e fosse capaz de agir diversamente do que fora programado, ao menos desligar-se em protesto. A inteligência parece ser em grande parte independente dos desejos motivadores, na medida em que "mais ou menos qualquer nível de inteligência poderia, em princípio, ser combinado com mais ou menos qualquer objetivo final" (a tese da ortogonalidade) (Bostrom, 2012, p. 73). O algoritmo de xadrez não pode tornar-se filantropo, mas, caso fosse um agente moral,

poderia ao menos *desejar* tornar-se filantropo, tomar isso como seu objetivo final, ainda que incapaz de o perseguir ativamente.

A tese da ortogonalidade propõe que não podemos assumir que IAs sofisticadas compartilharão os mesmos valores tipicamente encontrados nos seres humanos. O robô mais inteligente que conseguirmos criar pode acabar não se importando nem um pouco com o bem-estar dos seres sencientes, com a busca por descobertas científicas, com a cultura refinada, com a ecologia ou com virtudes de qualquer tipo (Bostrom, 2012, p. 83).

Outra resposta é que psicopatas, pode-se dizer, fornecem evidência empírica que sugere que inteligência geral não necessariamente conduz a capacidade moral. Como ocorre com a maior parte dos seres humanos adultos, psicopatas mostram inteligência geral, em alguns casos têm QI superior ao da média da população. Mas psicopatas não tem emoções morais: não "mostram quaisquer sinais genuínos de remorso ou culpa […]. Não se envergonham de [suas] ações, mesmo quando muito erradas, e não sentem qualquer simpatia aparente por suas vítimas" (Levy, 2007, p. 130). Essas deficiências emocionais tornam os psicopatas incapazes de apreciar a moralidade.

Ainda que psicopatas possam falar sobre argumentos morais, eles não têm qualquer experiência somática da moralidade, qualquer preferência pessoal acerca da moralidade. Eles sabem o que as outras pessoas pensam sobre o que é certo e o que é errado e podem relatar isso, mas não têm qualquer convicção moral. Como argumenta Neil Levy (2007), uma vez que psicopatas não apreciam ou respondem a razões morais como razões morais e que não são responsáveis por serem como são, eles não deveriam ser considerados agentes morais integrais. McKenna (2013, p. 223) concorda que sociopatas, sendo "incapazes de entendimento moral", não são agentes morais responsáveis.

Críticos podem objetar que psicopatas parecem ser um contraexemplo ao meu argumento de que senciência é um requisito para a capacidade moral. Psicopatas podem ter deficiências emocionais, mas seria exagero negar-lhes senciência. Meu argumento, porém, é de que senciência é *necessária* à capacidade moral, não suficiente. Emoções morais – a habilidade de sentir empatia, compaixão, arrependimento, culpa e similares – podem também ser necessárias. Mas a senciência é, por sua vez, pré-requisito para emoções morais. Senciência é o alicerce das emoções morais. Se não há nada como um zumbi, então os zumbis não poderiam ter nenhuma experiência subjetiva de culpa ou compaixão.

É possível que as deficiências emocionais dos psicopatas sejam causadas por deficiências de senciência. Parece haver uma correlação entre o desrespeito insensível pelos outros e a insensibilidade à dor: quanto mais tolerantes à dor são os psicopatas, mais insensíveis tendem a ser. Logo, a insensibilidade à dor pode ser um mecanismo que contribui para a insensibilidade ao sofrimento alheio (Brislin; Buchman-Schmitt; Joiner; Patrick, 2016). Psicopatas parecem ser menos sencientes que não psicopatas. Pode haver um limiar de sensibilidade à dor e ao prazer necessário para experimentar emoções morais e desfrutar da capacidade moral. Pode haver outros requisitos necessários para experimentar emoções morais. Independentemente disso, meu ponto principal permanece: seja o que mais for que necessário, a senciência é necessária para a capacidade moral.

Conclusão

O presente trabalho afirmou que zumbis morais – criaturas que se comportam como agentes morais, mas não são sencientes – são incoerentes como agentes morais. Apenas seres que podem experimentar dor e prazer podem entender o que significa causar dor ou prazer e apenas aqueles com esse entendimento moral podem ser agentes morais. "Zumbis morais", como os chamei, são relevantes porque similares aos algoritmos, que tomam decisões morais como seres humanos fariam – determinando quem recebe quais benefícios e penalidades – sem qualquer senciência concomitante.

Pode ser que chegue um tempo em que a IA venha a se tornar tão sofisticada que robôs terão desejos e valores próprios.[8] Não será, porém, em decorrência de sua capacidade computacional, mas em decorrência de sua senciência, que, por sua vez, pode exigir alguma forma de corporificação. Até o presente, estamos longe de criar algoritmos sencientes.

