Paakat: Revista de Tecnología y Sociedad
e-ISSN: 2007-3607
Universidad de Guadalajara
Sistema de Universidad Virtual
México
paakat@udgvirtual.udg.mx

Año 10, número 19, septiembre 2020-febrero 2021

A falácia lúdica das três leis: Ensaio sobre inteligência artificial, sociedade e o difícil problema da consciência

The playful fallacy of the three laws: Essay on artificial intelligence, society and the difficult problem of conscience

La lúdica falacia de las tres leyes: ensayo sobre inteligencia artificial, sociedad y el difícil problema de la conciencia

Alexandre Quaresma*
http://lattes.cnpq.br/6089915050124806
Pontificia Universidad Católica de São Paulo, Brasil

[Recibido 12/03/2019. Aceptado para su publicación 14/08/2020]
DOI http://dx.doi.org/10.32870/Pk.a10n19.478

Resumo

O objetivo desse ensaio é refletir de maneira crítica sobre a condição de nossos mais modernos sistemas de inteligência artificial, seus potenciais latentes, suas possibilidades técnicas úteis à sociedade, e também sobre suas intrínsecas limitações estruturais, extrapolando hipóteses acerca do hard problem da consciência e sua sonhada superação, especulando também sobre as possíveis interações desses sistemas inorgânicos pretensamente conscientes e superinteligentes com as sociedades contemporâneas de hoje e de amanhã, e suas possíveis consequências indesejáveis para os próprios corpos sociais envolvidos nessas relações – em grande medida – disruptivas. Para tanto, usaremos como objeto teórico as famigeradas "três leis da robótica" de Isaac Asimov, afim de demonstrar a fragilidade inverossímil de tal hipótese ficcional, quando confrontada com a realidade fática e também com o próprio estado atual da arte em inteligência artificial e sistemas computacionais. Ademais, ainda que o problema duro da consciência em

sistemas cibernético-informacionais permaneça inalterado em sua insolubilidade persistente, não é difícil imaginar robôs com vários níveis e graus de inteligência distintos interagindo com os seres humanos nos mais diversos setores e ambientes da vida cotidiana social, como de fato já começa a acontecer, e as subsequentes consequências e desdobramentos sociais. Por isso é muito difícil crer que meras "três leis" lúdicas, oriundas da ficção científica, pudessem ser instrumento suficiente e satisfatório para o enfrentamento das questões que envolvem seres humanos e sistemas inorgânicos dotados com inteligência artificial (IA).

Palavras-chave

Inteligências artificiais; cognição; hard problem; robótica; crítica da tecnologia.

Abstract

The objective of this essay is to reflect critically on the condition of our most modern artificial intelligence systems, their latent potentials, their technical possibilities useful to society, and also on their intrinsic structural limitations, extrapolating hypotheses about the hard problem of consciousness and its dreamed of overcoming, speculating also about the possible interactions of these supposedly conscious and superintelligent systems with contemporary societies of today and tomorrow, and their possible undesirable consequences for the very social bodies involved in these - to a large extent - disruptive relationships. To do so, we will use Isaac Asimov's infamous "three laws of robotics" as a theoretical object, in order to demonstrate the unlikely fragility of such a fictional hypothesis, when confronted with factual reality and also with the current state of the art in artificial intelligence and systems computational. Furthermore, even though the hard problem of consciousness in cyber-informational systems remains unchanged in its persistent insolubility, it is not difficult to imagine robots with different levels and degrees of intelligence interacting with humans in the most diverse sectors and environments of everyday social life, as it actually begins to happen, and the subsequent consequences and social developments. That is why it is very difficult to believe that mere playful "three laws", originating from science fiction, could be a sufficient and satisfactory instrument to face issues involving human beings and inorganic systems endowed with artificial intelligence (AI).

Keywords

Artificial intelligence; Cognition; Hard problem; Robotics; Criticism of technology.

Resumen

El objetivo de este ensayo es reflexionar críticamente sobre la condición de nuestros más modernos sistemas de inteligencia artificial, sus potenciales latentes, sus posibilidades técnicas útiles para la sociedad y, también, sobre sus intrínsecas limitaciones estructurales. Se extrapolan hipótesis acerca del problema difícil de la consciencia y su soñada superación, además de especular también sobre las posibles interacciones de esos sistemas inorgánicos supuestamente conscientes y superinteligentes con las sociedades contemporáneas de hoy y de mañana, y sus posibles consecuencias indeseables para los propios cuerpos sociales involucrados en estas relaciones (en gran medida) disruptivas. Para ello, usaremos como objeto teórico las famosas "tres leyes de la robótica" de Isaac Asimov, con la finalidad de demostrar la fragilidad inverosímil de tal hipótesis ficcional, sobre todo cuando es confrontada con la realidad fáctica y también con el propio estado actual del arte en lo referente a inteligencia artificial y sistemas computacionales. Aunque el problema duro de la consciencia en los sistemas cibernético-informacionales permanezca sin alteración en su irresolución persistente, no es difícil imaginar robots con varios niveles y grados de inteligencia distintos interactuando con los seres humanos en sus más diversas áreas y ambientes de la vida cotidiana social, como ya de hecho empieza a suceder, y sus subsecuentes consecuencias y desdoblamientos sociales. Por eso es muy difícil pensar que solo "tres leyes" lúdicas, originadas en la ciencia ficción, puedan ser un instrumento suficiente y satisfactorio para enfrentar las cuestiones que involucran a los seres humanos y a los sistemas inorgánicos dotados de inteligencia artificial (IA).

Palabras clave

Inteligencias artificiales; cognición; hard problem; robótica; crítica de la tecnología.

Introdução

> O ensaio [...] tem a ver com os pontos cegos de seus objetos. Ele quer desencavar, com seus conceitos, aquilo que não cabe em conceitos, ou aquilo que, através das contradições em que os conceitos se enredam, acaba revelando que a rede de objetividade desses conceitos é meramente um arranjo subjetivo
>
> Theodor Adorno (2003, p. 44)

Nada nem ninguém, poderá demover um computador, androide ou robô, pretensamente consciente, dotado com IA, de fazer algo acerca do qual esteja certo de que seja o mais apropriado a ser feito naquela situação específica em que ele se encontre num dado momento, segundo seu programa. Ainda que na realidade factual ele esteja provocando sinistros, catástrofes, danos sociais graves e até mesmo a morte de seres humanos devido às suas ações pretensamente inteligentes, mas hipoteticamente desastrosas.