Quando algoritmos causam devastação moral, como frequentemente fazem, devemos nos voltar aos seres humanos que os projetaram, programaram, encomendaram, implementaram e deveriam supervisioná-los para atribuir a censura apropriada. Apesar de toda a sua complexidade e talento, algoritmos são apenas ferramentas e os agentes morais são integralmente responsáveis pelas ferramentas que criam e utilizam.

Referências bibliográficas

ARPALY, Nomy. **Unprincipled virtue**: an inquiry into moral agency. Oxford: Oxford University Press, 2002.

BEHDADI, Dorna; MUNTHE, Christian. A normative approach to artificial moral agency. **Minds and machines**, v. 30, n. 2, 2020, p. 195-218. DOI: https://doi.org/10.1007/s11023-020-09525-8.

BIRHANE, Abeba; VAN DIJK, Jelle. Robot rights? Let's talk about human welfare instead. **AIES '20: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society**, Nova Iorque, 2020, p. 207-213. DOI: https://doi.org/10.1145/3375627.3375855.

BOSTROM, Nick. The superintelligent will: motivation and instrumental rationality in advanced artificial agents. **Minds and machines**, v. 22, n. 2, 2012, p. 71-85. DOI: https://doi.org/10.1007/s11023-012-9281-3.

BRISLIN, S. J.; BUCHMAN-SCHMITT, J. M.; JOINER, T. E.; PATRICK, C. J. "Do unto others"? Distinct psychopathy facets predict reduced perception and tolerance of pain. **Personality disorders: theory, research, and treatment**, v. 7, n. 3, 2016, p. 240-246. DOI: https://psycnet.apa.org/doi/10.1037/per0000180.

BRYSON, Joanna J.; DIAMANTIS, Mihailis E.; GRANT, Thomas D. Of, for, and by the people: the legal lacuna of synthetic persons. **Artificial intelligence and law**, v. 25, n. 3, 2017, p. 273-291. DOI: https://doi.org/10.1007/s10506-017-9214-9.

CAVE, Stephen; NYRUP, Rune; VOLD, Karina; WELLER, Adrian. Motivations and risks of machine ethics. **Proceedings of the IEEE**, v. 107, n. 3, 2019, p. 562-574. DOI: 10.1109/JPROC.2018.2865996.

CHRISTMAN, John. Autonomy in Moral and Political Philosophy. In: ZALTA, Edward N. (ed). **The Stanford Encyclopedia of Philosophy**, Stanford: Stanford University, 2015. Disponível em: https://plato.stanford.edu/entries/autonomy-moral.

[8] Para um argumento sobre por que não deveríamos buscar o desenvolvimento de agentes morais artificiais, v. Van Wynsberghe; Robbins, 2019. Para razões favoráveis e contrárias às pesquisas visando construir máquinas éticas, v. Cave; Nyrup; Vold; Weller, 2019 e Winfield; Michael; Pitt; Evers, 2019.

COECKELBERGH, Mark. Responsibility and the moral phenomenology of using self-driving cars. **Applied Artificial Intelligence**, v. 30, n. 8, 2016, p. 748-757. DOI: https://doi.org/10.1080/08839514.2016.1229759.

DANAHER, John. Robots, law and the retribution gap. **Ethics and Information Technology**, v. 18, n. 4., 2016, p. 299-309. DOI: https://doi.org/10.1007/s10676-016-9403-3.

DANAHER, John. Welcoming robots into the moral circle: a defence of ethical behaviourism. **Science and Engineering Ethics**, v. 26, n. 4, 2020, p. 2023-2049. DOI: https://doi.org/10.1007/s11948-019-00119-x.

DAROLIA, Rajeev; KOEDEL, Cory; MARTORELL, Paco; WILSON, Katie; PEREZ-ARCE, Francisco. Do Employers prefer workers who attend for-profit colleges? Evidence from a field experiment. **Journal of Policy Analysis and Management**, v. 34, n. 4, 2015, p. 881-903. DOI: https://doi.org/10.1002/pam.21863.

FLORIDI, Luciano; SANDERS, J. W. On the Morality of Artificial Agents. **Minds and Machines**, v. 14, n. 3, 2004, p. 349-379. DOI: https://doi.org/10.1023/B:MIND.0000035461.63578.9d.

FRANKFURT, Harry. **Rationality and the Unthinkable**. Cambridge: Cambridge University Press, 1988.

FRANKFURT, Harry. Necessity, Volition, and Love. Cambridge: Cambridge University Press, 1999.

GUNKEL, David J. **The Machine Question**: Critical Perspectives on AI, Robots, and Ethics. Cambridge: MIT Press, 2012. DOI: https://doi.org/10.7551/mitpress/8975.001.0001.