Se o sistema cibernético-informacional realmente estiver determinado acerca de que tal ou qual ação que empreende é a mais apropriada, segundo sua codificação interna pregressa original, segundo também seus parâmetros e calibragens prévias, não importará o argumento que lhe seja apresentado, nem o apelo emocional acerca do que possa acontecer caso ele prossiga com sua ação resoluta, pois o sistema necessariamente segue e sempre seguirá regras rígidas, determinísticas, fixas, sem quaisquer ambiguidades ou exceções possíveis, ainda que essas regras e procedimentos –no caso da IA– sejam agrupados e arranjados de tal modo por seus projetistas, que visem justamente dar a impressão de que exista alguma espécie de vida no interior do sistema computacional automatizado, ou mesmo um alguém subjetivo qualquer que pudesse experimentar consciência, subjetividade, que tivesse protagonismo ontofenomênico, animando a máquina, impulsionando e movendo suas ações.

Na verdade, se ignorarmos as aparências externas e deixarmos de lado os antropomorfismos, e observarmos com atenção a constituição e a lógica internas dos referidos constructos e sistemas, veremos que o que temos de fato em termos discerníveis é a execução contínua de uma série de programas, rotinas e protocolos, o que significa dizer, ações acéfalas, e que assim, no fundo, não poderá haver ali valores a se respeitar, sentimentos reais a se sentir, se compreender e se expressar, existência vívida, subjetividade experiencial, nem muito menos ainda um agente atuante com perspectiva de primeira pessoa, ancorado e conectado ao mundo e à cultura.

A máquina –o que vale dizer, o computador, o androide e o robô– apenas executa obedientemente o seu programa rígido e pré-determinado –isso, quando "as coisas funcionam bem", sem bugs ou panes–, e ainda que esse programa excepcional incluísse representações do que sejam valores humanos, experiências ou sentimentos, ou ainda ética, juízo e discernimento, o que haverá

se manifestando ali no constructo não serão reais valores, experiências ou sentimentos, não no verdadeiro sentido dos termos, nem muito menos ainda poderá haver ética, juízo e discernimento algum, mas sim a mera simulação computacional das representações abstratas de tais propriedades no interior da máquina, representações estas completamente acéfalas e destituídas justamente das qualidades principais que pretensamente deveriam sentir, exprimir e compreender.

E mesmo que impuséssemos uma série de regras e até mesmo um código de conduta protocolar de segurança, que visasse proteger os humanos nas relações sociais com tais sistemas inorgânicos, ainda assim, a interação de tais constructos com a realidade do ambiente societal e suas complexidades sempre gerará situações imprevisíveis a priori para o projetista e programador ou mesmo engenheiro de *software*, já que a realidade "lá fora", além do mundo abstrato das ideias e reflexões, é extremamente dinâmica e complexa, fluida, e essa fluidez complexa e dinâmica não é computável.[1]

Cândido seria aquele que, por desventura ou malogro, inocência crédula ou mesmo desconhecimento da especificidade técnica do assunto hora em tela, acreditasse que meras "três leis" lúdicas, oriundas de um romance de ficção científica[2] do século passado, pudessem –na vida real e de fato, no século XXI– nos proteger dos comportamentos imprevisíveis e/ou indesejáveis de nossos próprios constructos cibernético-informacionais dotados com inteligência artificial (IA).

Seria uma grande tolice crer que a inteligência artificial só nos trará benefícios e confortos no futuro, e nunca crises e problemas. Pensá-los (crises e problemas) nesse momento que antecede a possível resolução do chamado "problema duro ou difícil" da consciência (hard problem), pode ser não apenas salutar, como também útil, já que existe no horizonte até mesmo a possibilidade extrema do surgimento de uma superinteligência,[3] ou mesmo de várias, o que eleva potencialmente o risco para um outro patamar por hora pouco preciso ou mensurável, cujas consequências periclitantes não podemos de modo algum prever *a priori*.

Mas, ainda que consigamos criar e ainda dominar uma superinteligência, de maneira a explorá-la sempre e unicamente a nosso favor –isso tudo num cenário ótimo sem maus funcionamentos ou panes–, tal constructo ainda estaria irremediavelmente determinado por suas próprias estruturações intrínsecas relativas à sua codificação original, e tal codificação original por sua vez não permite e não prevê o incomputável. Computadores não computam o incompreendido, o irredutível, o irrepresentável, logo e por estas mesmas razões, eles não computam também –por motivos técnicos igualmente óbvios– o incomputável.

Assim, como pano de fundo de toda essa problematização teórica acerca da inteligência artificial (IA), está 1) a nossa incapacidade inconteste de compreender o funcionamento objetivo da mente biológica que gera a consciência; 2) a nossa incapacidade subsequente de representar algo que não compreendemos e que talvez nem seja passível um dia de uma representação formal, e 3) também a conclusão lógica da qual não podemos nos esquivar de que sem uma compreensão e representação formal é impossível reproduzir, copiar ou mesmo simular o fenômeno da mente consciente que se pretende replicar.

Em síntese, sem compreensão do fenômeno, e sem sua redução e representação formal, não pode e nem poderá haver reprodução, cópia ou simulação. Mesmo porque, como reproduzir, copiar ou mesmo simular algo que não compreendemos? E esse é, em linhas gerais, o famigerado hard problem, ou mais simplesmente o difícil problema de conseguir reproduzir o fenômeno da consciência em sistemas cibernético-informacionais, dentro do âmbito que se costuma chamar de IA forte, em contraposição à IA fraca, que são justamente os sistemas especialistas que absolutamente dispensam qualquer tipo de consciência, pois executam uma única função específica.

Antes de avançar, explicitamos que é óbvio que as tais leis mencionadas não podem e nem poderiam fazer sentido real para um robô pretensamente inteligente, nem muito menos ainda funcionar, no sentido prático e objetivo de tentar proteger os seres humanos nas relações com seus sistemas artificiais, sejam eles computadores, androides ou robôs. De modo que, nesse ensaio, tomaremos as "três leis" apenas como uma espécie de objeto teórico exemplificador do que não pode e nem poderia funcionar na prática, como também para poder provocar a problematização e especulação de cenários futuros possíveis envolvendo robôs superinteligentes e seres humanos, bem como refletir sobre os problemas que certamente emergirão junto com a interação mais intensa e disseminada com estes sistemas computacionais de IA.

Ipso facto, nosso objetivo será demonstrar o quão falaciosas são as "três leis da robótica" no que tange à vida real e as relações que travamos com nossos sistemas, usando ainda sua hipótese improvável e inverossímil para demonstrar, por consequência, a quantas anda o nosso estado atual da arte em sistemas de inteligência artificial. Tomemos então uma a uma as próprias leis e suas formulações enunciativas, e constatemos suas principais fragilidades e inconsistências não apenas filosóficas, mas também objetivamente práticas e técnicas.

Sobre a "primeira lei"

Um robô não pode prejudicar um ser humano ou, por inação, permitir que um ser humano sofra algum dano. Ora, o problema dessa primeira lei começa imediatamente com a própria formulação e a terminologia utilizada, e também com a definição do significado objetivo desta mesma terminologia na própria elaboração enunciativa da referida lei. Num só termo, tudo isso referenciado e enunciado nessa primeira lei dependeria necessariamente da superação da restritiva barreira do *hard problem* da consciência em sistemas cibernético-informacionais, o que sabidamente ainda não se concretizou e nem há previsões para a sua concretização.[4] E sem uma consciência instanciada no sistema, como fazer o robô compreender o que significa "prejudicar", por exemplo? Ou ainda explicar a ele o que seja causar dano, ou o significado da expressão "sofra dano"? Mais difícil ainda seria definir o que seja *ser* "um ser humano".

Nesse sentido, como definir e representar tal coisa para o robô e ainda simular artificialmente a percepção experiencial do que seja sofrer um dano ou causar dano, se sabemos que a máquina não sente e nem pode sentir sofrimento ou dor? Além disso, tudo é complexo demais e repleto de ambiguidades e armadilhas quando tratamos de IA. Muitas vezes, quando se atribui ao sistema a adjetivação de inteligente, tudo se resume apenas a meros jogos de palavras, figuras de expressão, e nada mais.

Ainda quanto à primeira "lei", o problema de definição é de fato monumental, já que a impossibilidade de definição precisa ou mesmo parcial da consciência biológica impõe uma outra impossibilidade, ou seja, a impossibilidade de representação, e esta, por sua vez, também impõe uma outra restrição, que é justamente a impossibilidade de cópia, reprodução ou simulação, já que para nós as perguntas centrais ainda persistem completamente em aberto: o que é afinal um ser humano? Como funciona a mente consciente?

Essas perguntas nos assombram e nos perseguem há milênios, sem que haja uma definição acabada e consensual sobre suas respostas, e talvez nunca existam. Sendo que, por outro lado, há infindáveis definições que concorrem juntas e até mesmo se somam ou se anulam num monumental arcabouço teórico, poroso e multifacetado que, repleto de ambiguidades e conflitos de hipóteses, tenta expressar o que seja ser um humano. O que significa dizer que existem infinitas definições, e todas elas dependem de um conjunto de crenças e valores contíguos e indissociáveis que as sustentam, numa teia simbólica complexa de intersubjetividade que nós humanos temos enorme facilidade em conceber, significar, representar, e ainda transmitir às novas gerações.

Mas o que seria afinal um ser humano da perspectiva hipotética de uma máquina computacional pretensamente consciente? Como implantar nela um conceito e um valor acompanhado de seu simbolismo e significado, sendo que

nem mesmo nós sabemos como defini-los com a precisão necessária à reprodução do fenômeno? E o termo "permitir", o que significaria do ponto de vista do sistema? Bem, a resposta mais lógica e razoável é que um código cibernético-informacional é, por essência, absolutamente restritivo em sua concepção e em seu ordenamento, e principalmente em sua execução, de maneira que ele vai permitir ou não permitir de acordo unicamente com sua codificação original, o que significa dizer que, ou ele sempre permite, ou ele nunca permite, conforme foi programado.

Nesse âmbito não pode haver nada em aberto e indefinido a ser decidido e definido pelo próprio sistema, pois o sistema não decide, não define, ele apenas executa sua programação original pregressa, e isso é tudo o que ele consegue e pode fazer.

No sentido de esclarecimento, hipotetizemos a título de um ou dois exemplos apenas –lembrando que poderíamos elencar diversos outros possíveis–, e digamos que há dois seres humanos armados e em conflito, que se encontram numa dada situação social de enfrentamento –um assaltante e um segurança privado de loja comercial, por exemplo, durante um roubo–, e que ambos no mesmo momento deem ordens semelhantes e conflitantes a um mesmo robô policial que possa intervir a favor de um dos dois. A quem o robô obedeceria, se a premissa maior é simplesmente "não pode prejudicar um humano"? Imaginemos uma outra breve situação hipotética onde haja dois humanos clamando por socorro diante de um mesmo perigo iminente e fatal, sendo que ambos não podem ser salvos ao mesmo tempo.

Quem deverá ser socorrido primeiro? Quais são os critérios que devem balizar a escolha do sistema diante de uma mesma e única diretriz gravada em seu código?5 A resposta desconcertante é que tal situação não é computável e nem previsível *a priori*, e é exatamente por isso que ainda não temos robôs conscientes trabalhando como policiais ou seguranças armados nas ruas e cidades do mundo. Por hora, apenas no mundo lúdico da ficção científica o robô pode ser consciente, ter subjetividade, superando assim com facilidade o hard problem, de modo a poder também distinguir num átimo a qual humano deveria obedecer.

De maneira que a "primeira lei" seria –se fosse real– absolutamente inócua e ineficaz, no que toca a tentar garantir a segurança dos seres humanos nessas novas formas de relações que emergem junto com a concepção e uso social desses constructos artificiais. E ainda que se persiga a meta de tentar tornar tratável o problema difícil da consciência, muito ainda se tem para trilhar, não apenas para conhecer a consciência biológica, como para reproduzi-la artificialmente, frisando que não possuímos nada nem semelhante à consciência no interior de nossos computadores, androides e robôs da atualidade, e isso é um impeditivo irremovível para obedecer qualquer lei, inclusive a "primeira".

Em síntese, e em termos de raciocínio lógico, o problema é que 1) não possuímos ainda robôs nem sistemas capazes de sopesar qualidades e conveniências de situações reais da vida cotidiana, pelo simples fato de não termos também robôs nem sistemas com níveis de consciência semelhantes ou sequer parecidos com os dos seres humanos (*hard problem*). E mesmo que tivéssemos robôs superpotentes e superinteligentes, 2) ainda não existe uma forma conhecida para a redutibilidade do fenômeno da consciência –nem da própria, nem da dos outros–, sendo a consciência um fenômeno absolutamente subjetivo, e necessariamente de primeiríssima pessoa, e que não pode ser mensurado nem experienciado por outrem.

O que nos leva à seguinte situação em grande medida conclusiva: 3) se não podemos reduzir o fenômeno da consciência e nem representá-lo em alguma linguagem formal conhecida, também não podemos nem poderemos imitar, reproduzir e nem mesmo simular tal fenômeno computacionalmente, o que por sua vez, inviabilizaria na prática a "primeira lei", tornando-a, como já foi dito, inútil no mundo real e factual.