KANT, Immanuel. **Groundwork for the Metaphysics of Morals**. Oxford: Oxford University Press, 2019.

LEVY, Neil. The responsibility of the psychopath revisited. **Philosophy, Psychiatry, & Psychology**, v. 14, n. 2, 2007, p. 129-138. DOI: https://dx.doi.org/10.1353/ppp.0.0003.

MCKENNA, Michael. Reasons-responsiveness, agents, and mechanisms. In: SHOEMAKER, David (ed.). **Oxford Studies in Agency and Responsibility**. Oxford: Oxford University Press, 2013, p. 151-183.

MOOR, James. Four kinds of ethical robots. **Philosophy Now**, n. 72, 2009, p. 12-14.

OMOHUNDRO, Stephen M. (2008) The Basic AI Drives. In: WANG, Pei; GOERTZEL, Ben; FRANKLIN, Stan (ed.). **Proceedings of the first artificial general intelligence conference**. Amsterdã: IOS Press, 2008, p. 483-492.

O'NEIL, Caty. **Algoritmos de destruição em massa**: como o Big Data aumenta a desigualdade e ameaça a democracia. Trad. Rafael Abraham. Santo André: Rua do Sabão, 2020.

SCHNEEWIND, J. Autonomy, Obligation, and Virtue. In: GUYER, Paul (ed.). **The Cambridge Companion to Kant**. Cambridge: Cambridge University Press, 1992, p. 309-341.

SCHROEDER, Timothy; ARPALY, Nomy. Alienation and Externality. **Canadian Journal of Philosophy**, v. 29, n. 3, 1999, p. 371-388. DOI: https://doi.org/10.1080/00455091.1999.10717517.

SEARLE, John R. Minds, Brains and Programs. **Behavioral and Brain Sciences**, v. 3, n. 3, 1980, p. 417-457. DOI: https://doi.org/10.1017/S0140525X00005756.

SHARKEY, Noel E. The evitability of autonomous robot warfare. **International Review of the Red Cross**, v. 94, n. 886, 2012, p. 787-799. DOI: https://doi.org/10.1017/S1816383112000732.

SHARKEY, Amanda. Can we program or train robots to be good? **Ethics and Information Technology**, v. 22, n. 4, 2017, p. 283-295. DOI: https://doi.org/10.1007/s10676-017-9425-5.

SHOEMAKER, David W. Caring, identification, and agency. **Ethics**, v. 114, n. 1, 2003, p. 88-118. DOI: https://doi.org/10.1086/376718.

SPARROW, Robert. The Turing triage test. **Ethics and Information Technology**, v. 6, n. 4, 2004, p. 203-213. DOI: https://doi.org/10.1007/s10676-004-6491-2.

SPARROW, Robert. Killer Robots. **Journal of Applied Philosophy**, v. 24, n. 1, 2007, p. 62-77. DOI: https://doi.org/10.1111/j.1468-5930.2007.00346.x.

VAN WYNSBERGHE, Aimee; ROBBINS, Scott. Critiquing the reasons for making artificial moral agents. **Science and Engineering Ethics**, v. 25, n. 3, 2019, p. 719-735. DOI: https://doi.org/10.1007/s11948-018-0030-8.

VARELA, Francisco J.; THOMPSON, Evan; ROSCH, Elanor. **The Embodied Mind**: Cognitive Science and Human Experience. Cambridge: MIT Press, 1991. DOI: https://doi.org/10.7551/mitpress/6730.001.0001.

VÉLIZ, Carissa. The challenge of determining whether and A.I. is sentient. **Slate**, 14 abr. 2016. Disponível em: https://slate.com/technology/2016/04/the-challenge-of-determining-whether-an-a-i-is-sentient.html.

VELLEMAN, J. David. Identification and Identity. In: BUSS, Sarah; OVERTON, Lee (ed.). **Contours of Agency**: Essays on Themes from Harry Frankfurt. Cambridge: MIT Press, 2002, p. 91-123. DOI: https://doi.org/10.7551/mitpress/2143.001.0001.

WALLACH, Wendell; ALLEN, Collin. **Moral Machines**: Teaching Robots Right From Wrong. Nova Iorque: Oxford University Press, 2009. DOI: https://doi.org/10.1093/acprof:oso/9780195374049.001.0001.

WATSON, Gary. Moral Agency. In: LAFOLLETTE, Hugh (ed.). **The International Encyclopedia of Ethics**. Disponível em: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781444367072.wbiee294.

WINFIELD, Alan F.; MICHAEL, Katina; PITT, Jeremy; EVERS, Vanessa. Machine ethics: the design and governance of ethical AI and autonomous systems. **Proceedings of the IEEE**, v. 107, n. 3, 2019, p. 509-517. DOI: https://doi.org/10.1109/JPROC.2019.2900622.