Sobre a "segunda lei"

Um robô deve obedecer às ordens que lhe são dadas por seres humanos, exceto quando tais ordens conflitarem com a Primeira Lei. Essa segunda lei nos parece um tanto tautológica, no sentido de que robôs, como máquinas computacionais que são, apenas e tão somente seguem ordens rígidas da lógica e da matemática que lhe são dadas por seres humanos *a priori*, e até aí não há e nem haveria necessidade de nenhuma diretriz nova ou mesmo lei, e é isso que elas fazem afinal: seguem seu código, executam suas rotinas rígidas e protocolos, e na verdade não é absolutamente disso que se trata. O real problema diz respeito às referidas "ordens que lhes são dadas por seres humanos" *verbalmente*, incluso aí todas as outras implicações que emergem junto com a asserção ordenativa em quaisquer atos de fala.

E de novo: como o sistema computacional poderá triar e identificar a natureza da ordem dada, ou ainda se a ordem é emitida de boa-fé ou não por seus interlocutores humanos, e como ele poderá identificar a voz de comando entre tantas vozes humanas possíveis que podem lhe ordenar tal ou qual ações ao mesmo tempo? Se ao contrário, o robô só obedecer a uma única voz pré-programada, aquela credenciada e pré-autorizada perante o sistema, tal diretriz seria inútil, pois uma voz não credenciada ou não autorizada não seria sequer considerada, e o sistema e sua operação mundana seriam muito mais restritos em sua utilidade social.

Mas, se um sistema complexo o suficiente puder interagir cognitivamente e em alto nível com os seres humanos, o fato de ser apenas "humano" –podendo

ser um bandido, terrorista ou vítima– já seria um dado suficiente para que o robô desse crédito a esse seu interlocutor e o obedecesse, conforme a famigerada "segunda lei" ordena? Ademais, esse segundo ditame previsto nessa hipotética "segunda lei" parece –premeditadamente– delegar ao próprio sistema computacional a capacidade de discernir diante de conflitos complexos e complicados, de deliberar conscientemente, de poder optar em meio ao conflito de ideias, teses e teorias, a hipótese mais provável em termos de benefício em relação ao que foi pré-implantado em seu código original.

O problema é que não existe um tal código com tamanha complexidade, e que ainda seja capaz de replicar o que seja uma consciência deliberando, ou o ato de "deliberar", pois não há ali um agente, uma intencionalidade, um sujeito, ou ainda uma perspectiva de mundo, de experiência subjetiva, ou outra forma de consciência qualquer.[6] O que há pragmaticamente é apenas uma máquina seguindo o seu programa pré-implantado, e isso é tudo o que pode acontecer no interior de um computador, androide ou robô.[7]

Enfim, não é crível que um sistema computacional essencialmente binário, que não tolera ambiguidades de nenhuma espécie, possa ter alguma maneira de lidar com a compreensão do que seja e signifique a "primeira lei" (*Um robô não pode prejudicar um ser humano ou, por inação, permitir que um ser humano sofra algum dano*), nem muito menos lidar com a tarefa monumental de tentar discernir entre possíveis ações futuras conflitantes, hipotéticas e concorrentes –o que acontece o tempo todo no cotidiano dos humanos–, e nem muito menos ainda poder extrair delas a melhor e mais apropriada ação a ser empreendida, à revelia das demais existentes e possíveis, como fazem trivialmente os organismos biológicos.

E síntese, a segunda lei –assim como a primeira– seria absolutamente inócua e ineficaz, principalmente no que tange a tentar proteger os seres humanos dos maus comportamentos dos sistemas computacionais, androides e robôs. E uma vez que a "primeira lei" também se mostrou igualmente ineficiente e inútil, a "segunda lei" –que se baseia na primeira– também não pode nem poderia funcionar e cumprir sua pretensa função de resguardar os seres humanos das ações danosas de, por exemplo, robôs descontrolados. Uma tríade de leis dessa natureza só poderia funcionar de fato na ficção científica, onde os robôs e os demais sistemas já igualaram ou superaram a capacidade cognitiva humana, e por isso conseguem produzir façanhas extraordinárias que, na vida real e por hora, apenas organismos biológicos puderam e podem alcançar.

Lembrando que a mais inteligente das inteligências artificiais chega a parecer uma pálida caricatura falhada de vida diante da capacidade cognitiva e autopoiética real de organismos biológicos bem menos complexos que seres humanos.[8] Mesmo porque, não seria crível a emergência de um fenômeno tão complexo e extraordinário como a consciência, ou mesmo seu engendramento,

por meio simplesmente de um arranjo tecnológico de sistemas inanimados, no interior de uma máquina de computar energizada com eletricidade. Notoriamente, nem a vida e nem a consciência enquanto fenômenos se resumem a apenas isso. Se o *hard problem* da consciência é um empecilho irremovível para a hipotética funcionalidade útil da "primeira lei", o mesmo se dá e se daria com a dita "segunda lei", e pelas mesmíssimas razões, estando assim irremediavelmente aprisionada no mundo ficcional, proibida de se adentrar na realidade factual mundana da concretude, sob pena de ser classificada apropriadamente como falácia.

Sobre a "terceira lei"

Um robô deve proteger sua própria existência, contanto que tal proteção não entre em conflito com a Primeira ou com a Segunda Lei. Livrando-nos definitivamente da falácia lúdica das "três leis", qual seria o robô capaz de tamanha potência cognitiva, senão o da ficção científica de Asimov? Capaz também de resolver crises e dirimir conflitos internos? Segundo consta na literatura especializada internacional, superar o *hard problem* da consciência continua sendo um desafio a ser vencido, e isso num futuro incerto que ainda não é passível de precisão alguma em termos de data.[9]

E sem uma consciência artificial complexa subjetiva e inerente ao próprio sistema, capaz e eficiente, não é possível "proteger sua própria existência", como enuncia a referida "terceira lei", mesmo porque não há "existência" alguma, e nem muito menos ainda uma forma de *saber* se sua ação pretensamente consciente entra "em conflito" ou não com quaisquer outras regras ou leis que se possa imaginar, pois, mais uma vez, é importante frisar: não existe "conflito" no interior do sistema computacional, mas apenas um arranjo tecnológico que executa uma série de rotinas, protocolos e comando específicos, onde não pode e nem poderia haver ambiguidades, dúvidas, divergências, sob pena de se perder a computabilidade do próprio sistema, como também sua operacionalidade objetiva e funcional.[10]

Em um sistema desse tipo, binário, ou é "0" ou é "1", e não tem mais "conversa fiada" acerca de variações ou dúvidas. Não existe um talvez "0", ou quem sabe "1". Isso é impensável e impossível em sistemas computacionais binários, da mesma forma que não há e não poderá haver também fracionamentos e variações de subpedaços de bits, pois não há como grafar na "fita virtual" da Máquina de Turing "1,5", por exemplo, nem muito menos "0,1", "0,2", "0,3", pois para sistemas de inteligência artificial, que por sua vez operam com computadores binários, toda a codificação tem de estar absolutamente clara e definida *a priori* nos bits, onde esteja matematicamente programado e

determinado também o que seja cada bit de informação, sua função, seu ordenamento, sem que se perca, troque de lugar ou altere, um só bit sequer.

Quanto a essa recursividade enganadora sugerida pela "terceira lei", de não poder incorrer na quebra da "primeira" e da "segunda" leis, a questão problemática continua sendo absolutamente a mesma, ou seja, se como demonstramos não é possível para o sistema seguir a "primeira lei" e nem a "segunda lei", com mais razão ainda sustentamos que ele não poderá também seguir a "terceira lei", já que delas dependeria necessariamente para poderem encontrar sua validade, o que –como vimos– não acontece. Repetindo: não pode haver conflito algum no interior do código binário original do sistema, na concepção e formatação do arranjo computacional, na engenharia e arquitetura de *software*, de maneira que o programa –reversamente– vai executar apenas o seu código determinístico e finito original, formalizado e representado pelo programador de antemão, e isso é tudo o que pode acontecer.

E assim, mais uma vez, retornamos forçosamente ao difícil problema da consciência, ou seja, ao *hard problem*, pois somente um robô extremamente sofisticado e desenvolvido, e de alguma maneira insólita consciente –não apenas de si, mas também do mundo e dos complexos sistemas de valores–, poderia enfrentar problemas dinâmicos e complexos como estes, que exigem uma complicada interpretação das famigeradas "leis".

Assim sendo, o termo "proteger", por exemplo –na expressão da máquina computacional–, não será absolutamente nada parecido com o que seja "proteger" para a compreensão trivial dos seres humanos, pois essa *proteção* –no caso a humana– está irremediavelmente atrelada a valores relativos à vida, à própria pessoa humana viva, à dignidade dessa pessoa, o respeito conferido culturalmente à ela, diz respeito também a um corpo biológico que performa no mundo físico e químico, que sente dor, fome, medo, afeto, e que também, no extremo, pode morrer a qualquer instante, e isso vai concedendo significado à própria existência, e tudo isso é absolutamente estranho e de fato inacessível para o sistema cibernético-informacional, seja ele um computador, androide ou robô. Quanto a isso, não há ainda qualquer solução em vista.[11]

Considerações finais

A inteligência expressa pelos sistemas artificiais que construímos e usamos reflete a nossa própria inteligência, no sentido de que são extensões de nossa própria mente inteligente em relação ao meio circundante, aos outros, ao mundo das ideias, mas daí a conseguir instanciar a nossa própria consciência vívida nestes mesmos sistemas artificiais, há um grande e vertiginoso abismo em termos de capacidade e competência técnica nos separando da realidade

factual de nossos dias. Ainda estamos longe de robôs conscientes, ainda que persigamos essa meta diuturnamente, e que o estado da arte também tenha evoluído de forma significativa nas últimas décadas, principalmente no que concerne à IA fraca, ou, mais simplesmente, aos sistemas especialistas.[12]

Em termos conceituais, seria ingenuidade querer crer que uma consciência artificial maquínica pudesse se estruturar de maneira idêntica à consciência biológica, em termos de processos internos, constituição fisioquímica e dinâmicas oscilatórias, algo que, inclusive, está fora de cogitação. Se e quando houver uma inteligência artificial pretensamente consciente, tal inteligência provavelmente se estruturará por meio de informação, energia e dinâmicas de sistemas computacionais complexos, mas como processos distintos dos que ocorrem no carnal, biológico, orgânico.

Além disso, consideramos igualmente ingenuidade crer –diante de tudo o que foi dito até aqui– que computadores, androides e robôs, nunca possam atingir o nível de consciência que os seres humanos possuem hoje, já que a bioevolução continua em curso de modo constante, desdobrando-se em diversas frentes nesse exato instante, e de fato não existem limites fixos para os graus e níveis que a complexidade cognitiva de um sistema inteligente, seja orgânico ou não, podem alcançar.[13] Vale frisar que a mente do ser humano mais rude, tosco e menos preparado intelectual e cognitivamente, está a anos luz à frente de nosso melhor computador com IA, principalmente no quesito experiência consciente, subjetiva, intencional, agenciadora, meta perseguida com obstinação pela IA forte desde os anos 1940.[14]

Mas, extrapolando, nesse sentido de complexidade crescente do estado da arte em IA, um constructo cibernético-informacional que fosse capaz por meio de seu programas de expressar inteligência, exprimir sentimentos, experimentar consciência –ainda que estruturada de forma absolutamente diferente da nossa consciência orgânica e biológica–, por meio da execução de diversos programas e subprogramas replicando os sentidos, por exemplo, talvez também estivesse sujeito a erros, falhas e panes ocasionais, como sabemos que acontece vez por outra com quaisquer equipamentos tecnológicos de nossa cultura, por mais sofisticados que eles sejam ou pareçam ser.

Não é conveniente elencar a lista de desastres e acidentes graves provocados por nossas próprias técnicas e tecnologias, porém, vale rememorar resumidamente que, com o melhor de nossa tecnologia em jogo, naves espaciais já explodiram pouco depois do lançamento, carros autônomos provocaram acidentes fatais com mortos e feridos, e aviões modernos de carreira foram derrubados por mau funcionamento de seus próprios pilotos automáticos –todos estes constructos dotados com sistemas modernos de IA–, levando à morte dezenas de pessoas.

E por mais que se desenvolva a prevenção e a segurança na inteligência artificial como um ideal constante, sempre há e sempre haverá margem para surpresas: o imponderável, o imprevisível, o aleatório, o emergente, as conjunções de fatores; como a nossa própria história pregressa tem nos ensinado tão bem até então. Não é razoável imaginar um mundo asséptico, limpo, ordenado, idealizado, onde todas as tecnologias, incluso a IA, funcionem sempre bem e dentro de seu próprio projeto original de concepção, sem esses tipos de desvio, inconvenientes e imprevistos. A tecnologia de IA, como qualquer outro objeto técnico de nossa cultura, nunca pode ser controlada por completo por seus criadores, pois uma vez que ingressa na cadeia causal dos acontecimentos ordinários do mundo cotidiano, projetam-se a uma espécie de vida e força próprias, de modo que passam a agir e atuar à revelia de nossa própria vontade objetiva, fugindo do controle.

É possível imaginar também para o futuro, um sistema cibernético-informacional dotado com uma forma específica de mente –de novo, onde a mente humana é medida e referência–, imitando a experiência consciente das pessoas, ou baseada na simulação daquilo que conhecemos delas e, com isso, é bem provável que esses sistemas estejam *ipso facto* susceptíveis à falibilidade, ao erro, aos descontroles emocionais, às crises de ansiedade, às falhas de caráter, e até mesmo à própria loucura e a insanidade.

Afinal de contas, a existência não é um fardo leve, e uma parcela significativa das populações mundiais tem problemas mentais e psicológicos de adaptação à vida e à própria existência, o que nos leva à seguinte questão: o que pensaria de si mesmo um programa de computador, androide ou robô, cuja expressão máxima para si mesmo fosse a sua própria existência vívida computacional, executando fiel e docilmente seu código algorítmico originário, enquanto é confrontado com a noção consciente de sua própria consciência experiencial da realidade em que está inserido? Ser um ser vivo e estar consciente de sua própria consciência não é algo fácil, e nós humanos sabemos por experiência própria.

Sem embargos, referimo-nos à hipótese inconveniente de um sistema instável e desequilibrado, de um computador contrariado e vil, de um androide magoado e vingativo, de um robô traído e furioso, de um autômato insano e descontrolado, com tudo aquilo que se é capaz de fazer em tais estados emocionais e em tais situações extremas. Enfim, acreditamos que tudo pode acontecer. Além disso, no que mais esses seres cibernético-informacionais superinteligentes e conscientes seriam também semelhantes a nós?

Experimentariam agressividade, cobiça, maldade? Embriagar-se-iam com seu próprio poder? Revoltar-se-iam contra nós, seus próprios criadores, que como pretensos deuses insuflamos vida no não vivo?[15] Há que se estar atento e observar com discernimento os acontecimentos, para que possamos estar à

altura dos desafios que estão por vir, oriundos de forças poderosas e desconhecidas que nós mesmos estamos colocando em ação.

Ademais, ainda que o problema duro da consciência em sistemas cibernético-informacionais (*hard problem*) permaneça inalterado em sua insolubilidade, não é difícil imaginar robôs com vários níveis e graus distintos de inteligência, interagindo com os seres humanos nos mais diversos setores e ambientes da vida cotidiana social. E, tendo em vista que nossas sociedades se automatizam e informatizam sempre mais, e que a tendência é que seres humanos interajam cada vez mais com sistemas inorgânicos e artificiais, de maneira que, com o próprio aumento das interações, e a complexidade de possibilidades circunstanciais de toda ordem envolvendo seres humanos e, por exemplo, robôs, torna-se relevante a discussão teórica sobre o difícil problema da consciência (*hard problem*) em sistemas cibernético-informacionais, e também sobre os riscos envolvidos nessa disruptiva tecnologização que experimentamos na atualidade.

Por isso, é muito difícil crer que meras "três leis" lúdicas, oriundas da ficção científica, pudessem ser instrumento suficiente e satisfatório para o enfrentamento das questões complexas que envolvem seres humanos e sistemas inorgânicos dotados com inteligência artificial (IA). As "três leis" não podem –e na verdade, não conseguem– nos proteger de nada simplesmente porque nós não sabemos ainda implantar num robô uma consciência suficiente e capaz de discernir e obedecê-las a contento, conforme seu enunciado solicita.

E não sabemos, porque essa consciência suficiente e capaz, cuja medida é inegavelmente o próprio ser humano, não é compreendida em sua dinamicidade complexa corpórea, em seu funcionamento objetivo, e em sua dinâmica interna no próprio ser humano, e nem muito menos ainda tudo isso é redutível e formalizável por meio de uma linguagem formal conhecida e disponível qualquer, incluso a matemática, que se arvora a trazer para si a constituição e estruturação original da própria realidade ao nosso redor, ainda que a biologia insista em lhe escapar, esquivando-se de sua régua puramente quantitativa e nunca qualitativa, sendo que, biologicamente, o mais importante não se quantifica, qualifica-se.

Referências

Adorno, W. Theodor (2003). *Notas de literatura I*. Trad. Jorge de Almeida. São Paulo: Editora 34.

Aleksander, Igor (2014). Machine consciousness: fact or fiction? In: *Footnote:* showcasing research with the power to change our world. 20 fev. Disponível em: footnote.co/machine-consciousness-fact-or-fiction. Acesso em: 17 ago. 2019.

Asimov, Isaac (2005). *Histórias de robôs 3*. Porto Alegre: L&PM Pocket.

Bostrom, Nick (2018). *Superinteligência*: caminhos, perigos e estratégias para um novo mundo. Rio de Janeiro: DarkSide Books.

Goldin, Dina & Wegner, Peter (2004). The origins of the Turing thesis myth. *Technical Report CS04-14*, Providence: Brown University. Disponível em: ftp.cs.brown.edu/pub/techreports/04/cs04-14.pdf. Acesso em: 17 ago. 2019.

Haikonen, Pentti (2003). *The cognitive approach to conscious machines*. Exeter: Imprint Academic.

Haikonen, Pentti (2007). *Robot brains circuits and systems for conscious machines.* Chichester: Wiley.

Kaku, Michio (2001) *Visões de futuro:* Como a ciência revolucionará a o século XXI. Rio de Janeiro: Rocco.

Manzotti, R. & Chella, A. (2018). Good old-fashioned artificial consciousness and the intermediate level fallacy. *Frontiers in Robotics and AI*, v. 39, n. 5. Disponível em: frontiersin.org/articles/10.3389/frobt.2018.00039/full. Acesso em: 18 ag. 2019.

Negrotti, Massimo (2012). From the natural brain to the artficial Mind. *In:* SWAN, Liz. (org.), *Origins of mind*, v. 8: biosemiotics, Drodrecht: Springer Science+Business Media, pp. 399-410.

Searle, John (2017). *Mente, cérebro e ciência.* Lisboa: Edições 70.

Varela, Francisco (2017). *Conhecer*: as ciências cognitivas, tendências e perspectivas. Lisboa: Instituto Piaget.

CÓMO CITAR ESTE ARTÍCULO:

Quaresma, A. (2020). A falácia lúdica das três leis: Ensaio sobre inteligência artificial, sociedade e o difícil problema da consciência. *Paakat: Revista de Tecnología y Sociedad*, *10*(19). http://dx.doi.org/10.32870/Pk.a10n19.478

* Mestre em Tecnologias da Inteligência e Design Digital pela Pontifícia Universidade Católica de São Paulo (PUC/SP), escritor ensaísta e filósofo brasileiro, pesquisador de tecnologias e consequências sociais, com especial interesse na crítica da tecnologia. Autor dos livros *Artificial Intelligences: Essays on Inorganic and Non-biological Systems* (org.), (2019); *Humano-Pós-Humano: Bioética, conflitos e dilemas da Pós-modernidade* (2014)*; Engenharia genética e suas implicações* (org.), (2014)*;* e *Nanotecnologias: Zênite ou Nadir?* (2011). E-mail: a-quaresma@hotmail.com

[1] Quanto a isso, Nick Bostrom (2018, p. 284) acrescenta que "uma máquina provavelmente precisaria ter a capacidade de representar o mundo de uma maneira que fosse ao menos tão rica e realista quanto a representação de mundo que um humano adulto normal possui. [...] Isso está muito além do alcance da IA contemporânea. [...] Uma vez que [...] os processos de planejamento se tornem suficientemente poderosos, também se tornarão potencialmente perigosos".

[2] As famigeradas "Três Leis da Robótica" foram concebidas pelo escritor de ficção científica Isaac Asimov, aqui citadas apud Michio Kaku (1997, p. 164): "1) Um robô não pode prejudicar um ser humano ou, por inação, permitir que um ser humano sofra algum dano. 2) Um robô deve obedecer às ordens que lhe são dadas por seres humanos, exceto quando tais ordens conflitarem com a Primeira Lei. 3) Um robô deve proteger sua própria existência, contanto que tal proteção não entre em conflito com a Primeira ou com a Segunda Lei".

[3] *Superinteligência* é o título do livro de Nick Bostrom (2018). Nele, o autor afirma em tom de alerta que, "se algum dia construirmos cérebros artificiais capazes de superar o cérebro humano em inteligência geral, então essa nova superinteligência poderia se tornar muito poderosa. E, assim como o destino dos gorilas depende mais dos humanos do que dos próprios gorilas, também o destino de nossa espécie dependeria das ações da superinteligência de máquina" (2018, p. 15). Nick Bostrom também (2018, p. 26) aponta: "Defina-se uma máquina ultrainteligente como uma máquina capaz de superar todas as atividades intelectuais de qualquer homem [ser humano], independentemente de quão genial ele seja. Já que o projeto de máquinas é uma dessas atividades intelectuais, uma máquina ultrainteligente poderia projetar máquinas ainda melhores; haveria então certamente uma 'explosão de inteligência', e a inteligência humana se tornaria desnecessária. Desse modo, a primeira máquina ultrainteligente é a última invenção que o homem precisará fazer, contanto que a máquina seja dócil o suficiente para nos dizer como mantê-la sob controle".

[4] Pentti Haikonen (2003, p. 257) afirma em tom de especulação que "a consciência do próprio corpo da máquina, o ponto de vista vantajoso e a autoimagem, são apenas metade da história da autoconsciência. Talvez ainda mais crucial seja a exigência da capacidade da máquina de perceber seu próprio conteúdo mental, o fluxo da fala interior e das imagens internas, o 'filme no cérebro' [numa aparente referência velada a António Damásio], as memórias e a história pessoal e, ainda mais importante, a capacidade de relatar a existência e a propriedade desse conteúdo mental para a máquina em si e para os outros, para poder ter 'pensamentos sobre coisas'".

[5] Nick Bostrom (2018, p. 260), nesse mesmo sentido, indaga: "como o robô poderia comparar um risco grande de que alguns poucos humanos sofram algum mal com um risco pequeno de que muitos humanos sofram algum 'mal'? E qual seria mesmo a definição precisa de 'mal'? Como o 'mal' da dor física deveria ser comparado ao 'mal' da feiura arquitetônica ou da injustiça social? Um sádico seria prejudicado se ele fosse impedido de atormentar sua vítima? Como definimos 'ser humano'? Por que nenhuma consideração é dada a outros seres consideráveis moralmente, tais como animais conscientes não humanos e mentes digitais?". E há mais problemas sérios a serem resolvidos. Como escreve Bostrom (2018, p. 338), "nossos valores e desejos que são aparentemente simples, possuem, na realidade, uma complexidade extrema. Como nosso programador conseguiria transferir essa complexidade para uma função de utilidade?". Todavia, indagamos juntamente com Nick Bostrom (2018, p. 335): "Como implantar algum valor em um agente artificial de modo que ele venha a buscar esse valor como seu objetivo final? Em qualquer domínio mais complicado que um jogo da velha, existem muitos estados possíveis (e histórias possíveis) para que uma enumeração exaustiva seja factível. Assim, um sistema de motivações não pode ser especificado de forma tabular".

[6] Igor Aleksander (2014) afirma que "Quando digo que estou consciente, refiro-me a uma coleção de estados mentais e capacidades que incluem: um sentimento de presença dentro de um mundo externo, a capacidade de lembrar experiências prévias com precisão, ou mesmo imaginar eventos que não aconteceram, a capacidade de decidir para onde direcionar meu foco, conhecimento das opções abertas para mim no futuro, e a capacidade de decidir quais ações tomar".

[7] Como nos informa Aleksander (2014), "as máquinas de hoje não têm mente própria; sua assim chamada inteligência é alcançada através do sangue, suor e lágrimas, de exércitos de programadores humanos brilhantes. Os seres humanos escrevem as regras indispensáveis que fazem com que as máquinas reconheçam sons, respondam a padrões visuais, tomem o próximo passo no xadrez e até sugiram quais ações comprar no mercado de ações. Embora essas máquinas sejam limitadas às tarefas para as quais foram projetadas, um ser consciente tem outra coisa: um sistema complexo de estados internos instanciado através de seus mecanismos neurais". "Outra solução viável para alcançar a consciência robô", informa-nos Manzotti e Chella (2018), "é oferecida pela enação, na medida em que sugere que a experiência é constituída por um corpo e suas interações com o mundo, e, portanto, pode ser implementado em artefatos (O'Regan e Nöe, 2001)". Manzotti e Chella (2018) acrescentam que "os estudiosos que trabalham na consciência robótica sugerem um nível intermediário –padrões sensoriomotores, informações, cognição, espaço de trabalho global– como uma possível explicação para a consciência. [...] O

que está faltando para isso é um nível que deve levar à consciência. De uma perspectiva epistêmica, é como se sugerissem uma explicação sem fornecer seu relacionamento com o que é explicado, isto é, a consciência".

[8] Manzotti e Chella (2018), "a consciência é um fato que precisa encontrar seu lugar na natureza". Nesse sentido, os mesmos autores argumentam que "é claro que encontrar a consciência dentro do mundo físico é necessário quando o objetivo é projetar um robô consciente. Um robô não tem qualquer outro recurso, senão aqueles oferecidos pelo mundo físico. Pode soar raso, mas sem dar ou tomar, todas as abordagens mencionadas correm de acordo com este princípio. Portanto, todas as soluções viáveis exigirão deixar de lado a premissa que até agora prejudicou qualquer progresso —i. e., o difícil problema com a crença geral de que a consciência é algo distinto do mundo físico. Temos que reconsiderar a questão desde o início". Segundo Francisco Varela (2017, pp. 45-46), o que se observou foi que "a inteligência mais profunda e mais fundamental era a do bebê que adquire linguagem a partir de um fluxo cotidiano de palavras dispersas, ou reconstitui ainda objetos significantes a partir de um fluxo difuso de luz. [...] A tarefa mais banal cumprida pelo mais pequeno dos insetos, será sempre efetuada mais rapidamente do que por intermédio da estratégia computacional proposta pela ortodoxia cognitivista. O mesmo acontece com a resistência do cérebro à deterioração, ou com a capacidade da cognição biológica para se adaptar a novos ambientes sem perder por isso competência".

[9] Nick Bostrom (2018, p. 25) sustenta que "máquinas com inteligência geral comparável à dos humanos –ou seja, dotadas de bom senso e capacidade real de aprender, raciocinar e planejar a superação de desafios complexos de processamento de informação em uma vasta gama de domínios naturais e abstratos– têm sido esperadas desde a invenção dos computadores, na década de 1940. Naquele tempo, o advento de tais máquinas era frequentemente esperado para os vinte anos seguintes". No entanto, é interessante notar que, como acrescenta o mesmo autor (2018, p. 25), com uma boa dose de sarcasmo, "desde então, a data estimada para o seu surgimento tem recuado numa razão de um ano a cada ano, fazendo com que ainda hoje futuristas interessados na possibilidade de uma inteligência artificial geral acreditem que máquinas inteligentes surgirão dentro de duas décadas".

[10] Ainda assim, "o objetivo do pesquisador" de IA, escreve Pentti Haikonen (2007, p. 185), "é desenvolver máquinas autônomas, robôs e sistemas que conheçam e entendam o que estão fazendo, e sejam capazes de planejar, ajustar e otimizar seu comportamento em relação às tarefas que lhes são dadas em ambientes em mudança".

[11] "Embora os cientistas práticos da computação há muito tempo tenham ampliado o conceito de algoritmos para além da computação de funções", informa-nos Dina Goldin e Peter Wegner (2004, p. 7), "a ciência da computação teórica manteve a cosmovisão matemática. Apesar do trabalho teórico de complexidade avançada que se aventura fora dessa visão de mundo, como algoritmos on-line e distribuídos, jogos Arthur-Merlin e provas interativas, nosso tratamento da Teoria da Computação no nível de graduação não mudou. Os princípios matemáticos continuam a enquadrar a computação como baseada em funções e a delimitar nossa noção de problema computacional de acordo". Em suma, as três leis, enquanto noções de problemas computacionais, são incomputáveis. John Searle (2017, p. 49) nos alerta que "Não interessa a boa qualidade da tecnologia ou a rapidez com que os cálculos são feitos pelo computador. Se é realmente um computador, as suas operações têm de definir-se sinteticamente, ao passo que a consciência, os pensamentos, os sentimentos, as emoções e todo o resto implicam mais do que uma sintaxe".

[12] Massimo Negrotti (2012, p. 406) nos informa que "um sistema especialista [...] é um tipo de *software* que é capaz de fornecer consultoria, em termos de explicação e previsão, para o usuário em um campo específico de conhecimento, como medicina, lei, ou outro qualquer. O sistema é capaz de fazer isso com uma taxa de sucesso aceitável graças à 'doação' de um perito humano, que decanta, como é seu conhecimento profissional em um banco de dados. Em seguida, o *software*, através de um conjunto de regras inferenciais e estatísticas incorporadas nele, torna-se capaz de entregar a sua consultoria como se fosse, dentro de certos limites, o próprio perito humano". "O ponto-chave", argumenta Negrotti (2012, p. 406), "é que o que é modelado em um sistema especialista não é um cérebro humano, nem uma suposta mente, mas os resultados finais –conhecimento e regras– que os seres humanos obtiveram depois de ter trabalhado durante séculos sobre as melhores maneiras de raciocinar com os fatos dentro de um determinado domínio. É por isso que nenhum programa AI tem sido capaz ainda de propor algum problema novo, embora muitos desses programas sejam, sem dúvida, úteis no domínio de resolução de problemas."

[13] Bostrom (2018, p. 91-92) sustenta em tom de alerta que "não existe razão para supor que o *Homo sapiens* atingiu seu ápice da capacidade de cognição possível em um sistema biológico. Longe de sermos espécie biologicamente possível mais inteligente, é melhor nos enxergarmos como, provavelmente, a mais estúpida das espécies biológicas capaz de iniciar uma civilização tecnológica –um nicho que preenchemos porque chegamos primeiro, e não por estarmos perfeitamente adaptados a ele. [...] O potencial *máximo* de uma máquina inteligente é, claramente, muito superior ao da inteligência orgânica. É possível visualizar o quão grande é essa divergência se levarmos em consideração a diferença de velocidade entre componentes eletrônicos e células nervosas: atualmente, os transistores já são capazes de operar numa velocidade 10 milhões de vezes mais rápida que a dos neurônios biológicos".

[14] John Searle (2017, p. 52-53) nos lembra que "*nenhum programa de computador é, por si só, suficiente para dar uma mente a um sistema. Os programas, em suma, não são mentes e por si mesmos não chegam para ter mentes.* Ora, esta é uma conclusão muito poderosa, porque significa que o projeto de tentar criar mentes unicamente mediante projetar programas está condenado, desde o início [grifos do autor]."

[15] Alguns cenários, como sustenta Bostrom (2018, p. 178), são de fato sombrios, mas, ainda assim, possíveis: "Possuímos dados que mostram que pessoas com um QI de 130 têm maior probabilidade do que pessoas com QI d 90 de se sobressair nos estudos e ter um melhor desempenho em uma gama de tipos de trabalho que demandam um alto nível de cognição. Mas suponhamos que fosse possível, de alguma forma, estabelecer que uma certa IA futura terá um QI de 6455: e daí? [e isso é de fato preocupante] Nós não teríamos a menor ideia do que essa IA realmente poderia fazer".